Social Network Analysis: WOM Marketing in Rural India

ASMITA* and SIMRAN*, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), India

Word of mouth marketing in rural India setting is explored in this paper. The graph of the social network in the village is obtained by asking the locals to identify the leaders in their village, thus the view is able to capture the use case of word of mouth marketing. This work tries different methodology to identify the seed nodes for information spreading in the network and evaluates them through the general threshold model of information spreading and see the number of informed nodes in the end.

Additional Key Words and Phrases: influential nodes,information spreading,village graph

ACM Reference Format:
Asmita and Simran. 2022. Social Network Analysis: WOM Marketing in Rural India. 1, 1 (May 2022), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM DEFINITION (MOTIVATION AND INTRODUCTION)

1.0.1 Motivation. Identifying the seed nodes who are most effective for information spread is key to ensure maximum diffusion of information in a social network. Once key nodes are identified, they can be informed and thus be helpful for informing the others in the graph. This can be especially helpful in rural settings, where there is lesser penetration of technology and ease of accessibility to knowledge. The seed nodes thus identified can be helpful in propagating information regarding important policy decisions, information of natural calamity, information for educative purposes, etc. The information that is seeded to individuals that are "highly central", diffuses faster.

1.0.2 Introduction. The work explores different methods of finding the seed nodes for information diffusion in the social network graph. The information diffusion is modelled on General threshold model. The different approaches to finding the seed nodes are evaluated by their capacity to spread information in the general threshold model. Initially traditional methods of degree centrality, page rank etc. are explored in order to identify the seed nodes. Finally community detection is done on latent representation of the graph obtained by "Deep walk" in order to obtain the best set of seed nodes.

2 LITERATURE REVIEW

[Banerjee et al. 2019] addressed two of the following questions in the paper. 1) Can highly central or influential nodes be identified, when we do not have the exact network data, but only via asking individuals to identify the most influential members of their community. 2) Are the members identified via the above method more effective in transmitting information as compared to randomly chosen individuals or even highly respected individuals in the community. In order to identify the central nodes in the community, a metric "diffusion centrality" is proposed which ranks the individuals in the community by counting how many times someone hears about the individual in the community. In order to check the effectiveness of the individuals identified by the above metric, two studies were conducted. 1) Study 1: Investigates the effectiveness of the central individuals in the promotion of a cell phone brand. The key finding was that gossip individuals (individuals identified by the above metric) are much more effective than randomly selected nodes at promotion. 2) Study 2: Investigates the effectiveness of incentivizing central individuals for the success of a vaccination campaign. Four kinds of seeds were evaluated in this case; 1) Random seed, 2) Gossip Seed, 3) Trusted Seed, 4) Trusted Gossip seeds. The key finding indicated that the gossip nodes performed best.

[Chen et al. 2020] tries to identify influential spreaders via a node embedding based approach.The contribution of the paper is two fold: 1) learns node representations and partitions target network into clusters based on the similarity between nodes in the embedding space, i.e implements a modified deep walk ; 2) Applies K- means to identify clusters in the node embedding space and calculates the cores within the clusters and sets them as effective spreaders. It then evaluates its method in the SIR model and finds their proposed models to exceeding the traditional baselines.

[Jiang et al. 2014] tries to resolve some of the problems that exits with the traditional machine learning models that are used to analyze the diffusion of information across social networks. Most of the existing works assumes that the training set is "statistically consistent" with that of the testing set. There are two major shortcomings in the existing approaches: 1) The results learned by the models is specific to only one particular network that it is trained on and might not generalize well with other networks. 2) The machine learning models often ignores the "actions and the decisions" of the users. In order to overcome the above shortcomings the paper proposes a game theoretic approach to do analysis of information diffusion . The spread of information in the network is heavily dependent upon a user's actions, whether to forward an information or not.The decision of an individual to forward an information or not depends on the user's interest on the piece of information and even on the actions of its neighbours. Such a dynamic behaviour of the individual can be modelled as an evolutionary game, where each user can be considered as a player and the player has two strategies:

[*] Both authors contributed equally to this research.

Authors' address: Asmita, asmita21115@iiitd.ac.in; Simran, simran21146@iiitd.ac.in, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), Okhla Industrial Estate, Phase III, New Delhi, India, 110020.

XXXX-XXXX/2022/5-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1) Forward the information, 2) not forward the information. Thus the players,strategies and payoff matrix in this problem was defined and the correspondence between the evolutionary game theory and information diffusion was highlighted.

