2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE) | 978-1-6654-2186-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAICE54393.2021.00068

A Graph Convolutional Network-Based Approach for Community Detection in Attributed Networks

Zaisheng Wang
China Jiliang University
Hangzhou, China
wzsxuexi@163.com

Xiaofeng Wang*
China Jiliang University
Hangzhou, China
* Corresponding author: xfwang@cjlu.edu.cn

Guodong Shen
China Jiliang University
Hangzhou, China
2316486500@qq.com

Zengjie Zhang
China Jiliang University
Hangzhou, China
1256389270@qq.com

Daying Quan
China Jiliang University
Hangzhou, China
qdy@cjlu.edu.cn

Jianhua Li
Shanghai Jiao Tong Univerity
Shanghai, China
lijh888@sjtu.edu.cn

***Abstract*—Community detection has attracted widespread attention since it helps reveal geometric structures and latent functions of complex networks. Recently community detection has been revisited with the development of network representation learning, many approaches have been presented, including graph convolutional network (GCN) based methods. Existing GCN-based community detection methods usually rely on a considerable number of prior labels to infer unknown nodes. To address this problem, we propose a new GCN-based method for community detection in attributed networks without any label information. Based on the local self-organization characteristics of the communities, we integrate a label sampling model and the shallow GCN architecture into an unsupervised learning framework, the former helps construct a balanced training set via a local expansion strategy to train GCN. Moreover, we reveal the underlying community structures by fusing topology and attribute information. Experimental results on several real-world networks indicate our method is effective compared with the state-of-the-art community detection algorithms.**

Keywords-Community detection; Attributed networks; Graph convolutional network; Label sampling

I. INTRODUCTION

In recent years, the network has become a general form of describing and modeling a variety of complex systems [1], such as physical systems, social networks, and biological networks. Network structures have been studied extensively under the notions of subgraphs, groups, and communities. Community detection or identification is to divide nodes in a network into groups where nodes are densely connected whereas sparely linked between groups. Community detection has attracted widespread attention since it helps reveal geometric structures and latent functions of complex networks. Many methods for community detection have been proposed, a majority of which are based on the information of network topological structures, such as hierarchical clustering, modularity optimization, statistical inference, and dynamic modeling [2]. Moreover, some methods were proposed to make use of node semantics or node attributes in addition to network topology to improve results, such as matrix factorization and multi-objective optimization [3].

As complex networks become increasingly large-scale and evolve dynamically, network data from multiple sources must be effectively integrated to identify more meaningful communities. As a result, traditional methods become difficult to tackle such complex network data due to their high computational cost and limited data fusion capabilities. More recently, the technique of network representation learning (NRL) was adopted to learn a low-dimensional representation of network structures from the high dimension network data, and many efforts were made to develop scalable and effective NRL technology directly designed for complex networks [4]. Many NRL methods have been proposed to learn node representations [5], such as random walk-based methods (e.g., DeepWalk [6], LINE [7] and struc2vec [8]), and Matrix Factorization based methods (e.g., M-NMF [3] And TADW [9]) and deep learning-based methods (e.g., DNGR [10] and SDNE [11]).

Community detection has been revisited with the development of network representation learning, and many approaches have been presented [12]. Recently, graph convolutional networks (GCN) have been introduced to address the problem of community detection on graphs [13]. GCN can effectively extract complex features from network topology and node attributes through a stack of convolution operations like CNN [13]. Jin et al. [14] proposed to solve semi-supervised community detection by integrating statistical modeling of GCN and Markov random field. Sun et al. [15] developed a network embedding framework with graph convolutional autoencoder to learn node representations for clustering tasks. However, existing GCN-based community detection methods usually rely on a considerable number of prior labels to infer unknown nodes.

978-1-6654-2186-7/21/$31.00 ©2021 IEEE
DOI 10.1109/ICAICE54393.2021.00068

In this work, we propose an unsupervised learning method for community detection in attribute networks based on GCN, which can solve the shortcoming of the original GCN model. We integrate the shallow GCN architecture and a label sampling model into an unsupervised learning framework, to reveal the underlying communities in the attribute network. Without any prior knowledge about community structure, a balanced label set with a small number of nodes is constructed by the local label sampling model, which makes GCN training more effective.

