© 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Target Marketing on Social Media

Mayank Dhingra, Mohit Gahlot

Computer Science, The NorthCap University, Gurugram, Haryana, India

ABSTRACT

It is very common for us to see advertisements on social media, but one cannot target each and every person on such a platform for the marketing of a product, due to obvious reasons increasing number of people on social networks and budget constraints. Hence reaching out to the right audience for a successful marketing campaign on social media can be a task. This paper discusses how SNA (social network analysis) can be used to reach out to maximum target audience within a fixed budget for such promotional activities. The approach discussed, majorly uses basic knowledge of network analysis and graph theory.

Keywords : Centrality Measures, Clustering, Digital Marketing, Social Network Analysis.

I. INTRODUCTION

Social network analysis (SNA) is a technique to study and evaluate how people, organizations or entities interact with each other within a larger system, how information flows within a network of such entities, and to determine which entities are important or play a key role in spreading information within a network. It is not just a methodology but provides a perspective on how a society functions, at large. This overview does not only focus on individual entities within a society and their attributes, or on broad social structures, it focuses on relationships between individuals or groups.

Such an analysis uses graph theory to represent the social networks/connections – people or entities are taken to be the nodes and the weights between nodes represent the strength of their connection, such graphs can be directed or undirected depending upon the social media platform.

For example – connections on Facebook are undirected (if you follow the friend request of a person, both of you can follow each other's timeline) whereas those on Instagram are directed. Adjacency lists and Matrices are generally used to represent graphs. Network theory helps to analyze such graphs, study the structural connections and identify influential entities using various network metrics, as discussed later.

Combined with other visualization techniques and analytical tools, the knowledge of graph theory and network theory, forms the basis of social network analysis. Social networks can be categorized in three broad classes. First is Egocentric networks which are basically networks obtained by focusing on connections of a single node/entity, it means that egocentric networks are created for individual nodes that include its connections with its direct connections as well as their indirect connections (up-to one or two levels). Second is Socio-centric networks which have clear boundaries. They mainly extract connections of a well-defined society or community that has certain features in common. Examples of such type of a network are students in a classroom or workers working in an organization. Open system networks are not closed, they do not have clear boundary lines which makes this type of network difficult to study. One example can be the

entire facebook community.Depending upon the application, the type of network to be studied is determined. Social network analysis finds its applications in various fields like Business organizations use social network analysis to analyze and improve communication flow within their organization, or with their partners, dealers and customers. It helps them fasten the business processes and expand their business. Security agencies and many governments analyze Social Networks to identify terrorist and criminal networks from traces of communication collected; and then figure out the exact origin of such activities and also the identity the leaders. Social Network Sites like Facebook use a few basic elements of SNA and many more. The paper is divided into 6 sections. First section is introduction toTarget Marketing on Social Media. Section 2 deals with explaining about how the data used in the paper was prepared. Section 3 deals with explaining various clustering techniques used in the paper for making clusters. In section 4, centrality measures used to find important nodes is explained. Section 5 & 6 explains approaches taken in the paper.

II. DATA PREPARATION

Information about relationships between various people on social media is not available, we created a dataset describing relationships on Facebook. It had 15038 rows, each describing a relationship between two people. The various attributes were the node IDs of the two people, the number of messages they had sent to each other in the past year, number of mutual friends, number of groups in common, number of times they've tagged each other on posts, and the number of same events attended by them.

All these can be used to judge the friend circles on social media, whether or not, two people are actually closely related or not.

A. Logic Behind Data Generation

Assuming total count of people (considered as target audience) to be 500. A friendship network of 50 people is studied. Logic used is:

- The nodes are numbered from 151 to 200, so they can be friends with any node from 1 to 500.
- The dataset is created using random probabilities for the initial phase.

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

- The number of friends of each node is randomly set within a range of 40- 150, probability of having 40-70 friends is set to 3/5 and that of having 71-150 friends is set to 2/5.
- 60% of the number of friends of a node are kept within a range of node numbers from 151 to 230 and the rest can be from 1 - 500

III. CLUSTERING IN SOCIAL NETWORK

Discovering closely knit clusters on social networks is done mainly by using clustering techniques for network theory.

Clustering is done either on the basis of network structure or the similarity between entities (i.e. similarity between features of entities). When groups are made on the structure of the networks, concepts of network theory are used and algorithms such as newmann girvan algorithm which uses basic betweenness of a node for clustering them can be used.

