Marketing Pool

Aarushi (2019csb1059)
Aditi Aggarwal (2019csb1063)
Gurkirpal Singh (2019csb1087)

November 23, 2020

Abstract

Our primary goal was to make the product viral in the market and minimize the money one has to spend on advertising his/her product to the desirable number of customers. For this we had to find a suitable ranking method to rank the population. We looked at several algorithms that have been designed to identify the most influential regulatory points within a network. However, all those methods did not address all the topological dimensions of a network which limits their applicability. To overcome this computational deficit,we used an algorithm termed Integrated Value of Influence (IVI), which integrates the most important and commonly used network centrality measures in an unbiased way. Next we had to minimize the cost of advertising assuming that the advertising fee is directly related to the popularity of a person. Here, instead of naively advertising to all popular people we devised a method to identify the most preferable person to advertise at a given point keeping the expenses in mind. In some cases we were able to achieve a cost difference of 20+% between advertising naively to popular people and the advertising method developed by us.

1 Introduction

The idea is that marketing doesn't have to be expensive, but it needs to be strategized smartly to provide big returns. The method is aimed to provide a possibility to use highly targeted marketing to reach exactly the desired amount of audience, minimizing the expenditure. Nodes with a simultaneously large number of connections and high spreading potential are the most influential or vital nodes in a network however they are also more expensive to hire for advertising a product. Hence it's always a tradeoff between the total advertising cost incurred and the reach of the product. The aim is to identify the popular nodes using the IVI method. IVI(integrated value of influence) is the method used to rank importance of nodes in a network. It is the synergistic product of the most important local (i.e., degree centrality and ClusterRank), semi-local (i.e., neighborhood connectivity and local H-index), and global (i.e., betweenness centrality and collective influence) centrality measures in a way that simultaneously removes positional biases. The total advertising expenses incurred are cut off by choosing the

most feasible person to advertise. The input to the algorithm is the dataset of people and their friendships, the desired value of reach and the threshold value of any individual node (threshold value is the ratio of acquaintances of a person adopting an idea).Every node is affected by its neighbors and by extending this property to whole graph, any product can get viral. We minimized the total nodes which were initially given some bribe such that after considering one's neighbors everyone adopts the technology.

1.1 Problem

The main problem which we addressed in our project is why most of the startups fail these days. The reason is that it does not reach the concerned audience. So how will it reach the concerned audience what we have to do? Do we have to spend more on the fancy advertisements? Do we have to do more initial investment for all these things? Or should we use superpowers(of what kind we will see)? What should we do?We will address all these problems.

1.2 Literature

An inspiration for this project was the class lectures corresponding to the topic " How to go viral ". We explored several research papers for finding a suitable method for ranking. We used a dataset as a potential friendship network to test our algorithm.
Our code is available at: [1] [2] [3].

1.3 New idea

We solved the problem by ranking the nodes on the basis of their importance in the network using IVI(Integrated value of influence) which is the synergistic product of hubness and spreading values of the node. We used the property that adoption of anything depends on what one's friends adopt. On the basis of this we find minimum people needed to be advertised so that everyone adopts our technology. We also have used an algorithm to reduce the costs of advertising by selecting specific people.

2 Method

2.1 Implementation details

2.1.1 Ranking Nodes

Most influential nodes can be identified by measurements of the centralities of a network, which themselves are calculated by analyzing the overall topology of the network. We tried to integrate Six most useful ones: Degree Centrality, ClusterRank,Collective Influence ,Neighborhood connectivity, Betweenness Centrality and Local-H Index.Each of these centrality measures captures a different topological dimension of the graph, including local, semi-local, and global topology.

Degree Centrality is the local centrality measure of a graph that is the number of ties a node has It is calculated as:

$$DC_i = \sum_{i \neq j} A_{ij}$$

where A is the representative of the adjacency matrix of the corresponding network and $A_{ij}=1$ if nodes i and j are connected ,else $A_{ij}=0$.