The rest of this article is organized as follows. In Section 2, we introduce related work briefly. Section 3 describes in detail the proposed method for community detection. We show experimental results on various real-world networks to verify the performance of the method proposed in Section 4. Finally, the conclusions are summarized in Section 5.

II. RELATED WORK

A. Community Detection

Community detection in complex networks is one of the most important themes of modern network science. Its purpose is to discover the underlying groups or vertex clusters in the network. In the past ten years, many community detection methods based on various similarity measures between nodes have been proposed. For example, the hierarchical clustering method divides graphs or aggregates nodes according to the similarity measure between nodes [16]. The optimization method based on modularization transforms the task of community identification into the problem of maximizing the modular function of community structure [17]. Meanwhile, many methods have been proposed to detect disjoint or overlapping communities more flexibly [2]. For example, the centrality sampling algorithm is based on various centrality measures (such as degree centrality, proximity centrality, etc.) to identify important nodes [18]. However, these methods only use the network topology to identify community structures through greedy optimization and cannot achieve higher accuracy requirements.

B. Network Representation Learning

Network representation learning is a new paradigm of learning, which has been proved to be effective in network analysis in recent years [4]. Recently, graph convolutional networks (GCN) have been introduced in network analysis tasks [19,20]. However, although the GCN-based method is helpful to find network topology and attribute information, it needs to provide a large number of node labels as a prerequisite. Therefore, some unsupervised methods have been proposed. Jin et al. [21] proposed an unsupervised model for community detection through GCN embedding. He et al. [22] developed a community-centric GCN model for unsupervised community detection. In this paper, by integrating the label sampling model and GCN into an unsupervised learning framework, we propose a new GCN-based approach for community detection.

III. METHODOLOGY

In the semi-supervised learning problem of complex networks, the graph convolutional network naturally integrates node attributes and topological information of a network via convolution operation. However, a large number of labels are required to train in the GCN model, because the graph convolution operation is essentially a local filter that needs a considerable number of labeled nodes for verification and model selection. In fact, even with a large number of training labels, GCN may not be able to effectively propagate label information to the global network, because most training nodes may be topologically close to each other and far away from the cluster center. In order to solve this problem, we use the label sampling model to construct a training set of planting labels as a supplementary model of GCN.

A. Local Label Sampling

Due to the uneven distribution of labeled nodes in the network, more training labels are required to train the multi-layer GCN model, to effectively propagate label information to the global network. We propose to construct a balanced label set by a label sampling model, which can optimize the GCN model on a small number of labels without additional labeled nodes for validation. We first employ the structure center location method to find the structural centers which usually distribute in different clusters and assign them unique labels. Then we add them to the training set as the initial labels. Taking the Karate network as an example, we identify the structural centers of two groups in the network, as shown by the color nodes in the right subplot in Figure 1.

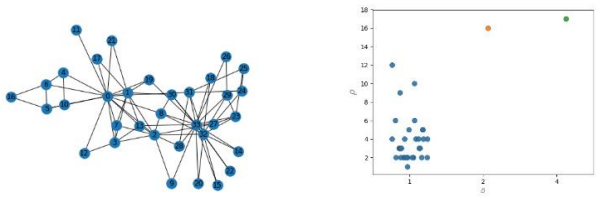

FIG 1. Structural center location on the Karate network

In order to accelerate the convergence of the GCN model, the nearest neighbors of structural centers are sampled. The sampling model we use is based on structural centrality, which measures the local density of nodes and the relative distance between nodes. Therefore, structural centers generally have high centrality and are distributed in different groups. In addition, the structure center is closely connected with other core nodes and peripheral nodes, which greatly improves the efficiency of label dissemination in communities.