When clustering on the basis of similarity, commonly used clustering algorithms such as KNN can also be used. More the number of features similar for two entities, lesser is the dissimilarity value and they are more probable to end up in the same group. 2 types of clustering algorithms used are:

A. **Louvain Clustering**

This method for community detection is used to extract communities from large networks created by Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. It is one of the most effective method to identify communities in large network. This method has a time complexity of O(n log n) and greedy optimization method is used to optimize the Modularity of the network.

1. Modularity Optimization

This method of community detection is used to optimize Modularity as the algorithm progresses. It measures the density of the edges inside communities and outside communities, and value is between -1 and 1.The best possible groups of nodes are made by optimizing the value. As it is impractical to iterate over all nodes into the group so the Heuristic algorithm is used for this. Louvain Method of community detection is a recursive process in which first small communities are detected by finding optimizing modularity locally and then each community is categorized as 1 node and the first step is repeated.

2. Algorithm

The value to be optimized is modularity which is defined as a value between -1 and 1 that estimates the density of links inside the communities compared to links between the communities. For a weighted graph, modularity is defined as:

Here, A_{ij} is an adjacency matrix which represents the edge weight between the nodes i and j;

ki and kj are the sum of weights of the edges connected to nodes i and j, respectively;

m is the total weight of the edges in the network;

ci is community of node i and cj is the communities of node j;

(ci,cj) is a Kronecker deltadelta function and its value is 1 when when 2 nodes are assigned to same community and 0 when they are from different community.

3. This method consists of 2 phases.

The First phase is Community Reassignment in which we iterate through each node of the network and find the change in the modularity if we place that node in its neighboring community. If the change is positive then we place that node in the community with maximum positive modularity change else we keep that node in the same community. We repeat this process with all the nodes of the community.

The Second phase is Coarse Graining in which we aggregate nodes of the same community that were discovered in the Reassignment phase. A new network is build in which the nodes are these communities. The edge weight between the 2 nodes in the new network is the sum of edge weights between integral nodes of each community.

These steps are repeated iteratively until network reaches the maximum modularity.

B. **Markov Clustering**

The Markov Clustering Algorithm, is a fast and scalable unsupervised clustering algorithm that is used for clustering nodes on the basis of the network structure, of the connection graphs.

It is based on the concept of random walks to determine groups.

Any clustering algorithm is said to perform well if the intra cluster distance is least and the inter cluster distance is huge. This is the main motivation behind this clustering method. (1

Hence in a graph it will be correct to presume that there will be dense links within a cluster, and lesser links between clusters. Which means beginning at a node, and then randomly traveling to a connected node it is more likely to stay within a cluster (reach a node within the same cluster) than travel between clusters (reach a node from a different cluster).

So by doing random walks upon the graph, it is feasible to discover where the flow tends to converge,

and therefore, where clusters are random walks on a graph are calculated using "Markov Chains".

Edge Weight has the following properties:

- Similarity or likeness between two nodes
- Considered as the bandwidth or connectivity.
- If an edge has more weight than the other, more flow will be flown over that edge.
- The total flow is proportional to the edge weight.

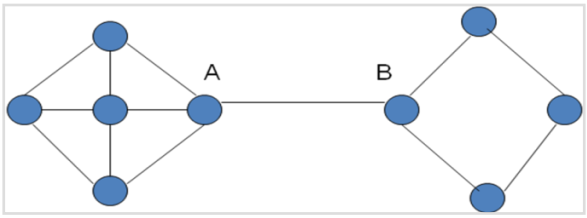

Figure 1. Cluster Identification using MCL

By visual examination the 2 clusters in the above graph can be clearly identified.

Considering the above network for identification of clusters, if one starts from the leftmost node in the network the probability of moving to one of its connected nodes is 1/3. Similar is the case for the next directly connected nodes. When the flow reaches the border point, it is likely to return back, than cross the border.

What that means is, on reaching Pt. A, there are 4 possibilities. Out of those three of the paths lead back to the cluster while the last leads to the other cluster and so it has a possibility of 1/4.

So that is broadly how the simulation of network flow helps in extracting closely knit clusters.

IV. CENTRALITY MEASURES

In graph theory and network analysis, indicators of centrality identify the most important vertices within a graph. Applications include identifying the most influential person(s) in a social network.

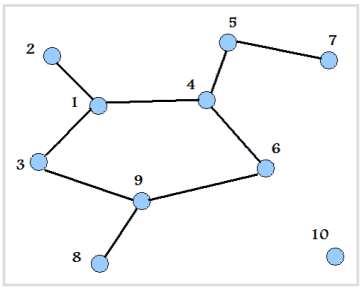

Figure 2. Network Graph

A. Degree

- It is the simplest indicator of how well connected a node is within a network.
- The degree of a node is the total number of edges incident on it.
- For a directed graph, the degree is the sum of in-degree and the out-degree edges
- For undirected graphs, total degree is simply the count of nodes it is directly connected to
- In the graph above, node 1 is connected to nodes 2, 3 and 5 and it is an undirected graph, so it has a degree of 3(number of direct connections).
- Node #10 isn't connected to any other node, so it has a degree of 0.

