ℓ^1-Graph Based Community Detection in Online Social Networks

Liang Huang[1], Ruixuan Li[1,⋆], Yuhua Li[1],
Xiwu Gu[1], Kunmei Wen[1], and Zhiyong Xu[2]

[1] School of Computer Science and Technology, Huazhong University
of Science and Technology, Wuhan 430074, China
[2] Department of Math and Computer Science, Suffolk University, Boston, USA
ellick@smail.hust.edu.cn, {rxli,idcliyuhua,guxiwu,kmwen}@hust.edu.cn,
zxu@mcs.suffolk.edu

Abstract. Detecting community structures in online social network is a challenging job for traditional algorithms, such as spectral clustering algorithms, due to the unprecedented large scale of the network. In this paper, we present an efficient algorithm for community detection in online social network, which chooses relatively small sample matrix to alleviate the computational cost. We use ℓ^1-graph to construct the similarity graph and integrate the graph laplacian with random walk in directed social network. The experimental results show the effectiveness of the proposed method.

Keywords: ℓ^1-graph, Spectral clustering algorithm, Graph Laplacian, Laplacian regularizer.

1 Introduction

Online social networks have attracted more and more attention in recent years. A typical social network consists of a set of nodes (represents the individual in the network) and a set of relational tie y_{ij} (measured on each ordered pair of nodes $i, j = 1, ..., n$). In these social networks, a common feature is the community structure. Detecting community structures in social networks is an issue of considerable practical interest that has received a great deal of attention.

Many algorithms for identifying the communities' structure have been proposed in the past few years. Newman [1] proposed an iterative, divisive method based on the progressive removal of links with the largest betweenness. Santo Fortunato [2] developed an algorithm of hierarchical clustering that consists of finding and removing the edge iteratively with the highest information centrality. In addition, Newman and Girvan [3] proposed a quantitative method called modularity to identify the network communities. It seems to be an effective method to detect communities in networks. However, Fortunato and Barthélemy [4] recently pointed out the serious resolution limits of this method and claimed that the size of a detected module depends on the whole network. To solve this

⋆ Corresponding author.

Q.Z. Sheng et al. (Eds.): APWeb 2012, LNCS 7235, pp. 644–651, 2012.

problem, Li [5] developed another quantitative methods called the modularity density. An alternative way to tackle the problem is by spectral clustering [6]. But the algorithm is only effective to handle small networks, when the network's scale gets bigger, it is incompetently for the community detection job.

Compared to the traditional social networks, online social networks have some new features. Firstly, the scale of online social networks are becoming more and more large due to the scale of the internet. The social networks adopted in early community detection algorithms, e.g. the famous benchmark social network Zachary karate club (consists of 34 people in the club), are small social networks. Yet the online social networks such as the LiveJournal consist of tens of millions members and uncountable relationships. Secondly, the relationships between people are becoming more complicated. The instantaneous and random interaction between millions of people have created the uncountable relationship in online social networks. Among these social networks, most are directed one that must be took into consideration.

We integrate the ℓ^1-graph [7] and Laplacian regularizer [8] in our work to handle the situation aforementioned. The work of Chung [9] is also included. This paper introduces these technologies for several good reasons. First, the ℓ^1-graph utilize the overall contextual information instead of only pairwise Euclidean distance as conventionally. The neighboring samples of a datum and the corresponding ingoing edge weights are both considered in the construction of the similarity graph. Second, the laplacian regularizer is suitable for the community detection job in online social network. The laplacian regularizer chooses relatively small sample sets of the given social network. This approach can reduce the computational cost and make it suitable for the online social network. At last, the traditional graph laplacian presume the social network is an undirected and weighted one, makes it unsuitable for the directed social network. As we known, our method is the first work that combine the ℓ^1-graph in the laplacian regularizer when applying in the online social networks. Our studies give a new insight to deal with these tasks and achieves better results compared to the previous outstanding clustering algorithm.

