Final Report on

Influencer Node Maximization (INM): Centrality-based influence maximization approach in a network.

Submitted By:

Saiman Dahal

11873444

Course: Elements of Network Science (CPT_S - 591)

Date of submission: 2024/04/30

Abstract:

Influence Maximization can be defined as the task of determining the maximum important nodes that are highly influential in the social network to maximize the influence spread. This can be a prominent algorithm in terms of marketing and product recommendations. There is a baseline greedy algorithm to maximize the spread of information over the network. In this project, I have developed a centrality-based algorithm implementing the concept of influence maximization. As a part of the implementation, I have taken the Amazon product dataset which consists of the link between the products based on the purchase. Calculating the information spread of a seed set and identifying the minimum seed set that can maximize the marketing information flow in the market are the two main areas of implementation. Here, I have compared my algorithmic approach with the baseline greedy approach.

Introduction:

Graphs are a simple tool to demonstrate the relationship between the entities. The relation between these entities can reflect several patterns and can aid the process of decision making. Influential maximization is the process of choosing the set of seeds from a network so that the maximum number of other nodes in the network is influenced. This algorithm can be extensively used in social networks, marketing, and recommendation systems. This algorithm has practical and successful applications in finding the key influencers in the given networks. Here, I propose the implementation of this algorithm in the Amazon product network to determine the influential products in the graph. The basic centrality approaches are to be used in this implementation. They are degree centrality, closeness centrality, and betweenness centrality to find a subset of initial spreaders.[1] For the influential maximization, an independent cascade model will be implemented. This algorithm determines the nodes that have the most influence in the network based on the centrality measures.

Amazon is the global leader of the e-commerce platform. With over 310 million users and more than 12 million products, the marketing of the products is always challenging. Reducing this product number to half can significantly reduce the time and cost of the marketing campaign. Identifying the most influential product within its network structure based on the customer purchase is the goal of the project. This can hence significantly impact product recommendations and overall sales of the company. Since the products have edges based on user preferences, the influential product with the highest centrality can directly influence users. Here, using this algorithm Amazon might concentrate its marketing efforts on promoting items that have a high potential to influence other purchases to optimize overall sales.

Problem Definition:

Marketing campaigns are costly and time-consuming. Amazon has millions of products, and it is not convenient to do marketing campaigns on all the products. Determining the key product is always necessary and challenging from such a huge dataset. It is a tough task to select a subset of highly crucial products for marketing in a network. Here, the goal is to determine such influential products in the network of products. If a smartphone buyer purchases a bunch of accessories like headphones, phone cover, charger, etc. while buying a single product then the marketing can be skipped on the accessories and instead can be targeted to the key product only. There can be a huge number of such key products that are vital in purchasing many other products. So, an effective sales strategy can be explored by selecting such important products. The Amazon product network contains the products as nodes and edges between them determine the purchase history. For example, if a user buys a smartphone and phone cover there is an edge between these two nodes (products). With the greedy approach providing maximum influence in the graph, I am looking to develop a different algorithm that can provide similar results. The greedy approach is time-consuming and computationally requires high resources[1]. So, my goal will be to mitigate the computation time and resources.

Product recommendations can also be effective with these kinds of results. For example, on the webpage of smartphones, we can keep phone covers and headphones as recommended products. We can keep all the products having an edge with smartphones as there is a maximum probability of purchasing the recommended products. There is an edge between these products. Also, more profit margins can be leveraged on these silent products. The influence product maximization algorithm determines these types of possible seeds that can be most influential in the network.

Centrality measures are always the key approach in determining the important node in the network. With the aim of effective product recommendation, the shortlisted products can significantly influence the decisions that customers make about what to buy. Further, by deploying the influence maximization techniques we can promote these influential products to maximize their impact on customer purchase behavior. Finding special nodes that have strong connections to the other nodes in the networks can be a good approach.

Algorithm:

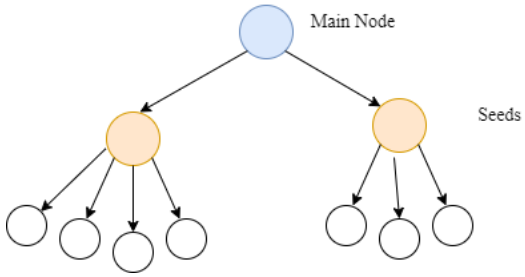
The influential maximization algorithm for a graph G (V, E) can be defined as the process of selecting the k seed nodes such that the seed set can maximize the influence over the network.

