Indonesian Journal of Electrical Engineering and Computer Science
Vol. 24, No. 3, December 2021, pp. 1604~1609
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v24.i3.pp1604-1609

Community-based framework for influence maximization problem in social networks

Mustafa K. Alasadi, Ghusoon Idan Arb
Faculty of Computer Science and Information Technology, University of Sumer, Al-Rifai, Iraq

Article Info

Article history:

Received Jul 26, 2021
Revised Oct 27, 2021
Accepted Nov 1, 2021

Keywords:

Community structure
Influence maximization
Information diffusion

ABSTRACT

Given a social graph, the influence maximization problem (IMP) is the act of selecting a group of nodes that cause maximum influence if they are considered as seed nodes of a diffusion process. IMP is an active research area in social network analysis due to its practical need in applications like viral marketing, target advertisement, and recommendation system. In this work, we propose an efficient solution for IMP based on the social network structure. The community structure is a property of real-world graphs. In fact, communities are often overlapping because of the involvement of users in many groups (family, workplace, and friends). These users are represented by overlapped nodes in the social graphs and they play a special role in the information diffusion process. This fact prompts us to propose a solution framework consisting of three phases: firstly, the community structure is discovered, secondly, the candidate seeds are generated, then lastly the set of final seed nodes are selected. The aim is to maximize the influence with the community diversity of influenced users. The study was validated using synthetic as well as real social network datasets. The experimental results show improvement over baseline methods and some important conclusions were reported.

Corresponding Author:

Mustafa K. Alasadi
Faculty of Computer Science and Information Technology, University of Sumer
Dhi-Qar, Al-Rifai, Iraq
Email: mk3_alasadi@yahoo.com

1. INTRODUCTION

Recently, online social networks (OSN) and their services have a direct effect on our daily life. People and organizations prefer using OSN (e.g. Twitter, Facebook, and Microblog) as a fast and easy medium for communicating and propagation of news, opinions, beliefs, and other types of information [1], [2]. OSN services allow individuals to create social relations such as friendship, through which they influence each other. Based on common interest and some other factors, the users of OSN and their social relations form a community structure, in which the similar users are densely linked to each other and loosely connected with other users [3]. In fact, the communities are usually overlapped since the people naturally belong and participate in multiple communities at the same time. The nodes (users) that belong to more than one community are known as overlapped nodes that play a critical role in information diffusion among the communities.

A graph $\mathrm{G} = (\mathrm{V}, \mathrm{E}, \mathrm{W})$ is usually used to represent and visualize the social relations, where V is a set of nodes defines the individuals and E is a set of links (edges) represent the relationships among the nodes, while the weight W represents the relationships strength. Each edge is represented by $e = (u, v) \in E$ and $u \neq v$, meaning the existence of a link between node u and node v. The similarity in user interest and social influence encourages organizations and companies to exploit this medium (OSN) for maximizing the spread of information related to their products. The influence maximization problem (IMP) was derived from the viral marketing, in which a product is given to some people for free in order to reach an expansive number of potential customers through the word-of-mouth impact [4], [5]. The influence maximization problem introduced and formulated by Kempe *et al.* in [6] and [7]. The problem is defined as finding a subset $S \in V$ with the size of k nodes from G, through which the diffusion of influence from seeds S will be maximized. Numerous researches and experiments have been conducted to solve the influence maximization problem by utilizing different solutions methodologies [8].

In order to simulate the diffusion process, all the empirical studies utilize the information diffusion model like the independent cascade model (IC), linear threshold model (LT) and other derived models. In this work, the IC model is used in which the diffusion process unfolds in a step-by-step fashion. Briefly, given a directed graph represents a social network, the diffusion process begins from an initial active set of nodes where: at time t, each seed node u from the set of initially active nodes, has a single attempt to activate each of previously inactive successor *neighbors* $v \in N_{out}(u);$ with probability $p_{u,v}$, where $p_{u,v} \in [0,1]$. If node v gets activated (switch to contagious node), then at time $\mathrm{t}+1$ it tries to activate the inactive neighbors. The process terminates when there are no more contagious nodes [9].

