A Novel Approach for Community Detection Using the Label Propagation Technique

Chapter *in* Studies in Computational Intelligence · September 2019

DOI: 10.1007/978-981-10-8797-4_14

Jyoti Shokeen
Maharshi Dayanand University

Chhavi Rana
Maharshi Dayanand University

A Novel Approach for Community Detection Using the Label Propagation Technique

Jyoti Shokeen, Chhavi Rana and Harkesh Sehrawat

Abstract In this chapter, we propose a new label propagation-based approach to detect community structure in social networks. This is a multiple label propagation technique in which a node can obtain labels of different communities, which allows researchers to discover overlapping communities. One important advantage of this approach is the updating of node labels over time that makes it dynamic. Given an underlying social network, we assume that each node receives a unique label id similar to its node id in the initial phase. We allow each node to accept multiple labels from its neighbors if each of the neighbors has a high common neighbor score, which naturally encompasses the idea of *overlapping communities*.

1 Introduction

The goal of social network analysis is to investigate and analyze the relationships between individuals or groups in a network with the aim of determining how these relationships arise and decline, and the consequences that emerge [1]. The widely studied research sphere of social network analysis is to comprehend the structure of networks in terms of communities, as the detection of communities helps us gain insights into the network. In the literature, a general definition of community within social networks is lacking. Several authors have attempted to define community, but the definition varies with the algorithms. In context of social networks, a

J. Shokeen (✉) · C. Rana · H. Sehrawat
University Institute of Engineering and Technology, Rohtak, Haryana, India
e-mail: jyotishokeen12@gmail.com

C. Rana
e-mail: chhavi1jan@yahoo.com

H. Sehrawat
e-mail: sehrawat_harkesh@yahoo.com

A. N. Krishna et al. (eds.), *Integrated Intelligent Computing, Communication and Security*, Studies in Computational Intelligence 771,
https://doi.org/10.1007/978-981-10-8797-4_14

community is a group of people sharing similar interests who communicate frequently with others in the group. In broad terms, a community has many intra-links and fewer inter-links. Generally, there exist two types of communities in social networks: disjoint communities and overlapping communities. When a node in a network belongs to only a single community, it is referred to as a disjoint community. When the nodes in a network belong to more than one community, the communities are termed as overlapping communities. There is a considerable amount of literature on community detection algorithms for social networks [2–9, 11, 12, 14]. Graph partitioning, clustering, label-propagation, clique, game theory and genetic algorithms are some of the techniques that have been used detect communities in social networks. Of these, label propagation-based algorithms have been proved to be the fastest [9].

The chapter is organized as followings. The first section gives a brief overview of community definitions in social networks. The second section examines the related work in detecting community structure using label propagation algorithms. A new methodology with some definitions is described in the third section. Our conclusions are drawn in the final section.

2 Related Work

Several community detection techniques have been implemented in the past 10 years. Many of these algorithms fail to detect communities in large networks or they exhibit issues related to complexity of time and space. An algorithm recently proposed by [9] gives promising results in terms of complexity in linear time. Label propagation is a community detection technique in which each node possesses a label that denotes the community to which it belongs. A node acquires the label given to most of its neighboring nodes. Prior knowledge about communities is not needed in label propagation-based algorithms. This section briefly explains these algorithms.

Raghavan et al. [9] were the first to propose label propagation algorithm that uses only network structure to detect both disjoint and overlapping communities. The algorithm is computationally inexpensive and takes approximately linear time, $O(m)$, where m is the number of edges. Unfortunately, it has some downsides, the first of which is the random initialization of node labels that results in unpredictable performance. Secondly, it follows a greedy approach for updating labels to produce tremendously large communities, due to which small communities are wiped out. Recently, many researchers have devised algorithms to address the problem of the original label propagation algorithm. Gregory [2] proposed the Community Overlap Propagation Algorithm (COPRA), which is an extended version of the original algorithm proposed by [9], but it was the earliest label propagation-based algorithm that detected overlapping communities. It permits each node to update its coefficients by calculating the average coefficients of the neighbor nodes, synchronously at each time step. The parameter v monitors the highest number of communities to which nodes can relate. However, this parameter is node independent. When some nodes in the network belong to a small number of communities and some nodes belong to a large number of communities, then it becomes tough for the algorithm to find an appropriate value of v that complies with both types of nodes simultaneously. The time per iteration for the algorithm is $O(vmlog(vm/n))$, and time taken by the initial and last steps is $O(v(m + n) + v^3n)$, where n, m and v refer to the number of nodes, edges and maximum number of communities per node, respectively. COPRA also works well for bipartite and weighted networks. The Speaker-Listener Label Propagation Algorithm (SLPA) [12] is another advancement over LPA [9] for discovering overlapping structures in social networks. Unlike the original LPA and COPRA, where the nodes forget the knowledge acquired in the preceding iterations, SLPA gives memory to each node to store received labels. The time complexity of SLPA is $O(Tm)$, where T and m denote the maximum number of iterations and total edges, respectively. By generalizing the interaction rules, SLPA can also fit for weighted and directed networks.