ClusterRank is a local ranking algorithm which takes into account not only the number of neighbors and the neighbors' influences, but also the clustering coefficient.Mathematically, the ClusterRank score si of node i is defined as:

$$s_i = f(c_i) \sum_{j \in} (k_j^{out}+1)$$

where the term f (c_i) accounts for the effect of i's local clustering and the term '+1' results from the contribution of j itself.
Here, $f(c_i) = 10^{-ci}$

Collective Influence is an adaptive algorithm which removes nodes progressively according to their current CI value, given by the following formula:

$$CI_\ell(i) = (k_i - 1) \sum_{j \in \partial B(i,\ell)} (k_j - 1), \qquad (1)$$

where *ki* is the degree of node *i*, $B(i, \ell)$ is the ball of radius ℓ centered on node *i*, and $\partial B(i, \ell)$ is the frontier of the ball, that is, the set of nodes at distance ℓ from *i* (the distance between two nodes is defined as the number of edges of the shortest path connecting them). At each step, the algorithm removes the node with the highest $CI\ell(i)$ value, and keeps doing so until the giant component is destroyed. A straightforward implementation of the algorithm consists in computing at each step $CI\ell(i)$ for each node *i*, and then removing the node with the largest $CI\ell$ value.

Neighborhood connectivity is average connectivity of neighbours of given vertex.

$$NC(x) = \sum_{k \in N(x)} |N(k)||N(x)|$$

Where N(x) is set of neighbours of vertex x

Betweenness Centrality is defined as the tendency of a node to be on the shortest path between nodes in a graph.If Smn is the number of shortest paths between nodes m and n,

and S_{mn}ð$_i$Þ is the number of shortest paths between nodes m and n that pass through node i,then the betweenness centrality of node i is

$$BC_i=\sum_{m\neq i;m\neq n;n\neq i}(S_{mn}(i)/S_{mn})$$

The LH-index method takes into account h-index values of the node itself and its neighbors, which is based on the idea that a node connecting to more influential nodes will also be influential. Mathematically, the H-index of node i is defined as:

$$H_i=\text{ֆ}(k_{i1};k_{i2};.;k_{iki})$$

where ֆ is an operator that calculates the H-index of node i based on the degree of its immediate neighbors, and k_i and k_i are the degrees of nodes i and j, respectively.
LH-index is defined mathematically as:

$$LH_i=H_i+\sum_{v\in N(i)}H_v$$

where N(i) is the set of neighbors of node i.

For integrating all these centrality measures, we recruited the Addition and multiplication functions, two mathematical operations of arithmetic.Using these, Spreading Score and Hubness Score were calculated for the nodes.
Spreading Score is the potential of vertices in spreading of information within a network and mathematically is defined as

$$Spreading\_score_i=(NC_i+CR_i)(BC_i+CI_i)$$

where NC_i,CR_i,BC_i and CI_i are range normalized neighborhood connectivity, ClusterRank, betweenness centrality, and collective influence of node i, respectively.
Hubness Score is the sovereignty of a vertex in its surrounding local territory.
Mathematically,

$$Hubness\_score_i=DC_i+LH_i$$

where DC_i and LH_i are range normalized degree centrality
and local H-index of node i, respectively.

IVI is the synergistic product of the most important local (i.e., degree centrality and ClusterRank), semi-local (i.e., neighborhood connectivity and local H-index), and global (i.e., betweenness centrality and collective influence) centrality measures in a way that simultaneously removes positional biases.

$$IVI_i=Hubness\_score_i * Spreading\_score_i$$

2.1.2* Minimizing the number of people to advertise

Every node either adopts a new idea or rejects it. Its decision is based on the fraction of its neighbours adopting the idea (also called the threshold value). To make everyone adopt our idea we ran the algorithm till everyone's threshold value was crossed and in the process we made several nodes accept the idea. There are 2 types of acceptances people make:

1. We advertise them and pay them to accept the idea
2. They are influenced by people around them and accept the idea

What we have to bother about is the first type of people and the second type would automatically accept it once they know enough people who already follow it.
Firstly we pay one node according to its ranking from the sorted ivi list and then mark all the second types of nodes which were influenced due to this node. We performed this recursively and found out the minimum number of nodes required to cover the graph and we did this same using various methods like pagerank, and various centrality measures.
We also plotted a graph of threshold vs minimum number of people to advertise to complete the graph using all centrality measures.

2.1.3* Minimizing the cost to advertise

Now, unlike in **2.1.2** we should be actually concerned about minimizing the total cost of advertising not the total number of people we advertise. **NOTE:** according to the function of fee distribution chosen by us the fee taken by each person to advertise our idea is directly proportional to his/her popularity. So here we can't directly select a person to advertise based on his popularity, we also have to consider the money he/she is demanding. So we choose the most effective node to be paid. Initially we consider the price of each node on the basis of their popularity in the network and after that we assign weight to each node.
For assigning weight to each node we sum the price of all the nodes which it can influence. This weight function changes after each node gets influenced. Handling this is a tough problem. After that we have to check which node we should pay in order to get maximum benefits and we solve this using a variation of the varying knapsack problem. We pay the node which has the highest ratio of weight vs price and this process is recursively repeated.
We also plotted a graph like how much people required to make the product viral to x% of the total population (where x is a user input).
And in this way we find out the minimum cost required in order to make the product successful in the market.