In this paper, we propose a framework to solve the influence maximization problem with higher diversity in activated nodes over the communities. Diversity means “more nodes belong to more categories” [10]. The framework integrates both the community and heuristic based solutions, where it consists of three phases: first, the community structure is discovered and all the overlapped nodes are defined, in the second phase, a set of candidate seeds is generated, lastly, the final seed nodes are selected. To signify the importance of a node, the emphasis is on the overlapped nodes and the nodes with the highest centrality measure. The IC model has been used to model information diffusion in experiments on real and synthetic networks. The contributions of this work are summarized by the following: i) Propose an efficient framework to tackle the IMP with better running time and scalability; ii) Integrate both community-based and heuristic-based solutions in one framework; and iii) Highlight the special role played by overlapped nodes to achieve community diversity of influenced users.

2. RELATED WORK

The independent cascade (IC) model and linear thresholds (LT) model are two seminal discrete influence propagation models [6]. Based on these two models, a variety of studies have been dedicated to address the IMP like [11]-[13], the aim is to select a limited number of nodes in social networks that could influence the maximum number of other nodes, summarized the recent works on IMP [5]. The study in [6], formulated the influence maximization as an optimization issue and demonstrated that the IMP is NP-hard.

The “path-based influence maximization (PB-IM)” is one of the promising solutions of IMP. Since evaluating the influence process is #P-hard and hard to solve in polynomial time, the work [14] introduced a scalable parallelizable approximation algorithm, “independent path algorithm (IPA)”. IPA defines the influence evaluation unit as a path from a seed node to another node and demonstrates that influence paths are independent of each other. Then, the influence approximation process only requires a series of simple arithmetic of influence paths. If all influence paths are collected, the algorithm does not need more time to approximate influence diffusion.

Ko *et al.* [15] propose a hybrid influence maximization method, by combining two solution approaches the Path-Based and community-based influence maximization. The hybrid method intends to address the performance problems of a simple Greedy method at micro and macro levels. With the Greedy approach, running Monte-Carlo simulations and iteratively evaluating the influence dissemination of each node after each seed selection step, is fairly expensive and leads to significant computational overhead. In their work, two technical solutions were introduced to improve the performance of the seed selection process. Centrality measure indicates how significant a node has in the social graph. Consequently, centrality based approaches claim that the users with high centrality are more fortunate to be influencers [16]. Numerous researches followed this approach to tackle the influence maximization problem [17]-[19].

Rahimkhani *et al.* [17] discover the community structure and use the betweenness centrality to distinguish the effective communities. The degree centrality was also used to select the set of candidate nodes from the influential communities. Finally, the proposed ComPath method finds the top-k most influential nodes. Despite the results that refer the ComPath reduces the number of investigated nodes by considering only the significant communities but still, it needs much processing time because all paths among the candidate nodes should be evaluated. In order to identify the top k influential nodes, Bozorgi *et al.* [20] tackle this problem from a community-based perspective. They study the local influence of users inside their communities as well as the global influence. By considering both the local as well as global diffusion ability. The proposed method (INCIM) finds the list of nodes’ diffusion values per community. In their experiment, the SLPA algorithm was utilized to detect the graph communities and linear threshold model (LT) is used for simulating the diffusion process. In spite of highlighting the role of community structure in solving the IMP, the work did not consider the overlapping of communities and ignored the semantic aspect.

3. THE SOLUTION FRAMEWORK OF INFLUENCE MAXIMIZATION PROBLEM

This section details the proposed community-based IMP approach. As mentioned previously the framework comprises three phases: i) overlapping community detection, ii) candidate seed generation, and iii) seed selection. Algorithm 1 summarizes all these phases.