B. Density

- The density of a graph is the number of existing edges in a graph divided by the number of maximum possible edges for the given number of nodes.
- A graph with higher density is said to be more strongly connected
- The more dense a network is, the better it resists link failures. In simple words, in a dense network the probability of finding a route from one node to another on an average is higher.
- In the above graph total no. of possible edges (for a 20 node graph):

$$[20 * (20 - 1)] / 2 = 380 / 2 = 190$$

But the graph has only 76 edges. Therefore, the density is 76 / 190 = 0.4

C. **Betweenness Centrality**

- As the name suggests, this centrality measure deals with the extent to which a node falls 'in between'
- Estimates the degree to which a particular node lies on the shortest paths (from one node to any other node in that graph).
- A node with high betweenness value implies that it is very central to the communication flow in a network. So if some information is given to that node, it is very likely to be spread across most of the network.

III. APPROACH

In layman terms, the approach can be explained using a very simple example of a student classroom at a school. Consider a teacher wishes to convey an important information to the students of a class on an urgent basis and cannot meet them due to some reason. What she would instinctively do is pass on the information to a few students, and the selection of those students is such that each of them along with their friend circles, covers up the entire class. So the only difference in our approach is that the number of people are very huge and groups do not explicitly exist as such.

Initially, the dataset is checked for any records with missing node Ids, such records, if found are simply removed from the dataset. If any other attribute has a missing value, it is set to zero. The entire record is not deleted because each row is a friendship link, and we cannot ignore such links, doing so might affect the final output.

All features describing the friendship level, are then normalized. Normalization here is necessary because we want to model the friendship level considering all the features are equal i.e they contribute equally to the friendship level/ factor, initially. The friendship level/factor is then calculated using a simple linear formula using the normalized features. These numbers are used as the weights for the graph that is formed for the social network we are using. Clustering is then done on the network graph to obtain clusters. To further identify crucial nodes, importance of a node is calculated as follows :

Importance = centrality measure of node * density of cluster

Nodes are then arranged in the descending order of their importance value.

The fixed budget available for the promotion activity is divided by the cost of targeting a single node, to get the top nodes from the above ordered list, to be targeted.

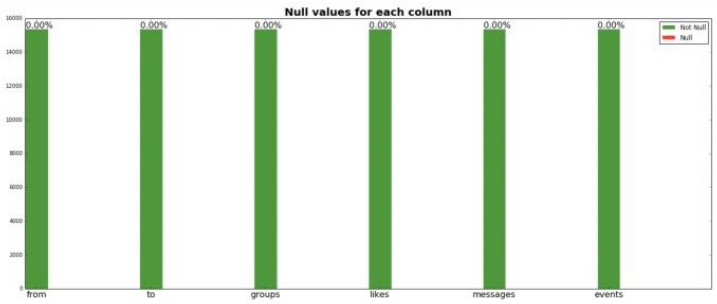

Figure 3. Normalized Data

IV. OUTPUT

For networking operations and functions NetworkX is used and for visualisation of data Matplotlib package are used.[10]

