Information Sciences 424 (2018) 55–68

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Adaptive modularity maximization via edge weighting scheme

Xiaoyan Lu[a], Konstantin Kuzmin[a], Mingming Chen[b], Boleslaw K. Szymanski[a,*]

[a] Department of Computer Science, Rensselaer Polytechnic Institute, USA
[b] Google Inc., United States

ARTICLE INFO

Article history:
Received 9 January 2017
Revised 29 July 2017
Accepted 27 September 2017
Available online 29 September 2017

Keywords:
Community detection
Scalability
Modularity maximization
Regularization

ABSTRACT

Modularity maximization is one of the state-of-the-art methods for community detection that has gained popularity in the last decade. Yet it suffers from the resolution limit problem by preferring under certain conditions large communities over small ones. To solve this problem, we propose to expand the meaning of the edges that are currently used to indicate propensity of nodes for sharing the same community. In our approach this is the role of edges with positive weights while edges with negative weights indicate aversion for putting their end-nodes into one community. We also present a novel regression model which assigns weights to the edges of a graph according to their local topological features to enhance the accuracy of modularity maximization algorithms. We construct artificial graphs based on the parameters sampled from a given unweighted network and train the regression model on ground truth communities of these artificial graphs in a supervised fashion. The extraction of local topological edge features can be done in linear time, making this process efficient. Experimental results on real and synthetic networks show that the state-of-the-art community detection algorithms improve their performance significantly by finding communities in the weighted graphs produced by our model.

1. Introduction

Community structures are observed across a wide variety of networks, including World Wide Web, Internet, collaboration, transportation, social and biochemical networks. Many important tasks, such as data extraction, link prediction, network evolution analysis, and graph mining are based on the community structures discovered in these networks.

Modularity maximization is one of the state-of-the-art methods for community detection that has gained popularity in the last decade. It aims at discovering the partition of the network which maximizes modularity [1], a widely used community quality measure proposed by Newman et al. Modularity measures the difference between the observed fraction of edges within a community and the fraction of edges expected in a random graph with the same number of nodes and the same degree sequence. Thus, high positive modularity indicates the quality of a community structure in the network. Although modularity maximization has been widely used in many applications, in certain cases it tends to merge small communities into large ones, giving rise to the so-called *resolution limit problem* [2]. In the literature, initially, it was assumed that community structure with maximum modularity is the best. Discovery of the resolution limit problem demonstrated that this is not the case. Another assumption is that the number of communities in the given graph is unknown.

* Corresponding author.
E-mail address: szymab@rpi.edu (B.K. Szymanski).

https://doi.org/10.1016/j.ins.2017.09.063
0020-0255/

In this paper, we propose to expand the meaning of the edges that are currently used to indicate propensity of nodes for sharing the same community. In our approach this is the role for edges with positive weights while edges with negative weights indicate aversion for putting end-nodes into one community. We also propose a novel feature-based edge weighting scheme that learns how the local topological features indicate whether a given edge is intra- or inter-community using small artificial graphs similar to a network in question. Further, we demonstrate that our proposed regression model assigns weights to edges in such a way that the state-of-the-art community detection algorithms achieve higher accuracy on the produced weighted graphs than they do on the original unweighted ones. Recent work [3] shows that edge weighting scheme is capable of decreasing the upper bound on the size of communities detectable by modularity maximization. A similar approach has been adapted in [4] where edges are weighted according to their centrality. In contrast to [3,4] where the edge weighting schemes are specified by experts, we develop a feature-based regression model and use labeled ground truth communities in artificial networks as training data to infer the suitable weights for edges of the input graph. These artificial networks are constructed to have degree distribution and clustering coefficient similar to the original unweighted networks. Considering the comprehensive definition of local community structures across different network instances, the regression model trained by ground truth community[1] in the artificial networks is therefore able to assign such weights to the edges that community detection is enhanced. Furthermore, the local topological features of edges can be extracted efficiently; so our model converts a graph into a weighted one in a time proportional to the number of edges in a network.

The experimental results on real and synthetic networks show that modularity maximization algorithms achieve higher accuracy on weighted graphs than on the original unweighted ones. For example, the optimal modularity obtained by the Fast Greedy algorithm [5] increases by at least 15% on an LFR benchmark [6]. We also show that our approach solves the resolution limit problem on the American college football network [7]. In addition, the state-of-the-art community detection algorithms, including the label propagation algorithm of Raghavan et al. [8], Newman's leading eigenvector method [9], algorithms based on random walks [10] and the multilevel algorithm of Blondel et al. [11], also improve their performance on the weighted graph produced by our approach, which validates the point that weighting graphs properly guides the algorithm to the desirable community detection results.

This paper is organized as follows. Section 2 introduces the related work on modularity maximization and edge weighting schemes. Section 3 discusses the effectiveness of the edge weighting scheme. The regression model is presented in Section 4, followed by the description of the key speedup improvements of the training algorithm. In Section 5, we describe the experimental results on real and synthetic networks. We close our work with conclusions presented in Section 6.

2. Related work

2.1. Modularity maximization

The goal of the modularity maximization is to discover community structure in a network by maximizing the modularity, defined as

$$Q(G,C)=\sum_{c_i\in C}\left[\frac{|E_{c_i}^{in}|}{|E|}-\left(\frac{d_{c_i}}{2|E|}\right)^2\right] \tag{1}$$

where $G=(V,E)$ is an unweighted, undirected graph with the node set V and the edge set E; $C=\{c_i\}$ is a partition of G into communities, c_i is the set of nodes in the ith community, d_{c_i} is the sum of degrees of nodes in c_i, $E_{c_i}^{in}$ denotes the set of edges residing within community c_i.

The modularity can be naturally extended to the networks with weighted edges by replacing the count of edges with the sum of their weights. Hence, the weighted modularity is defined as

$$Q^w(G^w,C)=\sum_{c_i\in C}\left[\frac{W_{c_i}^{in}}{W}-\left(\frac{W_{c_i}}{2W}\right)^2\right] \tag{2}$$

where W is the sum of weights of edges in the entire graph, $W_{c_i}^{in}$ is the sum of weights of edges within community c_i, and the weight of a community is defined as $W_{c_i}=2W_{c_i}^{in}+W_{c_i}^{out}$ where $W_{c_i}^{out}$ is the sum of weights of edges with exactly one endpoint inside c_i. The original definition of modularity is a special case of the weighted version when the weight of every edge is 1.

Many algorithms including [5,11–15] were proposed to discover communities in a network by maximizing the modularity. One interesting finding is that Newman's modularity measure is related to the broader family of spectral clustering methods [15]. There are two categories of spectral algorithms for maximizing modularity: one is based on the modularity matrix [1,9,16], the other is based on the Laplacian matrix of a network [15,17]. The first greedy algorithm, Fast Greedy [5], iteratively merges communities in the network to maximize the modularity. Initially, every node is a single community. In every step, two communities joining of which results in the largest modularity among all partitions created by temporary

[1] If ground truth communities are not available then thanks to the small size of the artificial graph, we use communities detected algorithmically as ground truth.