3 Proposed Work

Unlike other algorithms [9] in which nodes are initially given unique labels, we use quite a different strategy to assign labels to the nodes in the initial stage. We propose an Overlapping Multiple Label Propagation Algorithm (OMLPA) that assigns multiple labels to the nodes in the network to allow the detection of overlapping communities. A community is made of densely interconnected nodes, which implies that nodes sharing connections with a large number of nodes are likely to belong to the same community. The common neighbor similarity (CNS) [10], which is based on the Jaccard index, is used to detect the prime communities. The common neighbor (CN) is extensively used in social networks to find common neighbors between any two nodes. Nonetheless, the majority of algorithms like in [11] centered on CN assess common neighbors within one hop only. As social networks are dynamic in

Table 1 Notations used in algorithm

Notation	Description
v_i	Node i
$\Gamma(v_i)$	Set of neighbors of node i
$C(v_i)$	Community of node v_i
$C(v_i)(t)$	Community of node v_i at time t
$CN_t(v_i, v_j)$	Set of common neighbors of nodes v_i and v_j
$v_i.free$	Node i not belonging to any core
$L_t(v_i, v_j)$	Closeness between neighbors v_i and v_j at time t
$D_t(v_i)$	Change in degree of node v_i from time 1 to t
$CNS_t(v_i, v_j)$	Common Neighbor Score of nodes v_i and v_j at time t

nature, and the users (often known as nodes) change degree with time. Here, we use the concept of change in degree defined by Yao et al. [13]. Within a time interval $(1, 2, \ldots, t, \ldots, T)$, the change in degree due to influence by common neighbors is defined as follows:

$$D_t(v_c) = \frac{1}{\sum_{t=2}^{T} d_{t-1,t} / \Delta T},$$

where T and ΔT indicate the present time and the time interval between 1 and T, and $d_{t-1,t}$ represents the Euclidean distance of node v_c from time $t-1$ to t. At time T, node v_c is the common neighbor of node v_i and node v_j (Table 1).

Algorithm 1 Overlapping Multiple-Label Propagation Algorithm

Require: $G(v, e)$

Ensure: Community Structure

1: Initially propagate unique label (node id) to all nodes in network

$$C_{v_1}(0) = v_1, C_{v_2}(0) = v_2, ..., C_{v_n}(0) = v_n$$

2: Sort all nodes in decreasing order of their degree in the node set.
3: Set t=1
4: For each v_i in the node set
5: **if** $D(v_i) \geq 3$ and $v_i.free = true$ **then**
6: find v_j with the largest degree in $\Gamma[v_i]$ **and** $v_j.free = true$
7: **end if**
8: Compute $CN_t(v_i, v_j)$ at time t
9: Find $L_t(v_i, v_j)$ between the nodes in the set $CN_t(v_i, v_j)$
10: Calculate $D_t(v_c)$ of each common neighbor v_c in $CN_t(v_i, v_j)$
11: Compute $CNS_t(v_i, v_j)$ as follows:

$$CNS_t(v_i, v_j) = L_t(v_i, v_j). \sum_{v_m \in CN_t(v_i, v_j)} D_t(v_c)$$

12: **if** two nodes have equal highest CNS **then**
13: assign them to one group.
14: **else** $\{CNS(v_j) > CNS(v_k)\}$
15: then assign label of v_j to v_k.
16: **end if**
17: **if** a node has more than one neighbor with the same highest CNS **then**
18: then the node adopts multiple labels of all such nodes.
19: **end if**
20: **repeat**
21: $\frac{\text{edges between X and Y}}{\text{vertices in Y}} > \mu \times \frac{\text{edges in X}}{\text{vertices in X}}$ {Merge communities X and Y}
22: **until** each node possesses the same label as most of its neighboring nodes
23: **if** no node changes its community further **then**
24: algorithm stops
25: **else**
26: set $t = t + 1$ and **goto** step 3
27: **end if**