B. Community Detection via GCN

Given an attributed network $G = (V, E, W)$ over a node-set $V = \{v_1, v_2, ..., v_N\}$ with $|V| = N$, an edge set E with $e_{ij} = (v_i, v_j) \in E$, and a set of node attributes $W = \{w_1, w_2, ..., w_m\}$ with M dimensions, a labeling $l: V \rightarrow \{1, ..., C\}$ that denotes a partition P of all nodes into C communities is to be predicted for the network. For simple, we consider an undirected network G specified by an adjacent matrix $A = (a_{ij})_{N \times N}$ ($a_{ij} = 1$ if $e_{ij} \in E$, or 0 otherwise) which encodes structural connectivity over all nodes, and an attribute matrix $X \in R^{N \times M}$ which includes integrated attributes for all nodes. The resulting community structures in the attributed network generally show obvious local clustering characteristics in topology and dissimilarity in attributes.

Since Graph convolution is be proved essentially a Laplacian smoothing, multi-layer GCN model will reduce the clustering accuracy. Therefore, we use shallow GCN as the basic model of our unsupervised learning framework. The final output of the two-layer shallow GCN model is formalized as follows:

$$Z = \mathrm{softmax}(\tilde{A}\,\mathrm{ReLU}(\tilde{A}XW^{(0)})W^{(1)}) \quad (1)$$

where $\tilde{A}$ denote the normalized adjacency matrix by $\tilde{A} = D^{-1/2}\tilde{A}D^{-1/2}$, $\mathrm{softmax}(\cdot)$ and $\mathrm{ReLU}(\cdot)$ are two activation functions defined by $\mathrm{softmax}(x_i) = \frac{1}{\vartheta}\exp(x_i)$ with $\vartheta = \sum_i \exp(x_i)$ and $\mathrm{ReLU}(x) = \max(0, x)$, respectively. $W^{(0)}$ and $W^{(1)}$ are weight matrices that can be trained by gradient descent. Adam optimizer is used to train the GCN model. Here, we adopt the cross-entropy error over all labeled nodes to evaluate the loss:

$$Loss = -\sum_{i \in V_l}\sum_{f=1}^{F} Y_{if} \ln Z_{if} \quad (2)$$

where V_l denotes the set of indices corresponding to the labeled nodes, $Y \in R^{|V_l| \times F}$ is the prior community label indicator matrix, and F is the dimension of the output features which is equal to the number of communities. Then, we can train the GCN model with the constructed training set.

The difference from the original GCN model is that our GCN-based community detection model is used for clustering tasks and not for node classification. Moreover, our method directly optimizes the GCN on the balanced training set constructed by the label sampling process without additional data for verification. In order to meet the dense connectivity of the community structure, we only consider the connected components of the graph in the graph convolution.

C. Overview

The proposed method integrates the label sampling model and shallow GCN into an unsupervised learning framework to reveal the underlying community structure with topological and attribute information. It does not rely on prior label information and at the same time accelerates the convergence of the GCN model. Given the network *G = (V, E, X)*, we calculate the complexity from the two stages of label sampling and GCN training. In the experiment, the running time of the label sampling phase is negligible on a network with thousands of nodes. In the GCN training phase, we follow the settings of Kinf and Welling's and use full batch gradient descent, which scales linearly with the number of network edges. Therefore, the time complexity of the proposed method is comparable to that of GCN.

IV. EXPERIMENTS

A. Datasets and Baseline Methods

We experimented on 7 different types of real-world networks. The basic information of these data sets is shown in Table 1. The datasets Cornell, Texas, Washington, and Wisconsin come from Wang et al. [23], which describe the connections of webpages generated by college students from several universities. In addition, the three citation networks Cora, CiteSeer, and PubMed are from Kipf and Welling [20], and they are usually used for testing in network analysis.

We compare our method with various state-of-the-art community detection methods. The baseline methods include three types. The other includes two graph embedding methods DeepWalk [6] and MGAE [17] which both use graph neural networks for graph representation and a clustering algorithm for node clustering. The second type includes SCI [24] and NEM [25], which identify communities through network topology and attribute information. The above methods all perform node clustering in an unsupervised manner, and all nodes are considered in the performance evaluation. The last type includes the original GCN [20], DIG [26], and WSC [23], which are semi-supervised methods on attribute networks and usually use test sets to evaluate their performance. Moreover, we use a combination method (SCL+LEO) based on structure center location (SCL) and local expansion optimization (LEO) [27] for community detection in attributed networks, in order to reflect that GCN can obtain better performance based on the label sampling mechanism.

