Proceedings of the International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT 2023)
IEEE Xplore Part Number: CFP23CV1-ART; ISBN: 978-1-6654-7451-1

2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) | 978-1-6654-7451-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/IDCIoT56793.2023.10053447

Leveraging Deep Learning to Spot Communities for Influence Maximization in Social Networks

Shambhavi Mishra
Dept. of Information Technology & Computer Applications
Madan Mohan Malviya University of Technology
Gorakhpur, India
mishra.shambhavi33@gmail.com

Rajendra Kumar Dwivedi
Dept. of Information Technology & Computer Applications
Madan Mohan Malviya University of Technology
Gorakhpur, India
rajendra.gkp@gmail.com

***Abstract*—Groups play a crucial role in affecting decisions of individuals who are part of the group. When it comes to social networks the group here may be small with some 10-15 members or very big contacting more than 100 members. Thus, there is high possibility of individuals belonging to one or more groups in social networks. It thus becomes important to activate influential members of a group to ensure maximum information propagation. This work proposes a community-based seed selection algorithm. The communities are first identified node embedding which performs graph clustering. After which proportionate distribution of seed nodes is carried out to ensure fair selection. Mapping node features to lower dimensional space and similar nodes getting placed closer to each other proves a better technique for community detection and is also expandable if new nodes get introduced in the network.**

Keywords—Social Networks, Influence Maximization, Network Embedding, Community Detection, Diffusion Model.

I. Introduction

The problem of influence maximization over a given social network is not a new problem statement. The social network has been in existence from a long period and the strategy has been adopted long back where few companies offer free sample products to few of their influential members and they in turn with their socializing effects make the product spread across the network of which they are a part. There has always been a limit to number of free samples that may be distributed given that these few are those participants who tend to show higher potential of spreading the product or the scheme which needs to be promoted. Devising a strategy to obtain this set of participants for an online social network, often referred to as seed nodes in literatures, is our very purpose. Some of the applications include viral marketing, sales improvement, rumor blocking and control, spreading information which is in public interests, political promotion and many more.

A. Motivation

The dispersive characteristic and drastically increasing engagement of widespread users has made online social network an un-comparable platform for thriving individuals and groups to market their services or products much more effectively by reaching a good count of influenced users. Learning the network structure and analyzing the available connection specific details can give better set of nodes collectively maximizing the combined influence.

B. Contribution

The contribution of this work is as under:

An effective graph neural network based learning of network structure along with embedding knowledge obtained using Node2vec algorithm.

Intra community and inter community informed walks generation to be used in embedding.

C. Organization

Upcoming sections of the paper appear in the following sequence: section II is dedicated to the preliminaries which help better connect the context, in section III we have discussed already contributed work and their limitations, proposed work is presented section IV and section V and VI contain results and conclusions respectively.

II. Background

Before the discussion of strategies and their limitations, some preliminaries must be covered to better understand the underlying context and concepts.

A. Online Social Media Network

A social media network may be a Facebook friend's network or citations and publishers' network or any other possible network can all be represented in a graph form where links denote edges and nodes are people connected via those links. This network can have nodes of similar properties that have bidirectional relations like those in friendship networks or it may be unidirectional like followers in Twitter where A may follow B but B also following back A may not always be true. If G(V,E) be a network then V is a set of nodes and E denotes ordered or unordered pairs (u, v); where u, v belong to V.

B. Problem Definition

Influence maximization problem on a social media network modeled as a graph aims to find a subset of the vertices or nodes S, called the seed nodes that act as information spreaders. These nodes have more potential to influence their connections. Here S<V and depends on k which is the maximum number of seed nodes that can be selected for the network.

$$S = \arg\max \sigma(n)$$

where n is the proper subset of V and equal to the budget allotted for the network, S represents the final set of seeds obtained after considering candidate nodes and the expected number of active nodes [14] is represented by σ(S).