[Wicaksono et al. 2021] proposes eight different hypothesis for information diffusion. Each hypothesis is with respect to affect of a factor on consumers' use of an information source. The eight factors are: Age, Gender, Education, Mobile phone ownership, Direct social interaction, Social media ownership, Distance, and Nationality. It used binary logistic model to come to conclusion whether to accept the hypothesis or reject it. The factors that are really important are: Age, Education, Direct social interaction, and Nationality. The conclusion made was: "WOM is an important source and booster of information, even in years of digital communication.Consumers have trust in WoM information, but the reach of information is tighter than digital media. WoM is most commonly used for elderly peoples."

3 DATASET DESCRIPTION

"The diffusion of microfinance" dataset is being used for the implementation of this project. The dataset was collected by conducting a survey in rural Karnataka. A subset of individuals from 75 villages was asked the following questions:
1) who they borrow money from
2) give advice to
3) help with a decision
4) borrow kerosene or rice from
5) lend kerosene or rice to
6) lend money to
7) obtain medical advice from
8) engage socially with
9) are related to
10) go to temple with
11) invite to one's home
12) visit in another's home
This study takes graph from only two villages such that the nodes are individuals and an edge exits between the the ith and the jth individual if they performed any of the above activities together. The graph is directed and the adjacency matrix of the graph is symmetric. Figure 1 shows the snapshot of network for village 1, while Figure 2 shows the network for village 2.

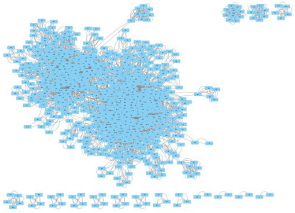

Fig. 1. Network Visualization For Village 1

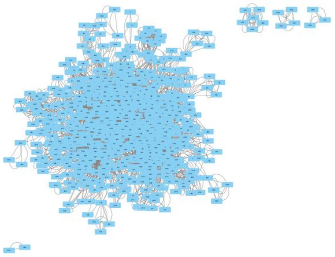

Fig. 2. Network Visualization For Village 2

4 BASELINES: OVERVIEW AND RESULTS

4.0.1 Degree Centrality. Higher the degree is, the more central or the more influential it is considered. It is intuitive to consider spreading the information to nodes with highest degree, as they would be able to spread the information to their immediate neighbors. It's basically choosing a person who deals with a lot of people.

4.0.2 Closeness Centrality. It is inverse of the sum of the shortest distance of a node to all other nodes in the graph. It is considered to be the seed via which information can be spread the fastest, as a higher closeness centrality score indicates a node closer to all the other nodes in the network.

4.0.3 Betweeness Centrality. Betweeness Centrality measures the extent to which a node falls between the paths to two other nodes. It thus acts as a connector node and can be considered useful in spreading information upto the widest distance within the network. A person with high betweeness centrality can spread a message as far as possible in the network.

4.0.4 PageRank. Page rank determines the number and the quality of the links passing through a node and thus ranks the nodes based on their importance.The more important a node is, there will be more number of links from other important nodes to it. Page rank gives a probability that a random walker in the graph will land on the node.

4.0.5 Katz Centrality. Katz centrality is used to measure the centrality of a node in directed network. Like page rank, it too takes into account the number of walks through a node while travelling from one node to other. It also takes into account the other nodes that connect via the current node along with its immediate neighbors.

The above metrics were used to rank nodes in the graph and the top ten from them was selected as the seed node. The diffusion capacity of the seed nodes were evaluated in the General threshold model of information diffusion and the results compared with our final solution can be found the table 3.

5 PROPOSED SOLUTION SKETCH

The information that is seeded to individuals that are "highly central", diffuses faster and wider.

Usually there are natural communities within village and people listen to most influential nodes within community. It is observed that influential nodes only form community, so it would be interesting to jointly find communities and influential nodes. The influential nodes form communites and influential nodes can be found by finding central nodes of community, so there is good chance that if both tasks are performed jointly, they would give better results.

For community detection, various algorithms like greedy modularity maximization, label propagation etc can be chosen but it was found that data mining methods give more promising results when node embeddings are fed as inputs. The node embeddings make sure that interconnectivity remains and thus it is safe to deploy data mining techniques on them. In order to better capture the net characteristics of the graph topology, Deep Walk was used to obtain the node embeddings. Deep Walk does random walk of specific length on the graph to thus obtain sequences of nodes visited on the graph. It then uses a skip gram model to learn the latent representation of these sequences. It is thus able to obtain a node embedding which the entire graph topology into account. Once the latent representation of 12 dimension is obtained from the graph, the model performs K-means clustering on the node embedding matrix. The number of clusters was set equal to number of influential people required. It would make it easier to return one most influential node from each cluster and directly report them. This approach is referred from [Chen et al. 2020] and helps to understand the community structure in the network. Though the idea of node embeddings and clustering on them wasn't used earlier. It is intuitive to think that the core node within each community would be the most influential in spreading the information within the community. PageRank is applied on each cluster to obtain the core node within the community.