TABLE I. THE BASIC INFORMATION ON THE REAL-WORLD NETWORKS

Dataset	Nodes	Edges	<k>	cc	Communities	Attributes
Cornell	195	286	2.93	0.16	5	1703
Texas	187	328	3.19	0.19	5	1703
Washington	230	446	3.63	0.20	5	1703
Wisconsin	265	530	3.62	0.21	5	1703
Cora	2708	5429	3.89	0.24	7	1433
Citeseer	3312	4732	2.81	0.14	6	3703
PubMed	19729	44338	4.49	0.06	3	500

In this table, <k> is the average degree of nodes, ''cc'' denotes the average clustering coefficient of the network, ''Attributes'' represents the dimension of attribute features information.

B. Evaluation Criteria

Community detection is essentially a network clustering problem, which divides nodes into different clusters based on topology and attribute information. In our experiment, the clustering accuracy (ACC) method is used to evaluate the performance of various algorithms. ACC measures the proportion of correctly clustered instances in total samples. In addition，the normalized mutual information (NMI) [2] is used in our experiment to evaluate the effectiveness of each method in community detection. The NMI index is a metric based on information theory and is widely used to measure the similarity

between the ground-truth community partition of a network and the detected partition from an algorithm.

C. Experimental Setups

In our experiment, for the citation networks including CiteSeer, Cora, and PubMed, we use the datasets provided by Yang et al. In training, we take a 2-layer GCN with randomly initialized weights and follow the same hyper-parameters as Kipf and Welling with learning rate 0.01, maximum epochs 200, dropout rate 0.5, *L2* regularization weight 5×10^{-4}, and 16 hidden units. For each run, we randomly divide labels into a small set for training, a large set for testing. We adopt Adam optimizer to train the GCN-based models and run experiments on TensorFlow. To ensure the stability of results, the proposed method and all compared algorithms have been independently run 10 times on each network. All the experiments are conducted on a PC with a 2.8 GHz, i7-7700H, Quad-core CPU, and 16 GB of RAM.

D. Experiment Results

In this section, we evaluate the effectiveness of the proposed method (namely LSGCN). We also compare the performance of LSGCN with other state-of-the-art algorithms on 7 real-world networks. The experimental results are shown in Table 2 and Table 3. In the experiment, the number of labeled nodes required by our method is far less than that required by other GCN-based algorithms to achieve comparable performance. In the experiment, we follow the criteria for choosing the number of neighbors proposed by Li et al. [27] during the label sampling. Given a network with the average degree $<k>$ and a GCN with the number of layers τ, the lower bound of t is determined by $<k>^{\tau} * t \approx N$. Its basis is to estimate how many labels are needed for a GCN model to propagate them to cover the entire graph.

In terms of clustering accuracy (ACC), LSGCN performs better than other baseline methods on these network datasets. The experimental comparison shows that the methods that use both topological structures and attribute information have better performance than the methods based only on network topological information. Based on the label sampling at the center of the structure, the accuracy of LSGCN is 9.7% and 6.1% higher than that of GCN and DIG respectively. This validates our analysis that based on the constructed small-scale training set, the GCN model can effectively propagate the labels to the entire graph. It can also be seen from the table that LSGCN has an average increase of 39.4% compared with SCL+LEO. The comparison shows that the community detection task of LSGCN on the attribute network can be performed higher.