C. Seed Nodes

The selection of a set of influencers in online social media network, is limited in budget, this means that from a given candidate set we can only pick k nodes where k denotes the budget allocated for the purpose. It is thus important to select nodes with maximum chances of spreading about the product. There are many considerations which have to be analysed such as number of neighbouring nodes, degree centrality[13] and other such parameters. The nodes or users finally shortlisted after considering all features are popularly termed as seed nodes. These seed nodes are the sources which work further to widespread required information using any of the diffusion models being followed in that network.

D. Diffusion Models

The propagation of a piece of information in a social network is commonly learned using linear threshold model or independent cascade model. After few seed-sets are obtained using the influence maximization method we pick the most effective seed set usually by estimating the number of activated nodes using one of these information diffusion models.

E. Representation Learning & Node Embedding

The idea of representation learning helps uncover various structural micro and macro level graph insights. Network embedding [7][16] is the low-dimensional representation of the network structural properties such that the nodes having similar properties are located nearer or closer in the embedding space. For a network with m nodes, the dimension of the embedding space is smaller than n. In our proposed work in this paper, we are using Node2vec algorithm that generates a vector representation based on walks generated [19][20]. Node2vec is preferred over other network embedding algorithms as the walk generation procedure here exploits both BFS (Breadth Fist Search) and DFS (Depth First Search), for microscopic as well as macro feature extraction of nodes which give a broader context about the overall network.

III. Related Work

This section covers few state-of-the-art methodologies to achieve the influence maximization task in a given network. Some of them use deep learning that has shown better results in community detection and network embedding that is preferred for generating lower dimensional representation of network.

Transfer learning is used by Khurshed Ali et. al [2] to minimize overall training time for deep reinforcement learning algorithm. This work addresses the major problem of unknown or partially known social networks [15]. The procedure involved querying known nodes which in turn reveal their neighbours. This is how the network is discovered. So the model learns the task of when to query a network and when to search of influential nodes in the network.

The real social networks are very complex and large. In order to get the subset of influential nodes the analysis of entire network as a whole is cumbersome task. Huan Li et. al[1] have tried to simplify this task by narrowing the network into some identified cliques. The seed nodes are then selected from these cliques. The method shows better results when compared to existing greedy methods devised for influence maximization. S. Singh [4] have proposed another group based approach. The method involves ant colony optimization for addressing influence maximization thus also emphasizing the group role in propagation.

Ling Wu et. al[8] have incorporated spatial proximity information in adjacency matrix by following matrix reconstruction. This reconstructed matrix works as input to autoencoder based convolutional neural network which then extracts spatial information of the network. Further the use of autoencoders and convolutional neural network helps in effective community finding [18] with better modularity value. Modularity is used as parameter to evaluate the community detection [10] process here.

Network embedding is opted by Nargess Vafaei et. al [16] where the node is represented in vector form and then the similarity between these nodes and the more popular or nodes occurring in number of vectors are considered for selection as influential nodes. Node embedding maps the nodes with similarity close in the embedding space. The distance between nodes usually determines the closeness between them.

Harshavardhan Kamarthi et. al [5] work with reinforcement learning where the agent learns to discover the unknown social network by surveying some of the known nodes of the network. It thus addressed the problem of influence maximization in initially unknown or partially unknown networks. The surveyed nodes here revealed the sub-graph of which they were a part resulting in discovery of larger network.

A more informative seed selection is proposed in the work by Shashank Sheshar Singh et. al [6]. The context or the interest of a node in the subject (or the product being promoted) is taken into consideration when selecting seed nodes from communities. They have called it C2IM or community-based context-aware influence maximization.

D. Sivaganesan et. al [22] have proposed technique in which they first use based influencing power for estimation for nodes. Two additional features are provided to aid in better estimating the influence including user's interest information and the social behaviour [21] that is the exchange of information it shows between its neighbors.