3.1. Overlapping community detection

The community (module) is a subgroup of nodes (users) that are densely connected and loosely connected to the rest of the nodes in the graph [21]. The users of social platforms are naturally participating in multiple communities and rarely divided into disconnected groups, resulting in overlapped community structure. The community structure is a related term defined as a group of modules or communities that exist in the graph, and denoted as, $CS = \{\mathrm{c}_1, \mathrm{c}_2, \mathrm{c}_3, \dots, \mathrm{c}_\mathrm{k}\}$, where CS stand for the community structure and $\mathrm{c}_1, \mathrm{c}_2, \mathrm{c}_3, \dots, \mathrm{c}_\mathrm{k}$ are the communities. For instance, in Figure 1, three communities are shown with different node colors as well as one overlapped node belong to c_1 and c_2, whereas community structure $CS = \{\mathrm{c}_1, \mathrm{c}_2, \mathrm{c}_3\}$. Through literature study, we find that the embedded information within the community structures is of great usefulness in solving the IMP. The link density-based technique clique percolation (CPM) presented in [22] used to discover the community structure. The overlapped nodes are in turn identified by analyzing the outcome of the community discovering process.

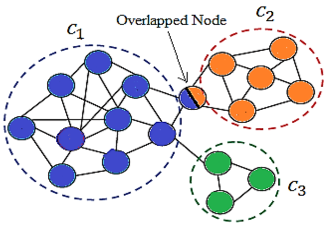

Figure 1. An example of community structure

3.2. Candidate seed generation

After discovering the community structure, the aim here is generating a set of candidate seed nodes (*SN*) taking into account the relative importance of the nodes. Limiting a number of seeds in this set, is a crucial point in IMP since the social networks are extremely large and select seeds with maximal influence in such a huge search space is a real challenge. In light of this fact, it is important to significantly decrease the number of candidate seeds. To achieve diversity in activated nodes across the entire graph communities, we generate two candidate sets of nodes, primary and secondary sets.

All the overlapped nodes are considered as a primary candidate set (*prim_cand*). As a secondary candidate seed, we chose from each non-overlapped community a node with highest degree centrality (*secn_cand*). Here the focus is on overlapped nodes because of the special role these nodes play in the information spreading process because they belong to a number of communities at the same time. These special nodes are considered as an interface among all the communities they belong to and provide more network connectivity which is very useful for solving the IMP.

Definition 1: Overlapped Sets (OS): is a collection of distinct sets, each of which contains the overlapped nodes among different communities. We notify it as: $OS = \{os1, os2, os3, \dots., osm\}$ where $os1 \cap os2 \cap os3$ .... $\cap\, osm = \emptyset$ and m is the number of overlapped sets. The primary candidate seed is defined as: $prim\_cand = \{os1 \cup os2 \cup \dots . \cup osm\}$, where $|prim\_cand| =$ number of overlapped nodes in graph G.

3.3. The seed selection

The third phase aims to select k nodes (from the candidate seed nodes) as final seed nodes (SN). This phase consists of two steps. In the first step, for each overlapped set (os_i) the degree centrality is calculated for all the nodes and the nodes are sorted accordingly. Then determine the number of nodes to be chosen as seed, where the number Ni is specified according to the size of the overlapped set (n_i) to the total number of overlapped nodes in graph G ($|prim\_cand|$) (1).

$$N_{os_i} = k * \frac{n_i}{\sum_{j=1}^{m} n_j} \tag{1}$$

At the end of this step, Ni top-degree centrality nodes are selected from each overlapped set (os_i). In the second step of this phase, the diffusion process is initiated with seed selected from the first step and the number of influenced nodes is recorded. Then for each node U in the secondary candidate set, the diffusion process is repeated with exchange the node with the lowest centrality from the set generated in the first step with node U. When the dissemination process is stopped and according to the influence volume, node U can be kept as a seed instead of the overlapped node.