On the other hand, we evaluate the NMI scores to evaluate the effectiveness of each method in community detection. As shown in Figure 3, The method LSMGCN outperforms other networks on 6 real-world networks, which further verifies the effectiveness of the local label sampling strategy. However, compared with the clustering accuracy (ACC), the NMI value is mostly between 40% and 70%, showing the difference in the performance of NMI in terms of scores. Through the analysis of these networks, it is found that the performance deviation of the NMI value is mainly caused by the difference in degree distribution and clustering coefficient. As shown in Table 1, these networks with low distribution and low clustering coefficients, such as Cornell, Citeseer, and PubMed, tend to perform poorly on NMI values.

TABLE II. COMPARISON OF ALGORITHMS ON REAL-WORLD NETWORKS IN TERMS OF CLUSTERING ACCURACY(ACC)

Dataset	SCL+LEO	SCI	NEM	DeepWalk	MGAE	GCN	DIG	WSC	LSGCN
Cornell	0.168	0.369	0.472	0.318	0.482	0.463	0.484	0.539	0.625
Texas	0.205	0.497	0.536	0.326	0.567	0.571	0.623	**0.775**	0.674
Washington	0.157	0.461	0.429	0.350	0.508	0.549	0.565	0.583	**0.643**
Wisconsin	0.166	0.464	0.634	0.287	0.588	0.556	0.577	0.619	**0.656**
Cora	0.073	0.417	0.576	0.467	0.634	0.815	0.817	0.537	**0.875**
Citeseer	0.060	0.344	0.495	0.362	0.636	0.703	0.712	0.476	**0.759**
PubMed	0.098	0.473	0.657	0.619	0.439	0.790	0.792	0.607	**0.812**
Avg	0.400	0.556	0.640	0.489	0.655	0.697	0.733	0.702	**0.794**

The proposed method is denoted by LSGCN for short. The bold value in each row denotes the best result on the corresponding dataset.

TABLE III. COMPARISON OF ALGORITHMS ON REAL-WORLD NETWORKS IN TERMS OF PARTITION SIMILARITY (NMI)

Dataset	SCL+LEO	SCI	NEM	DeepWalk	MGAE	GCN	DIG	WSC	LSGCN
Cornell	0.162	0.152	0.187	0.073	0.482	0.091	0.121	0.216	**0.337**
Texas	0.207	0.220	0.351	0.056	0.567	0.050	0.054	**0.537**	0.485
Washington	0.188	0.210	0.212	0.059	0.508	0.163	0.252	0.241	**0.403**
Wisconsin	0.198	0.185	0.380	0.043	0.588	0.178	0.257	0.314	**0.446**
Cora	0.425	0.178	0.441	0.327	0.634	0.545	**0.625**	0.525	0.614
Citeseer	0.313	0.092	0.243	0.097	0.636	0.423	0.454	0.353	**0.515**
PubMed	0.216	0.283	0.319	0.167	0.439	0.260	0.380	0.342	**0.408**
Avg	0.379	0.343	0.432	0.310	0.655	0.410	0.482	0.487	**0.583**

V. CONCLUSION

In this work, we propose a community detection method (LSGCN) based on graph convolutional network in attributed networks. Our method employs a structural center location approach and local expansion strategy to construct a balanced training set via introducing the label sampling model. Moreover, to reveals the underlying community structure by fusing topology and attribute information, we integrate the label sampling model and the shallow GCN model into an unsupervised learning framework. Experimental results on real-world networks demonstrate the validity of our method and perform better performance over state-of-the-art methods. In future work, we intend to combine LSGCN with the attention mechanism to improve clustering accuracy.

ACKNOWLEDGMENT

This research was supported by National Key Research and Development Program of China (2019YFB1707104), Natural Science Foundation of Zhejiang Province (LQ20F020021), and NSAF Joint Fund (U20B2048).

REFERENCES

[1] Newman, M. (2018) Networks. Oxford University Press.

[2] Fortunato, S., Hric, D. (2016) Community detection in networks: A user guide. Physics reports, 659, 1-44.

[3] Wang, X., Cui, P., Wang, et al. (2017) Community preserving network embedding. In: Proceedings of AAAI, pp. 203–209.

[4] Zhang, D., Jie, Y., Zhu, X., et al. (2017) Network representation learning: a survey. IEEE Trans. Big. 3–28.