Mohammad Mehdi Keikha et. al [3] have given a deep learning backed solution to influence maximization problem in interconnected networks. The capabilities of bridge nodes is exploited in the work. Deep learning is used as a network structural feature extractor to better understand node capabilities. The literature is the first to employ network embedding to influence maximization problem.

IV. Proposed Work

Here we divide the entire process into two steps, at first, we identify communities present in the network using network embedding backed model by extracting structural features. Once the communities are identified, the second step is to analyze and decide the distribution of seed nodes among each of the identified communities.

Therefore, there is a two-level path to meet the requirement of obtaining influential seed set for a given network. Many algorithms discussed above carried out the task of influence maximization using community identification-based approaches but the problem of priority wise seed selection based on a community overall strength and other network properties is not taken into consideration while picking up seed nodes from these communities. The steps mentioned here are explained under the headings below:

A. Node Embedding Backed Community Identification Task

Network embedding is highly popular and being effectively used in current solutions to influence maximization. It gives a lower dimensional transformed vector corresponding to high-dimensional network nodes representation. In our approach we are using it in order to better learn the structural features required to further form communities. Finally, we will have nodes with similar features in one community.

The input graph or network is first used to generate walks. These walks are alternate combination of nodes and edges present in graph. These walks are then passed through the skip-gram model which looks upon the walk as a sentence and the nodes as the words in those sentences. It results in a vector representation of these network nodes. Using these representations, we group the nodes together based on how similar they are based on the new lower dimensional representation.

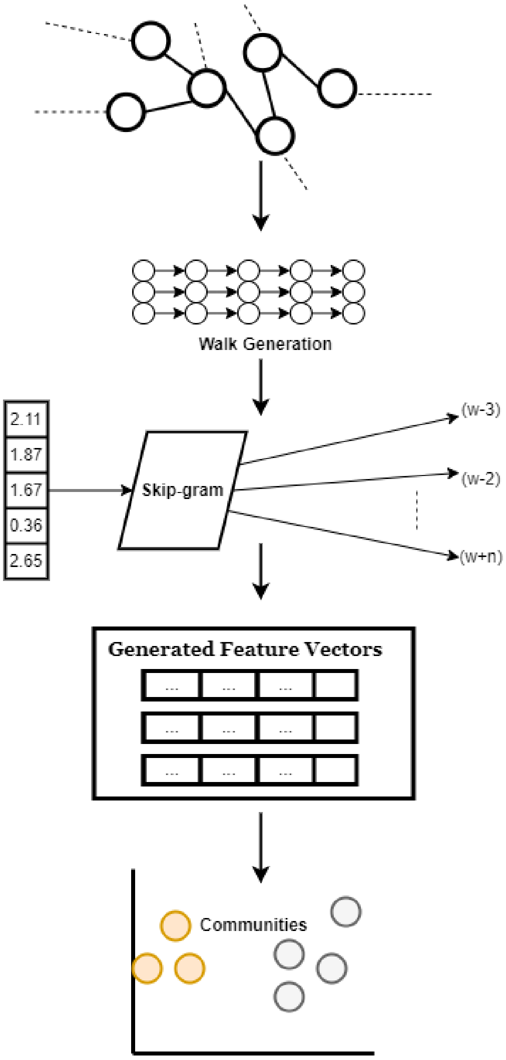

Fig. 1. Community Detection using Graph Embedding

The process flow is depicted in Fig. 1 where embedding model is given a network as input and it presents a vector representation containing the captured features. Node2vec [12] algorithm is used for node embedding in our case. It is capable of preserving graph structure and finds its application in problems like classification and community detection. Also the framework is designed to be a general solution for undirected or directed, weighted or un-weighted graphs. On the lines of Skip-gram architecture the network nodes can be treated as words and the walks generated depict the entire sentence formed by assembling these words in some order. The embedding is significantly dependent on random walks which in case of Node2vec depend on a factor called p that decides the algorithm's biasness at a particular stage in terms of considering DFS or BFS. If u is a node in the network then it can be represented with features in form of a vector by the following rule:

$$f : u \longrightarrow R^d$$

where R^d is the feature representation embedding and d, the dimension of the embedding space is usually very small from |V|.