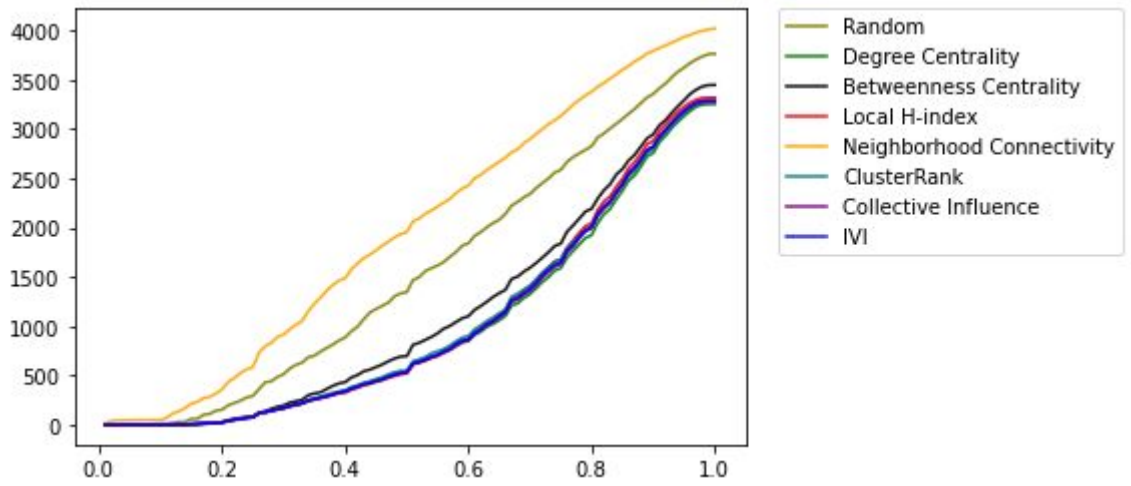
2.1.4 Comparing 2.1.2 and 2.1.3

We found the total costs incurred using **2.1.2** and using **2.1.3** and found their difference.
We then computed money saved and percentage of money saved.

We plotted a graph between input threshold and percentage of money saved at that threshold when advertising 100% of the population.

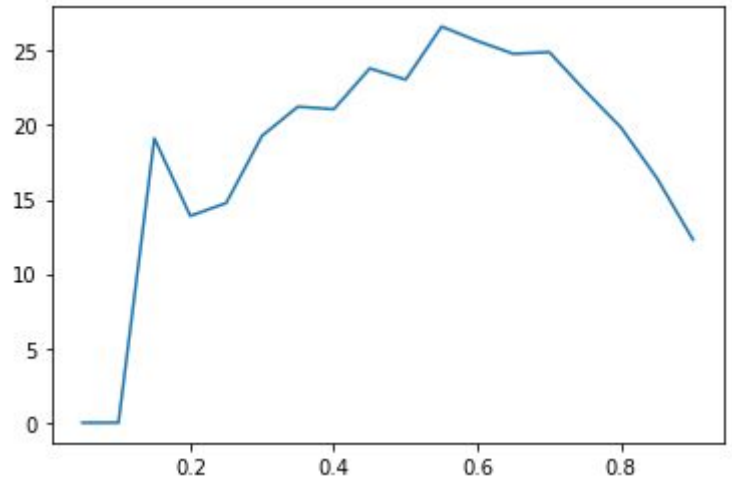
3 Results

3.1 Experiment findings



X-axis: Threshold Values
Y-axis: Number of persons required to advertise

In the above graph we plotted the threshold value on the x-axis and minimum number of people required to influence all the people on the Y-axis and then choose nodes according to degree centrality, betweenness centrality, IVI, random selection of nodes and observed this is almost the exponential graph.From the graph, we observe that **number of nodes required to be advertised are minimum when nodes are selected using IVI method**, hence concluding that IVI(which is a combined result of several centrality measures) is better than all other centrality measures used individually.