Algorithm 1: Community based IMP

```
Input: The Graph G(V, E, W); Number of seed nodes k
Output: k seed nodes
Sn ← {} //  empty seed set
prim_cand ← {}         // the primary candidate set
secn_cand ← {}          // the secondary candidate set
max=0
// 1: Community detection
$CS = CPM(G)$        // discover the community structure  $\{c_1, c_2, c_3, \dots, c_x\}$
OS= {Overlapped Set from $CS$}      // Definition 1
NC= {Non-overlapped Communities from $CS$}
// 2: candidate seed generation
For each osi  ∈ OS do
     prim_cand = prim_cand ∪ osi    // generate primary candidate set
      Find $N_{\mathrm{os_i}}$ // using equation 1
End
For each ci ∈ NC do
     secn_cand= secn_cand ∪ {the node with  highest degree centrality in ci }
End
//3: Seed Selection
For each node i ∈ prim_cand do
     store the degree centrality of i
End
For each osi ∈ OS do
     Sn= Sn ∪  {the top ($N_{os_i}$) degree centrality nodes from osi }
End
A (G) = ICSn(G)   // execute IC model with seed Sn , A(G) is set of activated nodes
max=|A (G)|
For each node U ∈ secn_cand do
     Sn´ = SN \ {the node with the lowest degree centrality from Sn}
     Sn´ = Sn´ ∪{ U }
     A (G)= ICSn´(G)   // execute IC model with seed Sn´
     If  MAX < |A (G)| then
          MAX = |A (G)|
          Sn = Sn´
     Else :
               Continue
End
Output: the node in Sn.
```

4. RESULTS AND DISCUSSION

To evaluate the proposed approach, the experiments were conducted on four datasets, two real-world datasets, and the others are synthetic, Table 1 detail the statistics of datasets. To investigate the role of community structures in IMP, the synthetic datasets are formed to have the same number of nodes and edges but with different community structures. The datasets-meme tracker dataset [23], in this data, the websites and bloggers are considered as users. The data formed by tracking the diffused content (news stories & blog posts) from about one M sources. Digg dataset [24] contains data about stories that appeared on Digg's front page in 2009 over a period of thirty days.

Community detection-using the CPM algorithm described in the introduction, the community structure is discovered for each dataset graph. Table 2 shows the result of the community detection process. In order to accomplish the candidate seed generation, information like overlapped nodes and nodes degree centrality is stored in a proper data structure. Figure 2 shows a sample of overlap communities discovered from the Digg dataset.

IMP and diffusion model: the well-known independent cascade model (IC) is utilized following the steps of Algorithm 1. For each dataset, the model parameters are specified based on our previous works described in [25], [26] and the diffusion process is initiated using the k seeds selected based on the proposed approach. For comparison purposes, we implement two baselines: a random election of seeds in addition to the degree heuristic method (DH), which only selects the top-k largest-degree nodes as the seed nodes. The results (number of activated nodes) are shown in Table 3.

Table 3 shows that the proposed community-based approach outperforms the baseline method in terms of a number of node activation. In addition to the volume of diffusion, we note from the experiments that spread of information reaches almost 75 % of the communities, and it is an important result in terms of diversity. In contrast, the diffusion in the baseline approaches has affected nodes only in specific communities and did not unfold globally across the network.

Table 1. The statistic of datasets

Dataset	Number of nodes	Number of links	Type
MemeTracker	960	2000	directed
Digg	2008	10333	
Syn_net 1 and 2	1000	5000	

Table 2. Results of overlap community detection

Dataset	Number of communities	Number of overlap nodes
MemeTracker	27	25
Digg	398	584
Syn_net 1	152	91
Syn_net 2	78	40