B. Picking Seed Nodes

After the communities are captured in the above step, the next step is to decide the seed set selection approach. Similarly, when considering communities, one at a time, the descending order of strength of each of them shows the order in which preference must be given when selecting seed nodes.

Dividing the total budget k in the ratio of strength is followed in the next step. Considering a case where $\{C_1, C_2, C_3, \ldots C_p\}$ be community set obtained after network embedding based community detection algorithm and the ratio of strength represented by each of them is given as $[x_1, x_2, x_3, \ldots, x_p]$, then the budget k must be allocated in such a manner as to incorporate and pick the communities that will prove to provide the maximum reach as far as the budget is concerned. This work thus focuses on more problem-oriented selection strategy where the effort ratio is distributed more specifically for each community in terms of number of nodes comprising the final seed set.

C. Algorithm

The algorithm first examines each community to determine the order of priority of each community. We consider the community strength for this purpose. The obtained number for each of the p identified communities is stored as a ratio in vector r[p]. Now we loop over each community in order of its priority and calculate for each of the nodes belonging to the community its impact score. In this proposed model we our considering average of three centrality measures to get the impact value or impact score for each of the nodes present in the community being examined. The measures include degree centrality, betweenness centrality [17] and closeness centrality.

Degree centrality of a node is calculated by the number of connections the node posses with other nodes of the network. Betweenness centrality refers to how often the node has appeared in a shortest path between nodes and the closeness centrality refers to how far the node has to travel in the network to reach another node, thus a low value of

Algorithm: Priority wise seed node selection

Input:

Community set {$C_1(n_1,e_1)$,$C_2(n_2,e_2)$,$C_3(n_3,e_3)$, ...,$Cp(n_p,e_p)$} , no. of seed nodes to be selected k

Output:

k seed nodes

```
for i=1 to p do
    V[n(Ci)] = no. of nodes per cluster
end for
for i=1 to p do
    arrange in decreasing order w.r.t V[n(Ci)]
end for
r[p] = ratio of elements in vector V[n(Ci)]
for i=1 to p do
    for each(node k in Ci) do
        a[k] =compute average(degree centrality + betweenness centrality + closeness centrality) and rank in descending order
    end for
    seed-set[] = select top r[i] nodes from obtained list
end for
```

D. Final Seed Set Computation

Once the above computation is done for each of the nodes, we select only top r[1] nodes for C1, top r[2] nodes for C2, …,top r[p] nodes for Cp. Since the weightage is already fixed by the vector r[n] depending on the centrality measures stated above.

V. Performance Evaluation

This section presents the dataset used to test and evaluate the overall performance of the approach used.

A. Dataset Description

To examine and evaluate the proposed approach we used a combination of synthetic and real-world networks. Table 1. Shows the structural details containing the number of nodes, number of edges and types of graphs or networks.

The network are categorized to be either having directed edges which show unidirectional relations, such as followers in Twitter, or it may have undirected edged showing mutual relations as in Facebook. One more attribute is shown in the table that clarifies the graphs that were used in evaluating the performance of node embedding approach in community detection.

Structural details of the dataset are shown in the following Table.1.

Email-Eu-core[11] dataset shows links between email users. It comes with node mapped to a community, so it becomes suitable to be used to check the accuracy of node2vec whether it attains those number of communities.

DBLP[9] dataset is a authors dataset containing community details. The citation relation is the main reason for the links between users. Authors who work together tend to cite their co-authors work and the reverse may also follow.

