A Simple and Effective Community Detection Method Combining Network Topology with Node Attributes

Dongxiao He, Yue Song, and Di Jin(✉)

College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
{hedongxiao,sy717389667,jindi}@tju.edu.cn

Abstract. Community detection is a fundamental problem in the study of complex networks. So far, extensive approaches, which use network topology alone or use both network topology and attribute information, have been designed to detect the community partitions of networks. However, existing approaches cannot work effectively for networks whose community structure does not match well with the ground-truth, or networks whose topological information contains serious noise, and networks where the difference of attribute similarity between nodes is tiny. Inspired by a force-directed network layout and community intuitive characteristics, we propose a simple while effective approach which utilizes attribute information to partition nodes into communities by maximizing network modularity. By using attributes as nodes to the network and the interaction between nodes, our novel method cannot only effectively improve community detection of networks, but also obtain the number of communities closer to the real one. Through extensive experiments on some real-world datasets, we demonstrate the superior performance of the new approach over some state-of-the-art approaches.

Keywords: Community detection · Modularity · Community structure

1 Introduction

Many complex systems can be abstracted as networks which consist of nodes and edges between nodes in modern network science. Complex networks are usually organized in communities, which have their own role and/or function such as groups of related individuals in social networks, functional modules of protein-protein interaction networks, etc. [6]. Discovering communities is useful for understanding what role they play in the network structure and dynamics, and has many real applications [7].

The primary objective of community detection is to partition the network into communities which correspond to functional modules composed of nodes. Generally, two types of information can be exploited for community detection. They are the network topology and attribute information on nodes. Therefore,

C. Douligeris et al. (Eds.): KSEM 2019, LNAI 11775, pp. 168–175, 2019.
https://doi.org/10.1007/978-3-030-29551-6_15

a wide variety of community detection methods based on different theories and technologies had been proposed recently [6–8,12]. Topology-oriented community detection methods can be mainly categorized into two types: measure-based methods [1,2,16,17] and probabilistic model based methods [10,13]. However, some of the real-world networks do not conform to the community structure property, and there is a lot of noise between nodes. Therefore, the above methods using network topology alone cannot work effectively in many cases. In order to improve the quality of community partitions detected, many existing methods take node attributes into account to supplement network topology information [3–5,15]. However, these methods usually need to manually adjust the effect between topology and attributes on community detection, which is typically difficult to achieve, and the similarity between nodes (calculated in these methods) in some networks are not significant. So these methods can not effectively play a role. Although some models [9,20] have been proposed, which do not need to adjust the parameters between two types of information, they cannot effectively partition nodes into communities for the network with serious noise.

Force-directed layout is a widely used technology in network layout. Its very essence is to turn structural proximities into visual proximities. Noack [18] has shown that the proximities express communities and proposes that layouts with optimal energy are consistent with clusterings with optimal modularity. The layout and clusterings of the whole network can be affected by adding or deleting nodes and edges because the structural proximity has been changed in the network. Since adding or deleting nodes and edges to the network can affect structural proximity, it may be ideal to change the community structure for community detection. But how to effectively add nodes and edges to the network for community detection is still a challenge.

To address these problems, inspired by a force-directed network layout and community intuitive characteristics, in this paper we propose a novel method with an augmented graph for community detection. Different from existing topology and attributes based approaches that using node attributes to calculate similarity, our new approach uses them as nodes to help detect communities of networks. In this way, our method not only avoids the problem that the quality of community partitions cannot be effectively improved due to the tiny difference between node attribute similarity, but also improves community structure of the network to obtain more accurate partitioning results through the interaction between nodes. We can then obtain community partition of the original network from the augmented graph using a fast and effective algorithm based on modularity optimization, e.g. Louvain. We present extensive experimental results on some real-world attributed networks to show that the new method is effective and outperforms some state-of-the-art approaches for community detection.

2 The Method

Assuming that an undirected and attributed network G is defined as a 3-tuple (V, E, X), where $V = \{v_1, v_2, \ldots, v_N\}$ is a set of N nodes, $E = \{(u, v) : u, v \in V, u \neq v\}$ is a set of e edges, which can be represented by an $N \times N$ adjacent matrix A with elements $A_{uv} = 1$ if u and v are connected, or 0 otherwise, and $X = \{x_1, x_2, \ldots, x_M\}$ is a set of M attributes, which is also binary matrix and its value is 1 if the node has the attribute, or 0 otherwise, then community detection is to partition the node set V of network G into K disjoint communities $P = \{C_1, C_2, \ldots, C_K\}$.