Thus:

IM (G, k) = max σ (e, G)

Here, σ refers to the information spread function.

S refers to the seed set. Thus, σ(S) gives the final propagation impact over the network.



Here is the algorithm of the baseline greedy approach:

- An empty seed set is initialized, and, on each iteration, the most influential node is found and added to the seed set.
- The most influential node needs the inner loop which iterates over candidate nodes and evaluates them by estimating their marginal gain.
- Finally, the seed set which has the highest influence is achieved. Their influence is propagated by Monte-Carlo simulation [4].

Algorithm: Greedy

Input: G (V, E) , k (number of seeds)

Process:

- Initialize S and R (number of iterations)
- For each vertex:
    - Cascade (v)
    - S = S ∪ argmax (s)
- Output S

Here s refers to the temporary seed set. Here for each seed set their influence is checked and the ultimate seed set with highest influence is selected as result.

The greedy approach is highly accurate but is less scalable. Time complexity is high and cannot be taken for any simple implementation. Instead of attempting to increase the running time of the greedy algorithms here I have implemented a new approach based on centrality measures to increase their influence spread which can be a promising solution to the influence maximization problem for large-scale social networks.

Considering the problems with the greedy approach here, I offer an enhanced version of the greedy method that ensures a relatively closer influence spread to the greedy approach. With low computation, I have implemented the influence maximization with close performance on propagation with the greedy.

INM algorithm:

Input 1: Graph G(V,E) , k (top nodes for centrality)

Process:

- Selection of centrality measures:
    - Calculation of degree, betweenness, and closeness centrality.
    - Selected nodes= 10% of total nodes (desc(centrality())
    - Seed set (S)= U(nodes from all the centrality measures)

Output: Seed set S

Input 2: Seed set S

Process:

- Independent cascade model on the seed nodes. Status of nodes: Active, activated and inactive.
- Taking the threshold as 0.8. Propagation of an edge will be successfully cascaded with the probability 0.8/indegree.
- On each cascade, the active nodes are set to be activated and the cascading is done until no nodes can activate any other nodes.

Output: Set of influenced nodes

With the new influence node maximization algorithm, the k seeds are selected based on the degree measures. Here the set of degree, betweenness, and closeness generates top n influencers based on their respective calculation. Now, the union of all 3 sets is taken making a big set of seeds. These are our final seed nodes.

With the seed nodes the information propagation is done using the information cascade method.

Implementation:

The main goal of our algorithm is to accelerate the running time of our algorithm keeping the influence spread close enough to the greedy approach. The implementation was done in the Amazon product network.

This project is mainly implementation-based. The main aim is to select subsets of products from the Amazon dataset. Several calculations and algorithms are to be developed in the building of the project. The given concept can be summarized as identifying a subset of nodes which are the products with maximum co-purchasing behavior with other products. Now using appropriate marketing campaigns on the products can minimize the number of nodes that are to be taken to diffuse the information.

Calculating the influence spread given a seed set is a crucial component of the greedy approach, and it turns out to be a challenging issue. With the Monte-Carlo simulations of the influence cascade model, for enough time to acquire an accurate estimate of the influence spread, as opposed to searching for a precise algorithm[2]. As such, with current computation resources

over the whole Amazon dataset, it could take me days to identify a small seed set. Considering this problem, I am using a subset of the network as a comparison between the greedy model and the INM model.

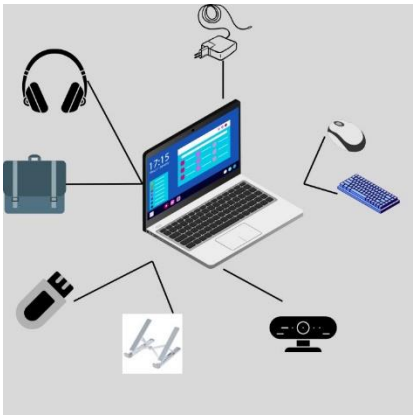
We can list the overall process as:

- **Start**
- **Input**: Amazon product co-purchasing network
- **Output**: Subset S with seeds and influence nodes.
- **Measures:**
    - Degree centrality
    - Betweenness Centrality
    - Closeness Centrality
- Select k nodes based on the given measures
- **Information Cascade Method**
- **End**

Detailed plan:

Dataset Collection: The dataset chosen consists of the user behavior on the product purchase. If on purchasing product i user frequently purchases product j then there is an edge between node i and node j. Data preprocessing is done in the dataset. The dataset is taken from the large database of Standford University. The dataset is based on customers behavior stating if the customer bought this item, then he/she also bought the other feature/product of the Amazon website. It is Amazon product co-purchasing network with 262111 nodes.