TABLE. 1 DETAILS OF DATASET USED IN THIS WORK

Name	#Nodes	#Edges	Undirected/ Directed	Used for Comparison of Community Detection Phase
email-Eu-core	1005	25571	Directed	Yes
DBLP	317080	10,49,866	Undirected	Yes
ego-Twitter	81,306	17,68,149	Directed	No
Com-Amazon	3,34,863	9,25,872	Undirected	Yes

Ego-Twitter is another dataset from SNAP giving a network of Twitter users. This dataset contains directed edges showing the unidirectional aspect where X following Y may not always lead to Y following back X.

Thus, the datasets are not all of the same type. This helps us to test our proposed approach on more likely heterogeneous real networks where node behave different from one another.

B. Results and Discussion

After the seed set is obtained by executing the algorithm on input graph, we have opted with the Independent Cascade (IC) model for realizing the actual information spread achieved by this algorithm generated seed nodes.

The IC model for information diffusion considers that initially a network has few active nodes, and these active nodes or influencers have to carry out the process of

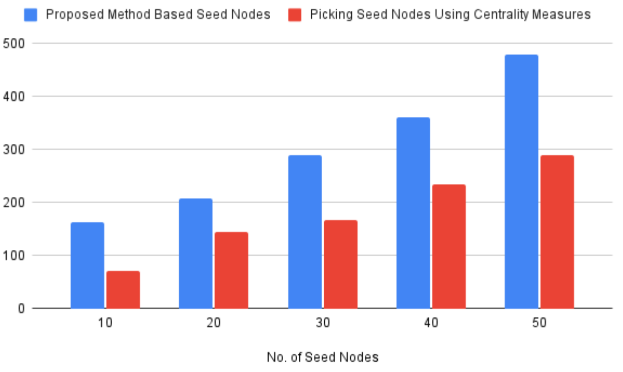

Fig. 2. Comparison between proposed model obtained activated nodes and nodes activated by directly applying centrality measures.

activating their neighbors with some probability. The process terminates when there is no more chance of any other nodes to get activated.

The seed set given by the model was run for 100 Monte Carlo simulations under IC diffusion model to the above numbers of activated nodes. The average number of activated users was taken to be the estimated activation power of the seed nodes thus examined.

To evaluate the community detection phase, we have used some synthetic datasets specifically to check the performance of node embedding. The email-Eu-core dataset also contains mapping of nodes to communities. We have used this data to compare the actual existing communities and generated communities by using the mentioned approach.

Making community specific decision beforehand seems to play a key role in the process of influence maximization. The portions of high engagements are more critically injected with these seed nodes than rest of the network. We have shown the comparison with method in which we directly apply centrality measures Fig. 2. shows the spread achieved by both the models with respect to varying number of seed nodes selected. The vertical axis shows the activated nodes.

The results clearly show that community plays an important role in the process of information propagation and influence maximization. The additional information which the node embedding generated helped tremendously in identifying better seed nodes.

VI. Conclusion and Future Work

The pro-active nature in proportionally distributing the final k value into communities was proposed in this work. It has shown its effectiveness over such models that are backed by community detection approaches. The community wise seed set helps in effectively dividing the bigger problem when considering the entire network as a whole. Although many such approaches have been proposed but network embedding has given effective result when finally analyzing the count of activated nodes or users that are not limited to few communities but cover network as one.

This distribution is executed by exploring centrality measures of nodes for this work but there is a scope of improvement by considering other such properties of nodes and edges to be specific and network in general. The method tries to address the scalability issue by working in parts so as to divide computation overhead when large networks are taken into consideration.

This arrangement can also be further employed to test its capability in approaching the same problem in inter-connected networks where bridge nodes often play crucial role. Drawing concepts of network coupling and community detection together can be worked upon to reach some conclusions.

References

[1] Li, H., Zhang, R., Zhao, Z. and Yuan, Y., 2019. An Efficient Influence Maximization Algorithm Based on Clique in Social Networks. IEEE Access, 7, pp.141083-141093.

[2] K. Ali, C.-Y. Wang, M.-Y. Yeh, and Y.-S. Chen, “Addressing competitive influence maximization on unknown social network with deep reinforcement learning,” 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020.