[5] Zhu, W., Wang, X., Cui, P. (2020) Deep learning for learning graph representations. In: W. Pedrycz, Shyi-Ming Chen (Eds.), Deep Learning: Concepts and Architectures, Springer, pp. 169–210.

[6] Perozzi, B., Al-Rfou, R., Skiena, S. (2014) Deepwalk: Online learning of social representations. In: Proceedings of SIGKDD, pp. 701–710.

[7] Tang, J., Qu, M., Wang, M., et al. (2015) Line: Large-scale information network embedding. In: Proceedings of WWW, pp. 1067–1077.

[8] Ribeiro, L.F., Saverese, P.H., Figueiredo, D.R. (2017) struc2vec: Learning node representations from structural identity. In: Proceedings of SIGKDD, pp. 385–394.

[9] Yang, C., Liu, Z., Zhao, D., et al. (2015) Network representation learning with rich text information. In: Proceedings of IJCAI, pp. 2111–2117.

[10] Cao, S., Lu, W., Xu, Q. (2016) Deep neural networks for learning graph representations. In: Proceedings of AAAI, pp. 1145–1152.

[11] Wang, D., Cui, P., Zhu, W. (2016) Structural deep network embedding. In: Proceedings of SIGKDD, pp. 1225–1234.

[12] Jin, D., Yu, Z., Jiao, P., et al. (2021) A survey of community detection approaches: From statistical modeling to deep learning. arXiv preprint arXiv:2101.01669.

[13] Zhou, J., Cui, G., Hu, S., et al. (2019) Graph neural networks: A review of methods and applications. AI Open, 1:57-81.

[14] Jin, D., Liu, Z., Li, et al. (2019) Graph convolutional networks meet markov random fields: Semi-supervised community detection in attribute networks. In: Proceedings of AAAI, pp.152–159.

[15] Sun, H., He, F., Huang, J., et al. (2020) Network embedding for community detection in attributed networks. ACM Trans. Knowl. Discov., 14:1–25.

[16] Peel, L., Larremore, D. B., Clauset, A. (2017) The ground truth about metadata and community detection in networks. Science advances, 3(5): e1602548.

[17] Blondel, V.D., Guillaume, J.L., Lambiotte, R., et al. (2008) Fast unfolding of communities in large networks. J. Stat. Mech., 10: P10008.

[18] Saxena, A., Iyengar, S. (2020) Centrality measures in complex networks: A survey. arXiv preprint arXiv:2011.07190.

[19] Defferrard, M., Bresson, X., Vandergheynst, P. (2016) Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceedings of NIPS, pp. 3844–3852.

[20] Kipf, T. N., Welling, M. (2016) Semi-supervised classification with graph convolutional networks. In: Proceedings of ICLR, https://arxiv.org/abs/ 1609.02907.

[21] Jin, D., Li, B., Jiao, P., et al. (2019) Community detection via joint graph convolutional network embedding in attribute network. , in: Proceedings of ICANN, pp. 594–606

[22] Zhang, B., Yu, Z., Zhang, W. (2020) Community-centric graph convolutional network for unsupervised community detection. In: Proceedings of IJCAI pp. 551-556.

[23] Wang, W., Liu, X., Jiao, P., et al. (2018) A unified weakly supervised framework for community detection and semantic matching. In: Proceedings of PAKDD, pp. 218–230.

[24] Wang, X., Jin, D., Cao, X., et al. (2016) Semantic community identification in large attribute networks. In: Proceedings of AAAI, pp. 265–271.

[25] He, D., Feng, Z., Jin, D., et al. (2017) Joint identification of network communities and semantics via integrative modeling of network topologies and node contents. In: Proceedings of AAAI, pp. 116–124.

[26] Li, Q., Han, Z., Wu, X. (2018) Deeper insights into graph convolutional networks for semi-supervised learning. In: Proceedings of AAAI, pp. 3538–3545.

[27] Luo, W., Lu, N., Ni, L., et al. (2020) Local community detection by the nearest nodes with greater centrality. Inf. Sci., 517: 377–392.