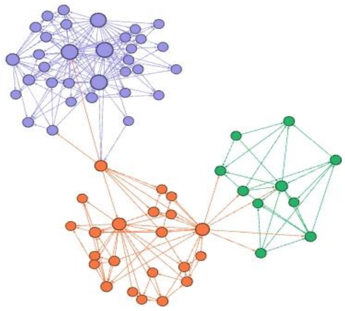

Figure 2. Sample of overlapped communities

Table 3. Comparison results with baselines based on number of activated nodes

Dataset	Proposed framework	Degree heuristic (DH)	Random selection
MemeTracker	230	187	49
Digg	387	302	71
Syn_net 1	283	160	66
Syn_net 2	185	172	54

5. CONCLUSION

Discovering the community structure of the underlying network is of high importance in solving IMP. This intermediate step brings down the problem to the community level and improves scalability. Selecting overlap nodes as seed nodes to trigger the diffusion increases the number of activated nodes (as compared with baseline methods) across diverse communities. The work shows the efficiency of combining community-based solutions with a heuristic approach since the proposed method depends on the overlapped nodes that have a high score of degree centrality measure.

REFERENCES

[1] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, “Advances in Social Media Research: Past, Present and Future,” *Inf. Syst. Front.*, vol. 20, no. 3, pp. 531–558, 2018, doi: 10.1007/s10796-017-9810-y.

[2] D. C. Cercel and S. Trausan-Matu, “Opinion propagation in online social networks: A survey,” *ACM Int. Conf. Proceeding Ser.*, Dec. 2016, 2014, doi: 10.1145/2611040.2611088.

[3] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, “Comparing community structure to characteristics in online collegiate social networks,” *SIAM Rev.*, vol. 53, no. 3, pp. 526–543, 2011, doi: 10.1137/080734315.

[4] J. Zhu, S. Ghosh, and W. Wu, “Group Influence Maximization Problem in Social Networks,” *IEEE Trans. Comput. Soc. Syst.*, vol. 6, no. 6, pp. 1156–1164, 2019, doi: 10.1109/TCSS.2019.2938575.

[5] Y. Li, J. Fan, Y. Wang, and K. L. Tan, “Influence Maximization on Social Graphs: A Survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1852–1872, 2018, doi: 10.1109/TKDE.2018.2807843.

[6] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” *Theory Comput.*, vol. 11, pp. 105–147, 2015, doi: 10.4086/toc.2015.v011a004.

[7] P. Domingos and M. Richardson, “Mining the network value of customers,” *Proc. Seventh ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 57–66, 2001, doi: 10.1145/502512.502525.

[8] S. Banerjee, M. Jenamani, and D. K. Pratihar, “A survey on influence maximization in a social network,” *Knowl. Inf. Syst.*, vol. 62, no. 9, pp. 3417–3455, 2020, doi: 10.1007/s10115-020-01461-4.

[9] D. Kempe, J. Kleinberg, and É. Tardos, “Influential nodes in a diffusion model for social networks,” *Lect. Notes Comput. Sci.*, vol. 3580, pp. 1127–1138, 2005, doi: 10.1007/11523468_91.

[10] J. Li, T. Cai, K. Deng, X. Wang, T. Sellis, and F. Xia, “Community-diversified influence maximization in social networks,” *Inf. Syst.*, vol. 92, no. March, 2020, doi: 10.1016/j.is.2020.101522.

[11] M. Jaouadi and L. Ben Romdhane, “Influence maximization problem in social networks: An overview,” *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, 2019, doi: 10.1109/AICCSA47632.2019.9035366.

[12] K. X. Yu Wang, Gao Cong, and Guojie Song, “Community-based Greedy Algorithm for Mining Top-K Influential Nodes in Mobile Social Networks Categories and Subject Descriptors,” *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 1039–1048, 2010, doi: 10.1145/1835804.1835935.

[13] S. Chen, J. Fan, G. Li, J. Feng, K. L. Tan, and J. Tang, “Online topic-aware influence maximization,” *Proc. VLDB Endow.*, vol. 8, no. 6, pp. 666–677, 2015, doi: 10.14778/2735703.2735706.