2.1 Motivations

Most of community detection methods mainly focus on detecting communities using network topology, and the premise is that the functional communities are consistent with community structure of networks. However, some of the real-world networks do not meet this property, and there is also typically serious noise in network topology. As shown in Table 1, we use a widely used metric modularity [17] to assess the community structure of seven widely used attributed networks [14]. We use Louvain method [1] to divide seven networks into communities to get the modularity values, and then compare it with the modularity values of the ground-truth communities. We can see that some of the real-world networks do not correspond well to what we usually say. In Table 1, it is obvious that the connections between communities are more denser for the first four networks, and the network structures contain serious noise for the latter three networks. It is not uncommon for two disconnected nodes to belong to the same community, and it is difficult to correctly divide a node connected to multiple communities for various reasons to the correct community by only relying on the topological structure. Therefore, the traditional methods based on network topology, such as modularity-based methods, cannot work effectively in this case.

Table 1. Datasets descriptions and comparison between modularity values obtained by Louvain method and ground-truth communities. Q is the Modularity value, which is between -1 and 1. The larger the value, the denser the connections within the community. N is the number of nodes, E the number of edges, M the number of attributes, K the real number of communities.

Metric	Methods	Datasets						
		Cornell	Texas	Washington	Wsicsonsin	Cora	Citeseer	UAI2010
Q	Louvain	0.647	0.552	0.561	0.639	0.809	0.883	0.455
	Ground-truth	−0.154	−0.215	−0.136	−0.118	0.640	0.543	0.258
N		195	183	217	262	2,708	3,327	3,067
E		283	276	366	459	5,278	4,676	28,308
M		1,588	1,498	1,578	1,623	1,433	3,703	4,973
K		5	5	5	5	7	6	19

As shown, the network topology reflects only one aspect of networks and usually contains noise. Fortunately, node attributes can be used to improve the results of community detection. However, this is also technically challenging to

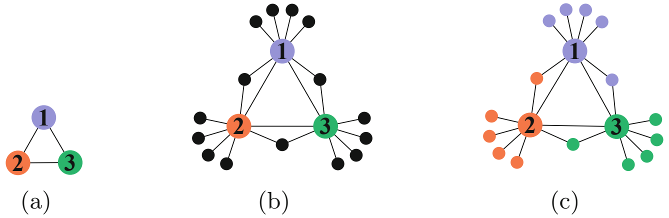


Fig. 1. Work flow of the proposed method. (a) is the original network, (b) the augmented network, and (c) the community partitions obtained by Louvain. We use large nodes to denote the original nodes in the original network and small nodes the new attribute nodes. Numbers are the ordinal numbers of nodes in the original network. Colors represent different communities. It is drawn using ForceAtlas2 [11].

effectively combine these two kinds of available information. Existing methods typically deal with network topology and node attributes separately. As a result, they need to adjust the impact of two types of information on community detection, which is difficult to achieve in real life. And most of the methods only deal with attribute information by calculating attribute similarity between nodes. However, there is no significant difference between the attribute similarity of a pair of nodes within community and that across communities in some networks in many cases. As shown in Fig. 1, Fig. 1(a) is a network with three nodes. We use the topology-based approach, such as modularity-based approach, to partition three nodes into the same community. When we improve the community partitioning result by calculating the node's attribute similarity (e.g., using cosine similarity), it does not play an effective role because the pairs of nodes have similar similarities. (The black nodes are attributes, and edges indicate that nodes have these attributes in Fig. 1(b).) Therefore, it still is not enough for community detection methods to help partition communities by calculating attribute similarity as existing methods done.

2.2 Augmented Graph Based Method

Here we develop a simple and effective community detection method that combines network topology and node attribute information by an augmented graph. The initial topological structure of the original network is augmented by new nodes called attribute nodes and new edges called attribute edges. Our method is mainly based on the following intuitive properties:

- Nodes that connected to each other are more likely to belong to the same community.
- Nodes sharing common attributes are more likely to be in the same community.
- Even if two nodes are not interconnected, they may belong to the same community through the interaction between original nodes and the interaction between original nodes and attribute nodes.

– Even if both the topological structure and attribute information contain a lot of noise, communities can be found well through the implicit interaction of these two kinds of nodes.

Algorithm 1. Augmented graph based approach.

```
Input:  $G(V, E, X)$: undirected and attribute network. $clust$: Louvain method.
Output:  A partition $P = \{C_1, C_2, \ldots, C_K\}$ of $V$.
$V_M = \{v_1, v_2, \ldots, v_M\}$ with $M = |X|$
$E_M \leftarrow \emptyset$
for $n = 1$ to $|V|$ do
    for $m = 1$ to $|X|$ do
        if $X_{nm} = 1$ then
            $E_M \leftarrow E_M \cup (v_n, v_m)$
        end if
    end for
end for
$V' \leftarrow V \cup V_M$ and $E' = E \cup E_M$ with $E_M \subseteq V \times V_M$
$G' = (V', E')$
$P' \leftarrow clust\,(G')$
$P \subset P'$        \\ $P$ is the partition of the set $V$
return  $P$
```