Consider the image shown. Here the nodes are the products and the edge between the nodes shows that, on the purchase of a laptop, the user has also purchased headphones and other accessories. The initial and the final node are both products.



Centrality based Measures:

An algorithm is developed to implement the centrality measures. Centrality measures can be taken as a metric to evaluate the importance of the node in a network. The seed set is generated based on the centrality measures. A certain percentage of the total nodes are selected to determine the top n nodes based on the centrality measures. The common nodes among them are the seed set.

The major three centrality measures are used: Here the measures used are:

- Degree centrality: It refers to the number of connections a node has with other nodes. In this case, this can be taken as the number of products purchased with other products. Higher degree shows the product is connected to many other products. Node with higher degree represents if the product is purchased several other products are also purchased.
- Betweenness centrality: It measures the number of shortest paths that pass through the node. Here, a product can be taken as a bridge between the other products. This measure is helpful in determining the key junction of the network. Products with higher betweenness can be considered a common product purchased frequently with other products. They are the influential products connecting various categories of the network.
- Closeness centrality: It measures the inverse of the shortest distance between the node to all other nodes. Here, the product's centrality is based on the average shortest distance between the nodes. Here, if the product has higher closeness centrality, then we can conclude the product is closely related in terms of co-purchase ties to many other products[3].

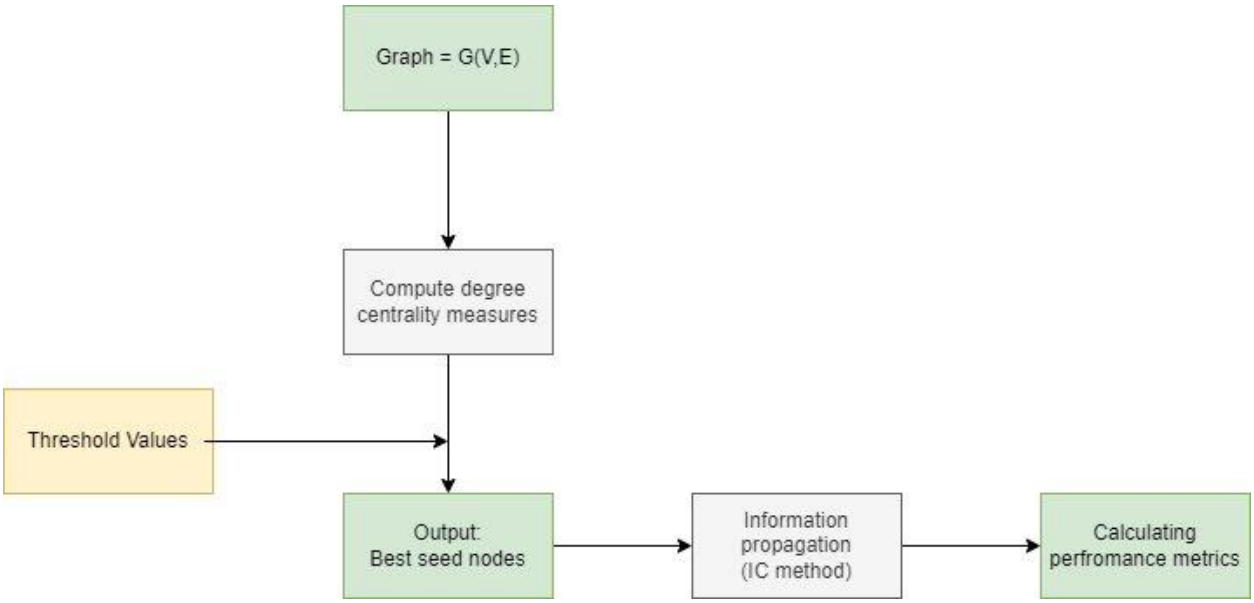
Based on these centrality measures, the number of potential seeds in the network are selected. The selected products are the highly influential nodes.

