Higher-Order Multi-Layer Community Detection

Ling Huang, Chang-Dong Wang, Hong-Yang Chao

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

huanglinghl@hotmail.com, changdongwang@hotmail.com, isschhy@mail.sysu.edu.cn

Abstract

In this paper, we define a new problem of multi-layer network community detection, namely higher-order multi-layer community detection. A multi-layer motif (M-Motif) approach is proposed, which discovers communities with good intra-layer higher-order community quality while preserving inter-layer higher-order community consistency. Experimental results have confirmed the superiority of the proposed method.

Introduction

Many efforts have been made in developing effective approaches for community detection in multi-layer network (Mucha et al. 2010; Zhang et al. 2018). However, they are conducted relying on lower-order microscopic proximity structure (e.g. first- and second-order proximities of nodes) that can be captured at the level of individual nodes and edges. Very recently, higher-order structure at the level of small network subgraphs has been introduced for discovering communities that preserve higher-order connectivity patterns. Since higher-order structure lies between lower-order microscopic proximity structure and mesoscopic community structure, it plays the role of building blocks for complex network (Benson, Gleich, and Leskovec 2016) and therefore helps discover communities that best avoid damaging building blocks. In this paper, we define a new problem, namely higher-order multi-layer community detection, and propose a novel approach called multi-layer motif (M-Motif), which for the first time discovers communities in multi-layer network from the perspective of preserving higher-order structure.

Background and Problem Statement

The input is a multi-layer network $\mathcal{G} = \{\mathcal{G}^1, \ldots, \mathcal{G}^v\}$ consisting of v layers with each layer $\mathcal{G}^p = \{\mathcal{V}, \mathcal{E}^p\}, \forall p = 1, 2, \ldots, v$ having the same set of n nodes $\mathcal{V}$ but slightly different linkage structure $\mathcal{E}^p$. The community structure in different layers may differ slightly due to different linkage structures across different layers. The goal is to discover the community structures $\mathcal{C} = \{\mathcal{C}^1, \ldots, \mathcal{C}^v\}$ with $\mathcal{C}^p = \{\mathcal{C}_1^p, \ldots, \mathcal{C}_c^p\}$ for the p-th layer, where $\mathcal{C}_j^p, \forall j = 1, \ldots, c$ is the set of nodes that are assigned to the j-th community

in the p-th layer and c is the number of communities. The community structure $\mathcal{C}^p$ should well preserve the topological structure in the p-th layer while the degree of consistency between community structures $\mathcal{C}^p$ and $\mathcal{C}^q$ should reflect the consistency level between the two layers.

One representative higher-order structure is motif, which is defined as follows (Benson, Gleich, and Leskovec 2016).

Definition 1 (Motif) *Motif is a dense subgraph occurring in complex networks at numbers that are significantly larger than those in randomized networks. The motif is denoted as $\mathbb{M} = \{\mathcal{V}_\mathbb{M}, \mathcal{E}_\mathbb{M}\}$ where $\mathcal{V}_\mathbb{M}$ and $\mathcal{E}_\mathbb{M}$ denote the node set consisting of m nodes and edge set consisting of e edges in the motif $\mathbb{M}$ respectively, with e being between $m - 1$ (a line motif) and $\frac{m(m-1)}{2}$ (a clique motif).*

Despite the bursty studies of motifs in community detection in single-layer network, the higher-order structure of multi-layer community detection remains largely unknown (Huang, Wang, and Chao 2018). This paper aims to study the following new problem.

Definition 2 *Higher-order multi-layer community detection aims to discover a set of v community structures, each for one individual layer, satisfying the following two requirements.*

- ***Requirement I** [Intra-layer Higher-order Community Quality]: For each layer p, the number of motif instances within each community $\mathcal{C}_l^p$ should be as large as possible compared with the expected number of motifs within community $\mathcal{C}_l^p$.*

- ***Requirement II** [Inter-layer Higher-order Community Consistency]: For each node i, if the motif structures containing node i are consistent across two views p and q, the community label should also be consistent.*

Requirement I ensures that the motif structure can be best preserved in the discovered communities. On the other hand, Requirement II ensures the consistency degree of community structures matches the consistency degree of higher-order structures across layers.

The Proposed Multi-layer Motif Approach

For each layer, the statistically significant motifs will be identified first by means of the Abundance measure as described in (Milo et al. 2004), which are denoted as the motif
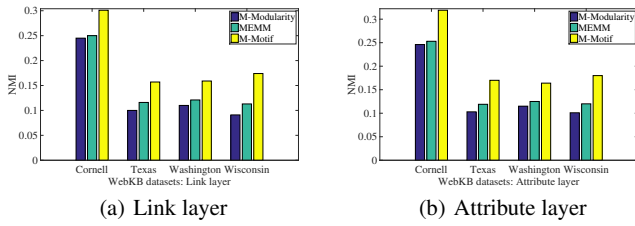
(a) Link layer   (b) Attribute layer

Figure 1: Comparison results on the four datasets.

set $\{\mathbb{M}\}$. Each motif $\mathbb{M}$ is taken as the building blocks of the multi-layer network and will be leveraged as higher-order structure for community detection.