The algorithm begins with assigning unique labels to every node in the network. The nodes are then arranged in decreasing order of their degree in the node set. A node v_i with degree (≥3) is chosen randomly from the network, provided that the node is free, i.e., the node is not yet assigned to any community. Then, a free node *j* with the largest degree in the neighborhood of node v_i is chosen. Both nodes *i* and *j* are added to the core. Then, the algorithm finds the common neighbors of nodes v_i and v_j. $L_t(v_i, v_j)$ calculates the closeness between the common neighbors of nodes v_i and v_j at time *t*. In each iteration, the degree $D_t(v_c)$ of each of the common neighbors of these nodes is calculated. The common neighbor score (CNS) of the nodes is calculated, which is computed as follows:

$$CNS_t(v_i, v_j) = L_t(v_i, v_j). \sum_{v_m \in CN_t(v_i, v_j)} D_t(v_c).$$

When two nodes in the network possess same highest CNS, then the nodes are assigned to the same group. But if the CNS of a node is greater than the CNS of another node, then the node with the smaller CNS attains the label of the node with the higher CNS. Also, if a node possesses more than one neighbor with the same highest CNS, then the node acquires more than one label of all such nodes. A situation may arise in a network where many small communities develop, making it more difficult to analyze the networks. In such a case, the algorithm imposes a condition on the formed communities in which smaller communities merge into larger communities, until each node acquires the label possessed by most of its neighboring nodes.

4 Conclusion and Future Work

With the revolution in analyzing social networks, it is quite intuitive to detect communities to gain insights into a network. In this chapter, we have proposed a label propagation-based community detection technique that uses an updating strategy of degree of nodes with time to fit real-world networks. The algorithm detects both disjoint and overlapping communities. Our method is a clear advance on current methods of detecting community structure. Due to the space issue, we aim to implement the algorithm on real-world networks as our future work. To further our research we plan to extend the algorithm for bipartite networks and weighted networks.

Acknowledgements The first author wishes to say thanks to the Council of Scientific and Industrial Research (CSIR) for financial assistance received in the form of JRF.

References

1. Breiger, R.L. 2004. The analysis of social networks. In *Handbook of Data Analysis*, ed. M. Hardy, and A. Bryman, 505–526. London: SAGE Publications.
2. Gregory, S. 2010. Finding overlapping communities in networks by label propagation. *New Journal of Physics* 12 (10): 103018.
3. Havvaei, E., and N. Deo. 2016. A game-theoretic approach for detection of overlapping communities in dynamic complex networks. 1: 113–324. arXiv:1603.00509.
4. Hu, W. 2013. Finding statistically significant communities in networks with weighted label propagation. *Social Network* 2: 138–146.
5. Lancichinetti, A., and S. Fortunato. 2010. Community detection algorithms: A comparative analysis. *Physical Review E* 80 (5): 056117. https://doi.org/10.1103/PhysRevE.80.056117.
6. Newman, M.E., and G. Reinert. 2016. Estimating the number of communities in a network. *Physical Review Letters* 117–125.
7. Newman, M.E., and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69 (2): 026113.
8. Pizzuti, C. 2009. Overlapped community detection in complex networks. In *Proceedings of the 11th Annual Conference On Genetic And Evolutionary Computation*, 859–866, ACM.
9. Raghavan, U.N., R. Albert, and S. Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76: 36106.
10. Steinhaeuser, K., and N.V. Chawla. 2008. Community detection in a large real-world social network. *Social Computing, Behaviora-Cultural Modeling Prediction*, 168–175. Berlin: Springer.
11. Wu, Z.H., Y.F. Lin, S. Gregory, H.Y. Wan, and S.F. Tian. 2012. Balanced multi-label propagation for overlapping community detection in social networks. *Journal of Computer Science Technology* 27: 468–479.
12. Xie, J., and B.K. Szymanski. 2012. Towards linear time overlapping community detection in social networks. In *Advanced Techniques in Knowledge Discovery Data Mining*, ed. P.-N. Tan, S. Chawla, C.K. Ho, and J. Bailey, 25–36. Berlin: Springer.
13. Yao, L., L. Wang, L. Pan, and K. Yao. 2016. Link prediction based on common-neighbors for dynamic social network. *Procedia Computer Science* 83: 82–89.
14. Zhou, L., K. Lu, P. Yang, L. Wang, and B. Kong. 2015. An approach for overlapping and hierarchical community detection in social networks based on coalition formation game theory. *Expert Systems with Applications* 42: 9634–9646.