This paper is organized as follows. We introduce our methodology in section 2. Including the details of ℓ^1-graph and the graph laplacian, etc. Subsection 2.3 is the detail of the regularized spectral clustering algorithm. We test the algorithm in the experiments, the results are compared with three algorithm in this section. Finally, the conclusion is given in Section 4.

2 Our Methodology

2.1 Construct Similarity Graph with ℓ^1-graph

There are several popular ways for graph construction like the KNN and the ε-ball, etc. However, these methods only consider the pairwise Euclidean distance without integrating the overall contextual information of the social networks. The ℓ^1-graph uses this information to construct robust and datum-adaptive similarity graph and achieves good results in our experiments.

The construction process of ℓ^1-graph is formally stated as follows.

1) **Inputs:** The sample data set denoted as the matrix X=$[x_1, x_1, ..., x_N]$, where $x_i \in \mathbb{R}^m$.

2) **Robust sparse representation:** The sparse coding can be computed by solving the ℓ^1-Norm optimization problem in (1),

$$\arg\min_{\alpha^i} \|\alpha^i\|, \qquad s.t. \quad x_i = B^i\alpha^i. \tag{1}$$

where matrix B^i=$[x_1, x_1, ..., x_N, I] \in \mathbb{R}^{m+N-1}$.

3) **Graph weight setting:** We can compute the graph weight matrix W for i-th training sample x_i in (2).

$$w_{ij} = \begin{cases} \alpha^i_j, & i > j, \\ 0, & i = j, \\ \alpha^i_{j-1}, & i < j. \end{cases} \tag{2}$$

2.2 Introduction of Undirected and Directed Graph Laplacian

Given the social network $G = (V,E)$, the similarity matrix of the social graph can be denoted with $W = w_{ij}(i, j = 1, ..., n)$, w_{ij} is the similarity between the nodes. Provided G is undirected that $w_{ij} = w_{ji}$, the degree d_i of $v_i \in V$ can be defined as $d_i = \sum_{j=1}^{n} w_{ij}$.

The simplest form of undirected graph laplacian matrix can be defined as

$$L = D - W.$$

D can be defined as the diagonal matrix with the degrees $d_1, ..., d_n$ on the diagonal. (3) is the unnormalized laplacian matrix.

There is another form of graph laplacian matrix called normalized graph laplacians which can be represented as follows

$$L_{sym} := D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2},$$
$$L_{rw} := D^{-1}L = I - D^{-1}W.$$

A random walk on the given directed graph for graph laplacian is defined in [9]. Given the weighted directed graph G (presume the directed graph G is strongly connected and aperiodic), the out-degree of v_i with d_i^+ can be denoted as follow:

$$d_i^+ = \sum_{j:(v_i, v_j \in E)} w(i, j).$$

The random walk over G with transition probability matrix P can be defined as follows:

$$P_{ij} = \begin{cases} \frac{w(i,j)}{d_i^+}, & if(v_i, v_j) \in E, \\ 0, & otherwise. \end{cases}$$

The above random walk has a unique stationary distribution $\mathit{\Pi} = (\pi_1, ..., \pi_1)^T$ with $\pi_i > 0$ for all i. The Laplacian of G is then defined as

$$\tilde{\mathbf{L}} = \mathbf{I}_n - \frac{\prod^{1/2} \mathbf{P} \prod^{-1/2} + \prod^{1/2} \mathbf{P}^T \prod^{-1/2}}{2}.$$

Where $\prod$ is the diagonal matrix $\prod = diag(\pi_i)$. The Laplacian matrix $\tilde{\mathbf{L}}$ constructed as above can be used in exactly the same way as in the undirected case.