Definition 3 *The motif-based inter-layer consistency of node i across layers p and q is defined as*

$$C_i^{p,q} = JaccardSim\big(\mathcal{I}_\mathbb{M}^p(i), \mathcal{I}_\mathbb{M}^q(i)\big) = \frac{|\mathcal{I}_\mathbb{M}^p(i) \cap \mathcal{I}_\mathbb{M}^q(i)|}{|\mathcal{I}_\mathbb{M}^p(i) \cup \mathcal{I}_\mathbb{M}^q(i)|}$$

where $\mathcal{I}_\mathbb{M}^p(i)$ denotes the instance set of motif $\mathbb{M}$ in layer p containing node i. To avoid trivial calculation, $C_i^{p,p} = 0, \forall i$.

For each layer $\mathcal{G}^p$, the adjacency matrix representing the edge structure $\mathcal{E}^p$ is denoted as $A^p \in \mathbb{R}^{n \times n}$. Let $k_i^p = \sum_{j=1}^n A_{i,j}^p$ denote the degree of node i in layer p and $k^p = \frac{1}{2}\sum_{i=1}^n k_i^p$ denote the total degree of all nodes in layer p. By further denoting the community label of node i in layer p as g_i^p, the objective function of the proposed model is

$$Q(\{g_i^p\}) = \frac{1}{2\mu} \sum_{p,q} \sum_{i_1,\dots,i_m}$$
$$\left[\Big(\prod_{(a,b)\in\mathcal{E}_\mathbb{M}} A_{i_a,i_b}^p - \gamma^p \prod_{(a,b)\in\mathcal{E}_\mathbb{M}} \frac{k_{i_a}^p k_{i_b}^q}{2k^p} \Big) \delta(p,q) \chi(i_1 < \cdots < i_m) \right.$$
$$\left. + C_{i_1}^{p,q} \delta(i_1,\dots,i_m) \right] \delta(g_{i_1}^p,\dots,g_{i_m}^p, g_{i_1}^q,\dots,g_{i_m}^q) \tag{1}$$

subject to: each layer is partitioned into c communities.

In the above objective function, $\gamma^p \in (0,1]$ is the resolution parameter for the p-th layer; $(a,b) \in \mathcal{E}_\mathbb{M}$ denotes the edge (a,b) of the motif $\mathbb{M}$ with a,b being the nodes of the motif; $\chi(i_1 < \cdots < i_m)$ equals 1 if $i_1 < i_2 < \cdots < i_m$ is true and 0 otherwise, which avoids both the overlapping nodes (invalid) and the redundant calculation of motif instance; and $\delta(*,\dots,*)$ is the Kronecker function which equals 1 when all the variables $*$ are equal to each other and 0 otherwise.

It can be proved that the proposed objective function Eq. (1) satisfies both Requirement I and Requirement II stated in Definition 2. An optimization framework can be developed that provably finds near-optimal communities by the "generalized Louvain" approach (Jeub et al. 2011 2017).

Experiments

For experimental purpose, the four WebKB datasets are used, namely Cornell, Texas, Washington and Wisconsin, which contain 195, 187, 230 and 265 web pages respectively. The web pages are interconnected via the hyperlinks,

based on which an undirected network layer can be constructed for each dataset (link layer). Each web page is described by a 0/1-valued word vector representing the absence/presence of each word from the dictionary of size 1703. For each dataset, if the cosine similarity of the representation vectors of two web pages is no smaller than 0.5, the two web pages are interconnected, based on which an undirected network layer can be constructed (attribute layer). In this way, a 2-layer network is constructed for each dataset.

Two existing multi-layer network community detection methods are used as baseline methods, namely multi-layer modularity (M-Modularity) (Mucha et al. 2010) and multi-layer edge mixture model (MEMM) (Zhang et al. 2018). For M-Motif, the 3-node clique motif is used. For all the methods, the parameters are set as default values or tuned as suggested by the authors. Since all the three methods generate the community structure for each layer, Figure 1 plots the results in terms of normalized mutual information (NMI) obtained by the three methods in each of the two layers. From the figure, we can see that M-Motif significantly outperforms the two baselines in generating larger values of NMI in both layers on the four testing datasets.

Conclusions

In this paper, we have presented a novel multi-layer motif approach for solving the newly defined higher-order multi-layer community detection problem. Experimental results have confirmed the effectiveness of the proposed method.

Acknowledgements

References

Benson, A. R.; Gleich, D. F.; and Leskovec, J. 2016. Higher-order organization of complex networks. *Science* 353(6295):163–166.

Huang, L.; Wang, C.-D.; and Chao, H.-Y. 2018. A harmonic motif modularity approach for multi-layer network community detection. In *ICDM*.

Jeub, L. G. S.; Bazzi, M.; Jutla, I. S.; and Mucha, P. J. (2011-2017). A generalized Louvain method for community detection implemented in MATLAB. http://netwiki.amath.unc.edu/GenLouvain.

Milo, R.; Itzkovitz, S.; Kashtan, N.; Levitt, R.; Shen-Orr, S.; Ayzenshtat, I.; Sheffer, M.; and Alon, U. 2004. Superfamilies of evolved and designed networks. *Science* 303(5663):1538–1542.

Mucha, P. J.; Richardson, T.; Macon, K.; Porter, M. A.; and Onnela, J. 2010. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science* 328(5980):876–878.

Zhang, H.; Wang, C.-D.; Lai, J.-H.; and Philip, S. Y. 2018. Community detection using multilayer edge mixture model. *Knowledge and Information Systems* 1–23.