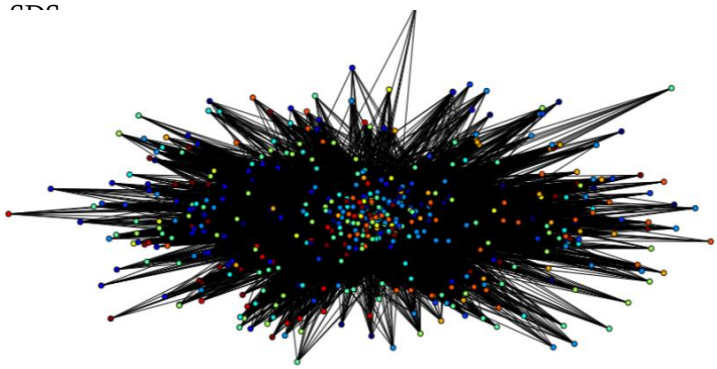

Figure 4. Identified clusters in the community

Figure 5. Individual Cluster 1

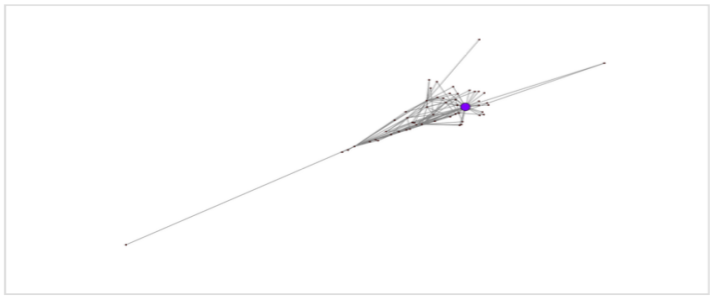

Figure 6. Individual Cluster 2

Figure 7. Individual Cluseter 3

```
247 298 403 387 372
31 303 60 210 363
182 59 172 461 128
388 199 41 107 152
254 72 337 334 2
499 487 472 68 467
190 222 65 96 287
282 82 5 223 415
206 244
6 2 8 2 6
7 8 7 9 8
9 7 2 6 5
5 9 7 7 2
2 4 4 5 7
8 1 2 4 5
5 8 9 2 5
5 6 7 7 8
9 9
target audience is
541
```

Figure 8. Top Nodes Identified in a Cluster

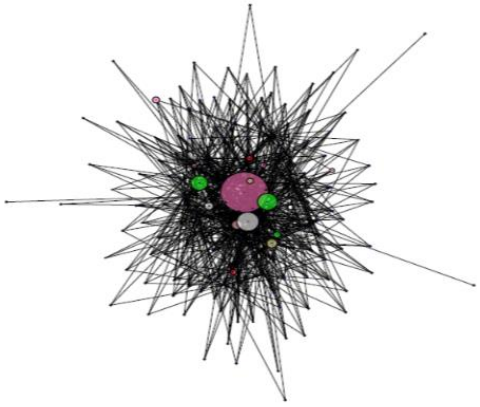

Figure 9. Target node IDs with their cluster

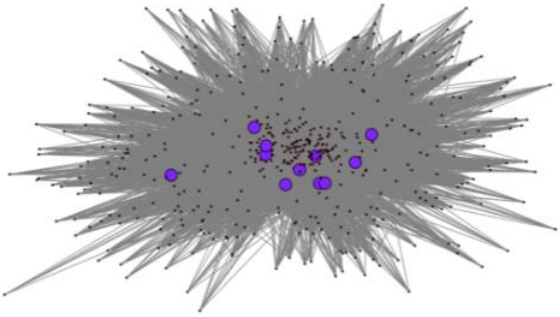

Figure 10. Most significant nodes identified in each cluster

V. CONCLUSION

The target audience for a budget of ₹ 10,000 is as shown in Fig 8. So, the actual number of people that could be reached with a budget ₹ 10,000 are 42, but with this approach we can reach 541 people with this budget.

VI. REFERENCES

[1]. Nicolas Dugue, Anthony Perez. Directed Louvain : maximizing modularity in directed networks.

[2]. Centrality and network flow, Stephen P. Borgatti, Department of Organization Studies,

Boston College, Carroll School of Management, Chestnut Hill, MA 02467, USA

[3]. Ching-Yung Lin directed Social Network Analysis in Enterprise, IBM T. J. Watson Research Center, Hawthorne, NY, USA

[4]. Dhanielly P. R. de Lima, José Francisco de M. Netto and Vitor Bremgartner directed Applying Social Network Analysis in a course supported by a LMS, Institute of Computing, Federal University of Amazonas (UFAM), Manaus, Brazil

[5]. Evelien Otte, Ronald Rousseau, Social network analysis: a powerful strategy, also for the information sciences, December 2002, SAGE journals

[6]. M. E. J. Newman, Modularity and community structure in networks, June 2006, journals of PNAS

[7]. Scott Emmons, Stephens Kobourov, Mike Gallant, Katy Borner, Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale, July 2016, PLOS journals

[8]. O. Green, R.McColl, D.A Bader, A Fast Algorithm for Streaming Betweenness Centrality, September 2012, IEEE

[9]. Tian Wang, Hamid Krim, Yannis Viniotis, A Generalized Markov Graph Model: Application to Social Network Analysis, February 2013, IEEE

[10]. Yedhu Sastri, Kuttyamma A.J, NetworkX and Matplotlib an Analysis, August 2013, International Journal of Scientific & Engineering Research

[11]. Andrea Landherr, Bettina Friedl, Julia Heidemann, A Critical Review of Centrality Measures in Social Networks, Semantic Scholars journal

[12]. Nina Mishra1,4 , Robert Schreiber2 , Isabelle Stanton, Clustering Social Networks, Stanford

[13]. Nicolas Dugu ́e, Anthony Perez, Directed Louvain : maximizing modularity in directed networks, hal archives ouvertes, 2015