General threshold Model was used in order to model the information spread across the network. The general threshold model takes the influence of the peers or the neighbors into account. Initially each node has a threshold value assigned as per uniform random distribution. The peers of the node will have positive influence i.e. will instigate the node to acquire the information and thus participate in further spreading, if the peer is an active node itself, or else if the peer is inactive it will add a negative influence to the node. The summation of the contribution from the peers is taken and if it is higher than the node's threshold, the node becomes active (participates in further spread of information) else remains inactive.

The effectiveness of the seed nodes was evaluated by setting them to "active" as they get informed and applying the general threshold model. The information spread in the general threshold is continued for max iterations of 10 and the percentage of active nodes (i.e informed nodes) are taken to be the metric of success of information spread.

The Table 1 shows the hyperparameters used in the model.

Table 1. HyperParameters used in model

Hyperparameter	**Values**
Number of Clusters (K in K-means)	10
Walk Length in DeepWalk	50
Dimension in DeepWalk	12
Window size in Deepwalk	4

6 RESULTS

To evaluate the baselines and final model, the 10 initial nodes chosen as per strategy spread the information to their neighbors till 10 iterations. In first iteration the seed nodes spread information to their one-hop neighbors, which further in second iteration spread information to their one-hop neighbors and this was carried on for 10 iterations. This value is not too low that information won't spread and not too high that information spread goes on forever. The results are shown in Table 3. It shows the percentage of informed nodes out to total nodes in a village for both village 1 and village 2 on final model. There are also some results reported for some experiments carried out to finalize final mechanism to find central nodes after cluster formation. The clustering was done using "greedy_modularity_communities" algorithm. The results for the same are reported in Table 2.

Table 2. %age of nodes informed when using greedy modularity community detection

Active node set	**%age of nodes informed**
Betweeness	19.63%
Vote Rank	19.41%
Degree	19.41%
Closeness	19.41%
Page rank	20.09%
EigenVector	17%
Load	19.63%

7 RESULT ANALYSIS

The Table 2 show that PageRank can be used to find central nodes from clusters detected by clustering mechanisms. Though for baselines, degree centrality proved to be really effective as shown in Table 3, but it isnt really effective after clustering. Table 2 shows that pagerank, load centrality or betweeness centrality can be used after clustering is done. Further Table 3 show that OurModel beats all the baselines. For village 1, OurModel was able to spread information to 2% more nodes than best baseline i.e. PageRank, while this %age increased to 5% more nodes than the baseline for Village 2 which is betweeness centrality. This shows that community detection when jointly learnt with influential nodes search is a nice method to solve the problem.

8 FUTURE SCOPE AND CONCLUSION

This section discusses the Future Scope of the project and followed by Conclusion.

8.1 Future Scope

The authors currently have explored node embedding and clustering as a better method to obtain the community structure of the network. However other deep learning methods and approaches can be used in order to identify the influential nodes. The community detection can also be done using the same. Deep learning can also be employed in modelling the flow of information in the graph, currently general threshold model is used to simulate the flow of

Table 3. Percentage of Informed Nodes

Approach	**Village 1**	**Village 2**
Degree Centrality	20.03%	18.38%
Closeness Centrality	19.21%	20.09%
Betweeness Centrality	19.87%	22.95%
Page Rank	20.86%	18.38%
Katz Centrality	Failed to converge (50k)	8.68%
OurModel	**22.85%** ↑	**25.57%** ↑

information. Information flow itself can be used a metric to find the optimal influential nodes via machine learning or deep learning approach.

8.2 Conclusion

Deepwalk was used to obtain the latent representation of the nodes. K-means clustering was done on the embeddings to obtain the community structure of the network. Further this community structure was exploited to find influential nodes from the community using Pagerank and thus the information spread using the general threshold was observed. The results thus obtained are promising and it is able perform much better than the traditional methods of obtaining seed nodes like degree centrality, closeness centrality, betweeness centrality etc. Hence more work can be done on obtaining better latent representations that model the flow of information in the network. The problem of community detection and influential node search can be performed jointly to give better results.

REFERENCES

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2019. Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies* 86, 6 (2019), 2453–2490.

Dongming Chen, Panpan Du, Bo Fang, Dongqi Wang, and Xinyu Huang. 2020. A Node Embedding-Based Influential Spreaders Identification Approach. *Mathematics* 8, 9 (2020). https://doi.org/10.3390/math8091554

Chunxiao Jiang, Yan Chen, and KJ Ray Liu. 2014. Modeling information diffusion dynamics over social networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1095–1099.

Tutur Wicaksono, Agus Dwi Nugroho, Zoltán Lakner, Anna Dunay, and Csaba Bálint Illés. 2021. Word of mouth, digital media, and open innovation at the agricultural SMEs. *Journal of Open Innovation: Technology, Market, and Complexity* 7, 1 (2021), 91.