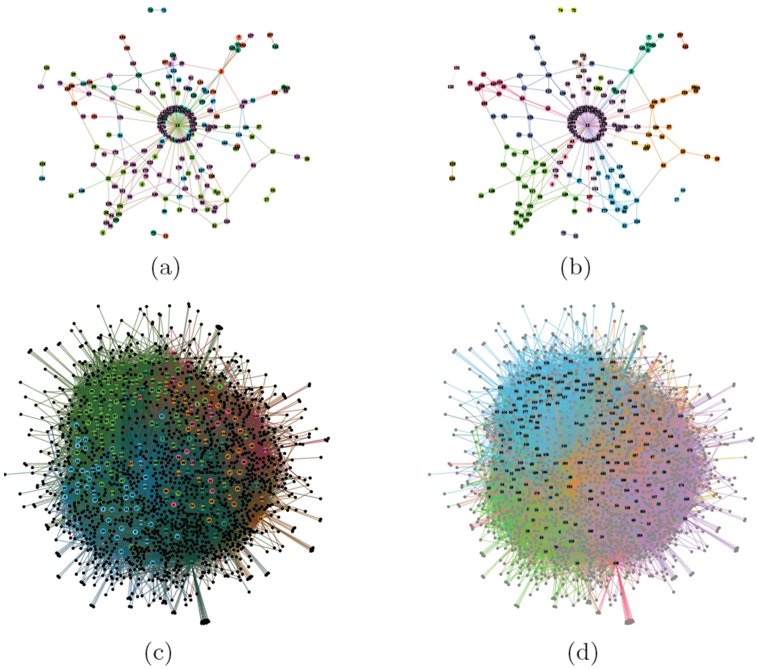
The process of the proposed new method is shown in Algorithm 1. We formally describe the work process of our method as follows. A set of attribute nodes is $V_M = \{v_1, v_2, \ldots, v_M\}$ and an attribute edge is added between node u and attribute node v_m if node u has the m-th attribute (As shown in Fig. 1(b)). There are totally M attribute nodes and $\sum_{u \in V, m \in V_M} X_{um}$ attribute edges added to the original network. In the augmented graph, two nodes are close if they are connected through many other original nodes, or if they share many common attribute nodes as neighbors. Once the augmented graph is created, we can apply some existing community detection methods, e.g., Louvain algorithm [1] based on modularity, to detect communities in the network (As shown in Fig. 1(c)).

3 Experiments

We now analyze why the new method (AUG for short) works. And then we compare it with several state-of-the-art community detection methods for evaluation. The datasets used are shown in Table 1.

3.1 Why Our Works

To validate the new method, we illustrate the original network and the augmented network of Cornell and their community partitions in Fig. 2. For the original network of the Cornell dataset in Fig. 2(a), the community structure derived is not consistent with ground-truth (i.e., the nodes within communities are sparsely connected, while ones between communities are densely connected), which results in a large number of nodes being misclassified (Fig. 2(b)). In comparison, AUG correctly assign most nodes to correct communities (Fig. 2(d)). This is because AUG uses attributes as nodes and edges which indeed improves community results (Fig. 2(c)). This validates that adding attributes as nodes and edges to the network can effectively complement the community characteristics into the network topology so as to find better communities.



(a)

(b)

(c)

(d)

Fig. 2. An example on Cornell. (a) is the original network and (c) the augmented network. In (a) and (b), the left (right) is the real partition (predicted partition). It is drawn the same as Fig. 1.

3.2 Comparison with Existing Methods

We compared the new method with two types of the state-of-the-art community detection methods. The first, including DCSBM [13] and NetMRF [10], uses network topology alone. The second includes PCLDC [20], SCI [19] and NEMBP [9], which use both topological and attribute information. All the methods compared require the number of communities to be specified, while our method does not. Here, we set their number of communities to the ground-truth, and ran these methods with their default parameters. We used Accuracy (AC) and Normalized Mutual Information (NMI) [7] as metrics for performance evaluation.

Table 2. Comparison of the results of different community detection methods in terms of AC and NMI. ORG is the method that original networks are partitioned by using Louvain. K' is the number of communities detected. Bold font denotes the best results.