[3] Keikha, M., Rahgozar, M., Asadpour, M. and Abdollahi, M., 2020. Influence maximization across heterogeneous interconnected networks based on deep learning. Expert Systems with Applications, 140, p.112905.

[4] Singh, S., Singh, K., Kumar, A. and Biswas, B., 2019. ACO-IM: maximizing influence in social networks using ant colony optimization. Soft Computing, 24(13), pp.10181-10203.

[5] H. Kamarthi, P. Vijayan, B. Wilder, B. Ravindran and M. Tambe, "Influence maximization in unknown social networks: Learning Policies for Effective Graph Sampling", *CoRR*, 2019.

[6] S. S. Singh, A. Kumar, K. Singh, and B. Biswas, “C2IM: Community based context-aware influence maximization in social networks,” ScienceDirect, 2018.

[7] H. Li, M. Xu, S. Bhowmick, C. Sun, Z. Jiang and J. Cui, "DISCO: Influence Maximization Meets Network Embedding and Deep Learning", *arXiv:1906.07378v1 [cs.SI]*, 2019.

[8] L. Wu, Q. Zhang, C. Chen, K. Guo and D. Wang, "Deep Learning Techniques for Community Detection in Social Networks", *IEEE Access*, vol. 8, pp. 96016-96026, 2020. Available: 10.1109/access.2020.2996001.

[9] J. Yang and J. Leskovec. Defining and Evaluating Network Communities based on Ground-truth. ICDM, 2012.

[10] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. "Local Higher-order Graph Clustering." In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017.

[11] J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

[12] Grover, A. and Leskovec, J. (2016) “Node2vec,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* [Preprint]. Available at: https://doi.org/10.1145/2939672.2939754.

[13] N. Shakeel and R. K. Dwivedi, "A Learning Based Influence Maximization across Multiple Social Networks," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2022, pp. 53-58, doi: 10.1109/Confluence52989.2022.9734145.

[14] N. Shakeel and R. K. Dwivedi, "A Comparative Analysis of the Techniques for Influence Maximization in Social Networks," 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), 2022, pp. 674-678, doi: 10.23919/INDIACom54597.2022.9763226.

[15] J. Ko, K. Lee, K. Shin and N. Park, "MONSTOR:An Inductive Approach for Estimating and Maximizing Influence over Unseen Social Networks", *CoRR*, 2020. [Accessed 20 September 2022].

[16] N. Vafaei, M. R. Keyvanpour and S. Vahab Shojaedini, "Influence Maximization in Social Media: Network Embedding for Extracting Structural Feature Vector," 2021 7th International Conference on Web Research (ICWR), 2021, pp. 35-40, doi: 10.1109/ICWR51868.2021.9443150.

[17] M.J. Newman, A measure of betweenness centrality based on random walks, Social Networks 27 (1) (2005) 39–54.

[18] S. Mishra, S. Singh, S. Mishra and B. Biswas, "TCD2: Tree-based community detection in dynamic social networks", *Expert Systems with Applications*, vol. 169, p. 114493, 2021. Available: 10.1016/j.eswa.2020.114493.

[19] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in KDD. ACM, 2014, pp. 701–710.

[20] S. Abu-El-Haija, B. Perozzi, R. Al-Rfou, and A. Alemi, “Watch your step: Learning graph embeddings through attention,” CoRR, vol. abs/1710.09599, 2017.

[21] N. Hudson and H. Khamfroush, "Behavioral Information Diffusion for Opinion Maximization in Online Social Networks," in IEEE

Transactions on Network Science and Engineering, vol. 8, no. 2, pp. 1259-1268, 1 April-June 2021, doi: 10.1109/TNSE.2020.3034094.

[22] Sivaganesan, D. "Novel Influence Maximization Algorithm for Social Network Behavior Management." Journal of ISMAC 3, no. 01 (2021): 60-68.