2.3 Regularized Spectral Clustering Algorithm

Several graph regularizers have been proposed in recent years [10,11]. These regularizers usually take the form

$$\mathcal{S}_G(f) = \mathbf{f}^T \mathbf{S} \mathbf{f}.$$

for an appropriate symmetric. The matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ derived from social matrix G. Generally, online social networks may scale up to tens of thousands of nodes and millions of edges. This makes it uncompetitive for standard spectral clustering algorithm due to the frequently matrix operations. The regularized spectral clustering algorithm uses the laplacian regularizer to overcome this shortcoming. By choosing sample nodes in social graph G, the introduced algorithm can alleviate the computational cost significantly through reducing the matrix operation down to acceptable scale.

Instead of using vectors as in spectral clustering algorithm [6], the regularized spectral clustering algorithm constructs a target function for clustering job. This target function then partitions the whole social network naturally. Given have an online social network $G = (V, E), V = v_1, v_2, ..., v_n$, the sample set $G' = (V', E') \subseteq G, V' = v'_1, v'_2, ..., v'_k, k \leq n$ will be chosen to construct the target function, and the function will be used to cluster the whole social network. Let $\mathbf{W}$ be the ℓ^1-weight of G', and L_n be the laplacian matrix of G'. The coefficient α can be calculated by solving the following nonlinear constraint problem (3), see in [12] for details.

$$\begin{aligned} \alpha = \arg \min_{\alpha \in \mathscr{R}^n} & \frac{1}{n^2} \alpha^T \mathbf{W} L_n \mathbf{W} \alpha + \gamma \alpha^T \mathbf{W} \alpha, \\ s.t. \quad & \frac{1}{n^2} \alpha^T \mathbf{W} D \mathbf{W} \alpha = 1, \\ & \alpha^T \mathbf{W} D \mathbf{1} = 0. \end{aligned} \tag{3}$$

$\alpha^{\mathbf{x}} = (\alpha^{\mathbf{x}}_1, ..., \alpha^{\mathbf{x}}_n) \in \mathscr{R}^n$ is the coefficient of the target function (4). $\mathbf{W}$ is the ℓ^1-graph weight and $\mathbf{1}$ is the vector of all ones. L_n,D is the corresponding laplacian and degree matrix, the details has been introduced in subsection 2.2. One can calculate the target function $\mathbf{f}$ by applying the coefficient in (4).

$$f(x) = \sum_{i=1}^{n} \alpha_i \mathbf{W}(x_i, x). \tag{4}$$

Algorithm 1. The methodology of our paper

Require:
The social matrix of social network graph $G = (V, E)$;

Ensure:
1: Random choose the appropriate sample set of the online social network;
2: Use the ℓ^1-graph to construct the similarity graph of the sample set;
3: Compute the graph laplacian matrix, if the social network is a directed one, use the method described in [9] to compute the directed graph laplacian matrix;
4: Compute the generalized eigenvectors and eigenvalues of the graph laplacian matrix;
5: Use the vectors of the first k eigenvalues to compute the corresponding coefficient α in (9);
6: Compute the target function values of the whole social network data correspond to the respective coefficient α;
7: Use the k vectors of the target function values to cluster the social network;
8: **return** The clusters of the given social network.

Let P be the projection onto the subspace of $\mathbb{R}^u$ orthogonal to $\mathbf{W1}$, one obtains the solution for the constrained problem in (3), which is given by the following generalized eigenvalue problem [8].

$$P(\gamma \mathbf{W} + \mathbf{W}L\mathbf{W})PV = \lambda P\mathbf{W}^2 PV. \tag{5}$$

The final solution is given by $\alpha = PV$, where v is the eigenvector corresponding to the eigenvalues. Similar to the standard spectral clustering algorithm, we choose the eigenvectors of the smallest k eigenvalues to get the k cluster results.

3 Experiments

We test our method in three different social networks, the Arxiv HEP-PH collaboration network [13] and two benchmark social networks [14,15]. The concept of modularity function [1] is employed to quantify the clustering results.