X-axis: Threshold Values
Y-axis: Money Saved Percentage

In the above graph we plotted the threshold values on the x-axis and on the Y-axis we plotted the money saved percentage against the naive approach.

```
You need to advertise atleast 335 people for 100% reach
You need to advertise atleast 335 people for 100 % reach
Money spent in a naive manner: 1297120
Money spent after optimising the approach: 1023865
money saved: 273255
profit %  21.06628530899223

In [37]:
```

Sample Output

This is a snapshot of our findings for threshold value 0.4 and 100% reach. In this 335 people are required to be advertised for 100% reach of the product. If those persons are selected by naive manner, money required is 1297120 units while by optimised approach, money required is 1023865 units.
Money saved : 273255 units
Percentage of money saved: 21%

Various methods:-	IVI	ClusterRank	Degree centrality	Collective influence	Betweenness centrality	Local h-index	Pagerank
1st	1912	107	107	107	107	1912	3437
2nd	107	1912	1684	1912	1684	107	107
3rd	2347	1684	1912	1684	3437	2347	1684
4th	2543	2543	3437	2347	1912	2543	0
5th	1684	2347	0	2543	1085	2366	1912
6th	2266	1941	2543	2266	0	2233	348
7th	1985	1888	2347	1985	698	1985	686
8th	2233	1800	1888	2233	567	2142	3980
9th	2142	1985	1800	2142	58	2206	414
10th	1941	2266	1663	2206	563	2218	698

The above table lists the 10 most influential nodes according to various methods according to our code's output. Name of every method is specified in the top most row.

Threshold values	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Total nodes	1	24	169	335	531	867	1378	2006	2822	3291

This table gives minimum nodes required for 100% reach at a particular threshold value to influence all the nodes in a network.

3.2 Interpretation of findings

We ranked the nodes according to their importance in the network. And the most surprising thing is that it is completely unbiased even better than the pagerank.
We also observed that the neighbors of the most influential node are also very important. We plotted the graph of threshold versus minimum number of people required in order to make some technology viral and it is very surprising that the **graph comes out to be exponential**. We chose the nodes on the basis of various methods and found that the number of nodes needed to be advertised obtained using the ivi method were very less as compared to other methods like pagerank,degree centrality and other centrality measures and random node selection methods.
Later we found the minimum cost required for the advertisement, and we observed that the result which we got was 20% more profitable than the result which we calculated from the naive approach.

4 Conclusion

This seems like a fairly reasonable method for computing the people who should be advertised since we have been able to make a significant reduction in advertising costs. However, further research could be helpful in making the algorithm even more effective and less time consuming.
We all learnt a lot from this project. How to work together as a team! We spent too much time in some areas.Looking back we have, at times spent too long trying to find the perfect solutions to problems. The knowledge which we gained from this project can't be obtained theoretically. We also enjoyed a lot while doing this project.
This project was a major learning experience in terms of graphs, how just changing one node everything gets changed and how everyone is so connected in the network and various other influencing techniques.

4.1 Team Work

One of the most difficult tasks was to develop a basic idea about what was to be done. We held several meetings regarding the same. We rejected several ideas in the process and finally one of us came up with this idea. We looked up on the internet but were unable to find much information about the already developed methods/ algorithms about the same. This did not hinder our belief in this idea and we thought of discussing it and developing it gradually on our own.
In terms of contribution no one did more and no one did less. Every second of each person, contributed to this project was valuable and fruitful.
We effectively managed to do most of the work collectively but sometimes we were able to distribute the work specially while plotting graphs which were time consuming.
We believe that the project was more thinking, exploring and discussing than coding.

One thing we feel that we could have optimised was the efficient use of time. We had planned to implement the algorithm on our class dataset but the lack of proper time management and planning resulted in a failure here.
Overall we are very satisfied with what we have done but we don't want to stop it here and do aim to explore other related aspects in future.

4 References

1.https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077455

2.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6037365/

3.https://www.researchgate.net/deref/http%3A%2F%2Frefhub.elsevier.com%2FS2666-3899%2820%2930063-5%2Fsref11

4.https://www.researchgate.net/publication/342374701_Integrated_Value_of_Influence_An_Integrative_Method_for_the_Identification_of_the_Most_Influential_Nodes_within_Networks

5. https://www.nature.com/articles/nature14604

6. https://www.sciencedirect.com/science/article/abs/pii/037887339190017N

7. (PDF) Calling Dunbar's Numbers

8.https://www.nature.com/articles/ncomms10168#:~:text=The%20coreness%20of%20a%20node,k%E2%89%A41%20to%20appear.