Metrics	Methods	Datasets						
		Cornell	Texas	Washington	Wsicsonsin	Cora	Citeseer	UAI2010
AC (%)	ORG	24.1	28.42	23.5	23.66	41.51	20.29	32.21
	DCSBM	37.9	48.1	31.8	32.8	38.5	26.6	2.6
	NetMRF	31.8	30.6	35	28.6	58.1	22.2	31.1
	PCLDC	30.3	38.8	30	30.2	34.1	24.9	28.8
	SCI	36.9	49.7	46.1	46.4	41.7	34.4	29.5
	NEMBP	47.2	53.6	42.9	63.4	57.6	49.5	**46.3**
	AUG	**54.87**	**57.92**	**62.21**	**71.37**	**61.26**	**58.25**	38.93
NMI (%)	ORG	14.35	8.19	10.22	8.83	42.58	32.41	28.18
	DCSBM	9.7	16.6	9.9	3.1	17.1	4.1	31.2
	NetMRF	7.3	5.5	5.8	3.2	37.2	1.2	25.8
	PCLDC	7.2	10.4	5.7	5	17.5	3	26.9
	SCI	6.8	12.5	6.8	13.3	17.8	9.2	23.4
	NEMBP	18.7	**35.1**	21.2	38	**44.1**	24.3	**47.2**
	AUG	**26.32**	30.34	**36.88**	**41.61**	42.83	**32.45**	35.32
K'	ORG	19	12	15	14	97	461	19
	AUG	6	6	7	5	10	9	6

As shown in Table 2, AUG is far superior to ORG, confirming that the new method can effectively improve the quality of community structure detected by adding attributes as nodes to the original network. Moreover, compared with ORG, AUG can find the number of communities (Table 2) which is much closer to that of ground truth communities (Table 1). Moreover, compared with existing methods, AUG has the best performance on 6 and 4 of 7 networks in terms of AC and NMI, respectively. Generally, the methods that use both topology and attribute information perform better than those using topology alone. We like to highlight that among the methods using both two types of information, our AUG is on average 26.8%, 17.2%, and 6.3% more accurate than PCLDC, SCI and NMEBP in AC; and 24.3%, 22.3% and 2.5% more accurate than PCLDC, SCI and NMEBP in NMI. These results further validates the superiority of our new approach over others in finding communities.

4 Conclusion

In this paper, we proposed an efficient and simple method for community detection by adding attributes as nodes to the original network. It is able to make fully use of network topology and attribute information especially when the topology contains serious noise. The extensive experimental results demonstrated that our method outperformed the state-of-the-art approaches for community detection and validated the superior performance of our method on 7 real-world networks.

Acknowledgments. This work was supported by Natural Science Foundation of China (61876128, 61772361).

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech-Theory E **P10008**, 1–12 (2008)
2. Chakraborty, T., Dalmia, A., Mukherjee, A., Ganguly, N.: Metrics for community analysis: a survey. ACM Comput. Surv. **50**(4), 54:1–54:37 (2017)
3. Combe, D., Largeron, C., Géry, M., Egyed-Zsigmond, E.: I-Louvain: an attributed graph clustering method. In: Fromont, E., De Bie, T., van Leeuwen, M. (eds.) IDA 2015. LNCS, vol. 9385, pp. 181–192. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24465-5_16
4. Falih, I., Grozavu, N., Kanawati, R., Bennani, Y.: Anca : Attributed network clustering algorithm. In: Complex Networks, pp. 241–252 (2018)
5. Falih, I., Grozavu, N., Kanawati, R., Bennani, Y.: Community detection in attributed network. In: Proceedings of the WWW, pp. 1299–1306 (2018)
6. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
7. Fortunato, S., Hric, D.: Community detection in networks: a user guide. Phys. Rep. **659**, 1–44 (2016)
8. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
9. He, D., Feng, Z., Jin, D., Wang, X., Zhang, W.: Joint identification of network communities and semantics via integrative modeling of network topologies and node contents. In: Proceedings of the AAAI, pp. 116–124 (2017)
10. He, D., You, X., Feng, Z., Jin, D., Yang, X., Zhang, W.: A network-specific Markov random field approach to community detection. In: Proceedings of the AAAI, pp. 306–313 (2018)
11. Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PLoS One **9**(6), e98679 (2014)
12. Jin, D., Wang, H., Dang, J., He, D., Zhang, W.: Detect overlapping communities via ranking node popularities. In: Proceedings of the AAAI, pp. 172–178 (2016)
13. Karrer, B., Newman, M.E.J.: Stochastic blockmodels and community structure in networks. Phys. Rev. E **83**, 016107 (2011)
14. Leskovec, J.: Stanford network analysis project (2016). http://snap.standford.edu
15. Muslim, N.: A combination approach to community detection in social networks by utilizing structural and attribute data. Soc. Networking **05**, 11–15 (2016)
16. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**, 8577–8582 (2006)
17. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
18. Noack, A.: Modularity clustering is force-directed layout. Phys. Rev. E **79**, 026102 (2009)
19. Wang, X., Jin, D., Cao, X., Yang, L., Zhang, W.: Semantic community identification in large attribute networks. In: Proceedings of the AAAI, pp. 265–271 (2016)
20. Yang, T., Jin, R., Chi, Y., Zhu, S.: Combining link and content for community detection: a discriminative approach. In: Proceedings of the ACM SIGKDD, pp. 927–936 (2009)