3.1 Two Benchmark Social Networks

In this subsection, four algorithms are tested for the comparison including Kmeans, standard spectral clustering algorithm , regular spectral clustering algorithm with Gaussian similarity graph ($exp(-\|x_i - x_j\|^2/(2\sigma^2))$, $\|x_i - x_j\|$ is the Euclidean distance which be defined in this paper as the shortest distance between node i and j), and our method. These algorithms are applied in two benchmark social networks, the Zachary karate club network [14] with 34 nodes and Dolphin network [15] with 62 nodes. We sample the full networks for regular spectral clustering algorithm in the experiments for comparison.

Figure 1 shows the modularity scores of the four algorithms in Zachary and Dolphin network. Our method outperform the other algorithms in the Zachary

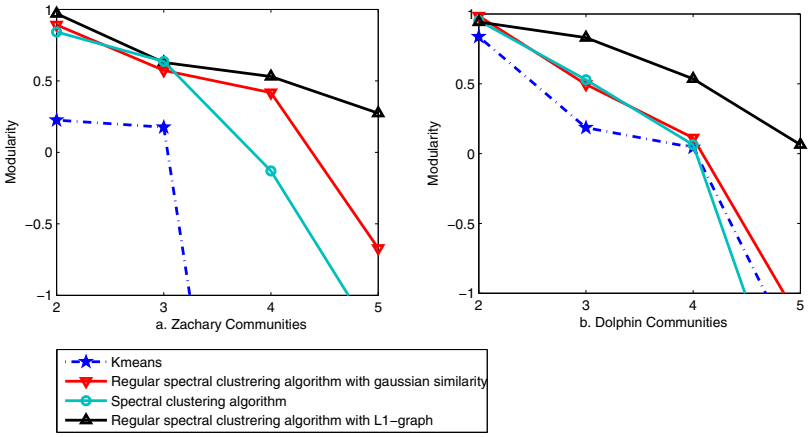

Fig. 1. The modularity of four algorithms applied in Zachary and Dolphin network

network for each community numbers. The standard clustering algorithm achieves the highest score in the dolphin network for but drop fast when the community numbers increase. Our method still win the highest scores for nearly every community numbers in Dolphin network, and gives the most stable performance in Zachary and Dolphin networks both. Kmeans gives the poorest performance in two networks compared to the other algorithms.

3.2 Arxiv HEP-PH Collaboration Network

Arxiv HEP-PH (High Energy Physics - Phenomenology) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to High Energy Physics - Phenomenology category. We choose the data covers papers in the period from January 1993 to April 2003 (124 months). A collaboration network with 12,008 nodes and 237,010 edges is constructed in this way.

We choose training sets with size $m \in \{800, 2000\}$ in the experiments, the standard clustering algorithm is employed for the comparison. The most time-consuming part of our method is the computation of the ℓ^1-graph and the shortest distance. In the experiments, it takes about 9 minutes for 800 sample nodes when applying our method from step 2 to 7 in algorithm 1, and 17 minutes for 2000 sample nodes. The standard clustering algorithm takes exceed half an hour for the similar steps. 46 communities are detected in wiki-vote social network corresponding to 500 sample nodes, and 91 communities for 2000 sample nodes. Meanwhile, the standard clustering algorithm detect 77 communities which need to calculate the whole network. We can see from Figure 2, our method can achieves better performance with small sample network compared to the spectral clustering algorithm. When the size of the social network scale up, more time will be spent in standard spectral clustering algorithm with poor performance.