[14] J. Kim, S. K. Kim, and H. Yu, “Scalable and parallelizable processing of influence maximization for large-scale social networks?,” *Proc. - Int. Conf. Data Eng.*, pp. 266–277, 2013.

[15] Y. Y. Ko, K. J. Cho, and S. W. Kim, “Efficient and effective influence maximization in social networks: A hybrid-approach,” *Information Sciences,* vol. 465. pp. 144–161, 2018, doi: 10.1016/j.ins.2018.07.003.

[16] M. Ahsan, T. Singh, and M. Kumari, “Influential node detection in social network during community detection,” *Proc. - 2015 Int. Conf. Cogn. Comput. Inf. Process. CCIP 2015*, 2015, doi: 10.1109/CCIP.2015.7100737.

[17] K. Rahimkhani, A. Aleahmad, M. Rahgozar, and A. Moeini, “A fast algorithm for finding most influential people based on the linear threshold model,” *Expert Syst. Appl.,* vol. 42, no. 3, pp. 1353–1361, 2015, doi: 10.1016/j.eswa.2014.09.037.

[18] X. Li and Q. Sun, “Identifying and ranking influential nodes in complex networks based on dynamic node strength,” *Algorithms,* vol. 14, no. 3, pp. 1–11, 2021, doi: 10.3390/a14030082.

[19] Y. H. Fu, C. Y. Huang, and C. T. Sun, “Using global diversity and local topology features to identify influential network spreaders,” *Phys. A Stat. Mech. its Appl.*, vol. 433, pp. 344–355, 2015, doi: 10.1016/j.physa.2015.03.042.

[20] A. Bozorgi, H. Haghighi, M. Sadegh Zahedi, and M. Rezvani, “INCIM: A community-based algorithm for influence maximization problem under the linear threshold model,” *Inf. Process. Manag.*, vol. 52, no. 6, pp. 1188–1199, 2016, doi: 10.1016/j.ipm.2016.05.006.

[21] Z. Dhouioui and J. Akaichi, “Overlapping community detection in social networks,” *Proc. - 2013 IEEE Int. Conf. Bioinforma. Biomed. IEEE BIBM 2013*, vol. 45, no. 4, pp. 17–23, 2013, doi: 10.1109/BIBM.2013.6732729.

[22] G. Palla, I. J. Farkas, P. Pollner, I. Derényi, and T. Vicsek, “Directed network modules,” *New J. Phys.*, vol. 9, 2007, doi: 10.1088/1367-2630/9/6/186.

[23] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle:. *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining,* 2009, doi: 10.1145/1557019.1557077.

[24] T. Hogg and K. Lerman, “Social dynamics of Digg,” *ICWSM 2010 - Proc. 4th Int. AAAI Conf. Weblogs Soc. Media,* pp. 247–250.

[25] M. K. Alasadi and H. N. Almamory, “Diffusion model based on shared friends-aware independent cascade.” *2nd International Science Conference. IOP Conf. Series: Journal of Physics: Conf. Series,* 2019, doi: 10.1088/1742-6596/1294/4/042006.

[26] M. K. Mahdi and H. N. Almamory, “Modeling the information diffusion of overlapped nodes using SFA-ICBDM,” *Int. J. Recent Technol. Eng*., vol. 8, no. 2, pp. 709–713, 2019, doi: 10.35940/ijrte.B1710.078219.

BIOGRAPHIES OF AUTHORS

Dr. Mustafa K. Alasadi, From Iraq, Kerbala city, 1986, obtained the Ph.D in Computer Science from the University of Babylon-College of Information Technology, 2020. He is currently a lecturer at Computer Science Department, Sumer University, Iraq. His research focuses on social networks and recommendation systems.

Ghusoon Idan Arb, From Iraq, Dhi-Qar city, 1987, obtained the master degree in Computer Engener, 2014. She is currently a lecturer at Computer Science Department, Sumer University, Iraq. Her research focuses on Cloud computing, GIS, Web design.