Information Propagation Measure:

For the information propagation, I have used the independent cascade mode. The independent cascade method starts with the initial set of seeds and propagates to the other nodes activating the inactive ones. Independent cascade is one of the diffusion models for information spread through discrete steps. The propagation is done based on probability. The probability is calculated for each edge and based on that propagation from the source node to the target node along that edge is determined. Once a node is influenced and gets the information, the node can influence its neighbors and so on. Here in this project, the threshold probability is taken as 0.8. The seeds of the network are activated at first and it triggers the activation of the neighboring nodes at discrete time steps. The independent cascade method is implemented as a part of influential maximization. This method allows the diffusion of the information in the network. The process continues until all nodes are activated[1].

Independent cascade model on the seed nodes. Status of nodes:

- Active
- Activated and
- Inactive.



The overall flow of the influence node maximization model can be seen in the figure above. Here, the workflow starts using graph G. First the centrality measures are determined, further using threshold values the seed set is determined. Finally, information propagation is done, and the performance metrics are calculated.

The performance metrices used are:

- Influence spread
- Running Time
- Memory Footprints.

The comparison of performance is done on the proposed algorithm with various node count and the running time comparison is done with the greedy algorithm.

Testing and Documentation: Overall testing of the approach is done. Various data and threshold values are considered while testing. The model is tested on different node counts and separate results are obtained. Further, documentation work is done.

All the code of the project will be done in Python programming language. Necessary libraries like networkx and others are used as a part of the implementation. Jupyter Notebook and Google Colab are used as the implementation platform. Code is available on GitHub with link:

https://github.com/saimandahal/InfluenceNodeMaximization

The dataset used is available on the link:

https://snap.stanford.edu/data/amazon0302.html

Results:

The plot below shows the amazon product network on a subset of nodes. Here:

Total nodes: 814

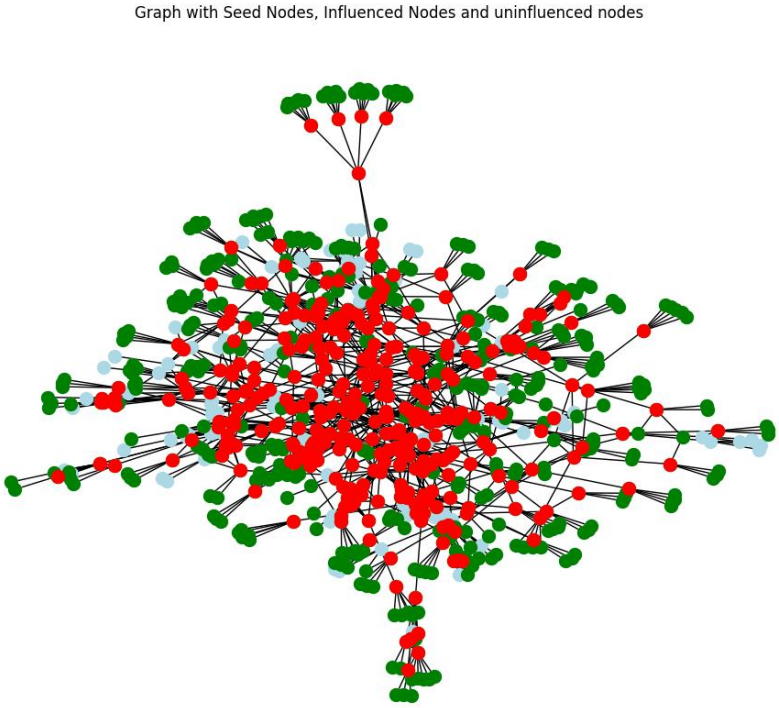
Total edges: 1665

Seed nodes (red color nodes) = 334

Influence nodes (green color nodes) = 709

Ratio of seed: total = 0.41

Ratio of influence: total = 0.87



Similarly, a larger graph with more complex structure is shown below:

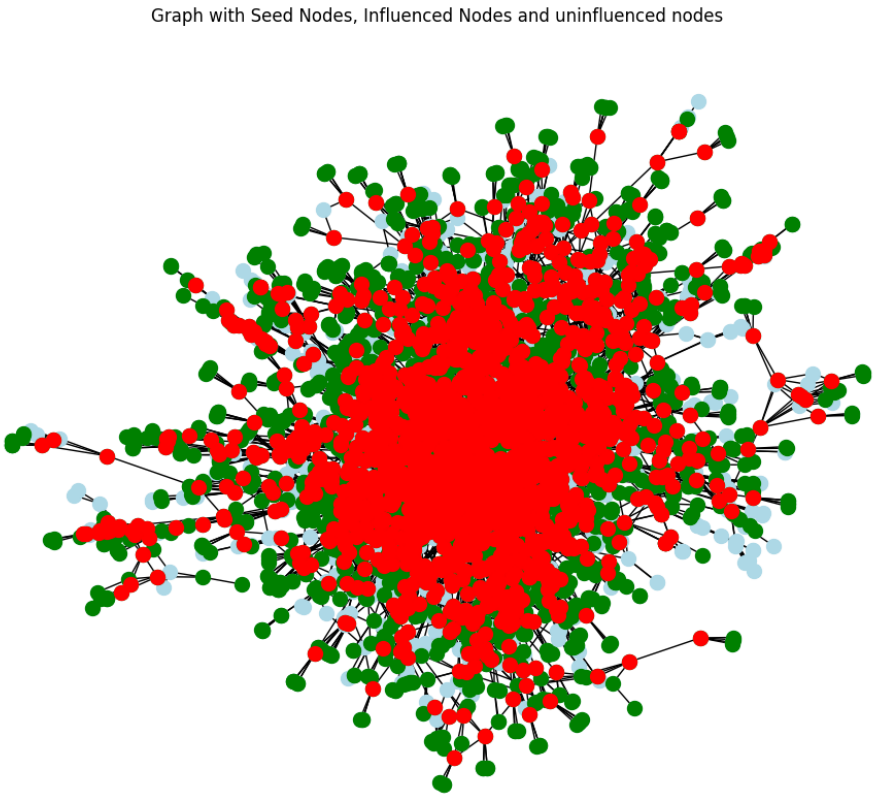
Total nodes: 11164

Seed nodes (red color nodes) = 4501

Influence nodes (green color nodes) = 8993

Ratio of seed: total = 0.4

Ratio of influence: total = 0.8



The graph shows the overall result of our model. The red color nodes are the seed nodes, while the green color nodes are the influenced nodes. More than 80% of the nodes are influenced in both cases. The model is scalable and covers the maximum possible nodes in the network.

Further, the marketing information spread is fast and scalable. Here, the results properly showed that I am successful in finding an appropriate candidate subset of nodes for spreading the marketing information of the Amazon company. Also, the performance criteria are compared. The running time and memory footprints are less comparable to other IM algorithms. Thus, using this algorithm the marketing information can be distributed effectively over fewer products

Performance comparison:

Total Nodes	Centrality nodes	Seed Nodes	Influenced Nodes	Time taken (s)	Memory Used (MB)
11164	500	**1013**	**3525**	1196	126
11164	1000	**1978**	**5100**	1160	126
11164	2500	**4501**	**8993**	1255	126
11164	4000	**6248**	**10491**	1264	127

Here, the term centrality nodes are the nodes that are taken for the centrality measures. The first row 500 resembles the top 500 nodes of 3 sets: degree, betweenness and closeness are taken and the common from these 3 sets of 500 nodes are taken as the seed nodes.

We can see that the more seed nodes the more influence propagation. The time and memory usage are quite similar while varying the seed node count. Thus, the tradeoff can be visible from the seed set point of view. More marketing campaigns on the larger seed set implies more information spread over the network. However, the larger seed set implies more marketing costs. Hence, the appropriate number of seed sets can be considered based on the resources present.

Comparison between greedy approach and the INM approach:

Here, the whole amazon network has over 292000 nodes, so using a greedy approach in this was not possible. So, I have taken a subset of the graph. Since random sampling can create some false results with much outlier information, I have done the sample of the top 300 nodes. Here, the nodes are taken serially, starting from node 0 to node 300.