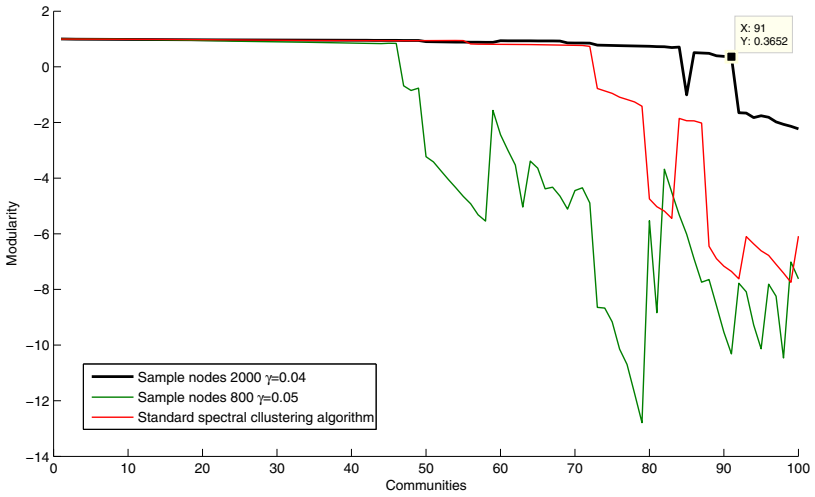

Fig. 2. The three modularity curves of our method (800, 2000 samples) and standard spectral clustering algorithm applied in Arxiv HEP-PH collaboration network

4 Conclusion

The issues of detecting community structures in large online social networks are increasingly common. Our work builds on recent work on ℓ^1-graph and regularization which aimed at detecting community structure in complex networks in an efficient way. The experimental results indicate that our method can achieve better performance with relatively small computational cost.

The method will not only allow for the extension of community structure analysis to some of the very large networks, but will also provide a useful tool for visualizing and understanding the structure of these networks. We hope that this approach will be employed successfully in the search and study of communities in social networks, and will help to uncover new interesting properties in this area.

However, the social networks adopted in our paper are smaller one that consist of tens of thousands nodes. The bottleneck of our method is the computation capability of the computer. We plan to integrate our method into the distributed computing environment such as Map-reduce, etc. We believe that better performance can be achieved in the extreme large online social network like the LiveJounal and delicious social network in this way.

Acknowledgment. This work is supported by National Natural Science Foundation of China under grants 61173170 and 60873225, and Innovation Fund of Huazhong University of Science and Technology under grants 2011TS135 and 2010MS068.

References

1. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E. 69(2), 26113 (2004)
2. Fortunato, S., Latora, V., Marchiori, M.: Method to find community structures based on information centrality. Phys. Rev. E. 70(5), 56104 (2004)
3. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E. 69(2), 66133 (2004)
4. Barthélemy, M., Fortunato, S.: Resolution limit in community detection. Proc. Natl. Acad. Sci. 104, 36–41 (2007)
5. Li, Z.P., Chen, L.N., Zhang, S.H., Zhang, X.S., Wang, R.S.: Quantitative function for community detection. Phys. Rev. E. 77, 36109 (2008)
6. Luxburg, U.V.: A Tutorial on Spectral Clustering. Tech. Rep. 149, Max Planck Institute for Biological Cybernetics. (2006)
7. Cheng, B., Huang, T.S., Yang, J.C., Yan, S.C.: Learning With ℓ^1-graph for Image Analysis. IEEE Trans. on Image Processing 19, 858–866 (2010)
8. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. J. Machine Learning Research 7, 2399–2434 (2006)
9. Chung, F.R.K.: Laplacians and the Cheeger inequality for directed graphs. Annals of Combinatorics 9, 1–19 (2005)
10. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15(6), 1373–1396 (2002)
11. Smola, A.J., Kondor, R.: Kernels and regularization on grphs. In: Proc. of the 16th Annual Conference on Learning Theory (2003)
12. Cao, Y., Chen, D.R.: Consistency of regularized spectral clustering. Applied and Computational Harmonic Analysis 30, 319–336 (2011)
13. Stanford University Stanford Network Analysis Project, `http://snap.stanford.edu/ncp/`
14. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. 33, 452–473 (1977)
15. Boisseau, O.J., Dawson, S.M., Haase, P., Lusseau, D., Schneider, K., Slooten, E.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behav. Ecol. Sociobiol. 54, 396–405 (2003)