Total nodes = 393

Model	Seed Nodes	Influenced Nodes	Ratio (influence: total)	Time taken (s)
INM	133	**342**	0.88	**0.44**
Greedy	80	**393**	**1**	2202

We can see that the influence of 100% greedy approach has the best influence with the 80 seed sets, but the time taken to complete this operation was 2202 seconds (about 36 and a half minutes). For a small network, the time taken is high. The performance from INM was 88%, which states that 88% of all the nodes of the network were influenced. However, the time taken to propagate the influence was a mere 0.44 seconds. So, we can confirm that our new model INM with relatively close performance to the greedy approach can accelerate the influence maximization algorithm.

The code was run in Google Colab platform under Google GPU.

The result above can be a good optimization algorithm for the greedy approach. The model is scalable and covers the maximum possible nodes in the network. The running time and memory

footprints are less comparable to other IM algorithms. Marketing information can be distributed effectively over fewer products.

In terms of Amazon company, this result can be effective in developing marketing strategies and can benefit the entire Amazon ecosystem. In a competitive e-commerce business, Amazon can maximize sales growth and optimize resource allocation because of the data gathered from influential maximization. Finding an appropriate candidate subset of nodes for spreading the marketing information of the Amazon company. Thus, the marketing information spread can be now fast and scalable.

Related Works:

Influence maximization is one of the prominent algorithms in social networks and many other networks. There are various related works in this area. The greedy-based influence maximization has a significant implementation in various areas like marketing, traffic, healthcare, and many more. This project expands on the previous research by improving the performance of the greedy method and presenting degree-based heuristics to balance the computational efficiency and influence spread. With the development of a new algorithm, INM can be a good approach to determining influencers in the network. It has a relatively close influence spread compared to the greedy approach. Microsoft's paper on influence maximization has also provided some measures to improve the greedy approach lowering the time taken in its seed selection steps[1]. Further, it explains the degree centrality based heuristics to determine the seed nodes. Similarly, another work is on the centrality-based influence blocking mechanisms. This paper has utilized the centrality measures to determine the seed nodes and used the diffusion model for information spread.[3]

Future Work:

An incremental model of seed set selection is my future work. For now, only one class of seed set is determined so a parent seed node can contain its child in the seed set as well. So, the concept of seed class is the future work. The main parent node can be considered as the parent and child seed node and so on.

Well, with the new INM model, information propagation is not enough, this can be an area of exploration. Reaching the information propagation close to the greedy approach is the ultimate solution. Still, the spread ratio is behind the greedy approach, so this can be the other future work.

More centrality measures need to be explored. Also, incremental sampling is to be explored as a part of sampling since it is computationally expensive while implementing the greedy approach in whole dataset.

Conclusion:

The influential node maximization algorithm has been implemented successfully. The implementation provides some valuable insights for effective marketing strategies. Identifying the most influential products over the product network can enhance product recommendation and optimize the resources. Effective implementation of the proposed methodology in the Amazon customer behavior dataset has been achieved. With successful implementation, marketing information can be distributed effectively over fewer products instead of reaching a large audience inside the network. Thus, my project has successfully implemented the influence maximization algorithm for finding an appropriate candidate subset of nodes for spreading the marketing information of the Amazon company. Three different centrality measures are used to determine the seeds and further, the diffusion model is used for the information propagation. Hence, the algorithm is properly implemented maintaining the influence spread guarantee.

Bibliography:

[1] Wei Chen, Yajun Wang ,Siyu Yang. Efficient Influence Maximization in Social Networks. https://www.microsoft.com/en-us/research/uploads/prod/2016/06/kdd09_influence-1.pdf.2009.

[2] Xinran He, Guojie Song, Wei Chen, Qingye Jiang. Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model.https://arxiv.org/pdf/1110.4723 . 2011

[3] Niloofar Arazkhani, Mohammad Reza Meybodi, Alireza Rezvanian. Influence Blocking Maximization in Social Network Using Centrality Measures. https://ieeexplore.ieee.org/document/8734920 2019.

[4] Yuxin Ye, Yunliang Chen, Wei Han. Influence maximization in social networks: Theories, methods and challenges.https://www.sciencedirect.com/science/article/pii/S2590005622000972 . 2022.

Appendix:

Code is available at: https://github.com/saimandahal/InfluenceNodeMaximization