Winter 12-6-2015

Neighborhood Overlapped Propagation Algorithm For Community Detection Based On Label Time-Sequence

Yu-ling Hong

Qishan Zhang

NEIGHBORHOOD OVERLAPPED PROPAGATION ALGORITHM FOR COMMUNITY DETECTION BASED ON LABEL TIME-SEQUENCE

Hong Yu-ling, College of Economics and Management, China, ailife8@qq.com
Zhang Qi-shan, College of Economics and Management, China, zhang_qs@foxmail.com

ABSTRACT

The community detection algorithms based on label propagation (LPA) receive broad attention for the advantages of near-linear complexity and no prerequisite for any object function or cluster number. However, the propagation of labels contains uncertainty and randomness, which affects the accuracy and stability of the LPA algorithm. In this study, we propose an efficient detection method based on COPRA with Time-sequence (COPRA_TS). Firstly, the labels are sorted according to a new label importance measure. Then, the label of each vertex is updated according to time-sequence topology measure. The experiments on both the artificial datasets and the real-world datasets demonstrate that the quality of communities discovered by COPRA_TS algorithm is improved with a better stability. At last some future research topics are given.

Key words: Community Detection; Label Propagation; Neighborhood Topology; Label Time-sequence

INTRODUCTION OF MING THE SOCIAL NETWORKS

In a social system, individuals tend to group with others who are like-minded or with whom they interact more regularly and intensely than others. Examples include the Internet, the world-wide-web, social and biological systems of various kinds, and many others [2][21][27]. This process leads to the formation of communities. Community discovery is a classical problem in social network analysis, where the goal is to discover related groups of members such that intra-community associations are denser than the associations between communities. Furthermore, actors with interests and purposes in different fields result in overlapped communities. For instance, overlapping features can be observed in scientific collaboration networks in which scientists participate in multiple disciplines [23]. This, in fact, is quite evident today. Community detection has diverse applications including the prediction of forthcoming events, activities or developments, business intelligence, campaign management, infrastructure management, churn prediction, etc.

Generally, a community is a sub-graph of a collection of members in a social network. Many complex systems in nature and society can be described in terms of networks or graphs. Complex networks are usually characterized by several distinctive properties: power law degree distribution, short path length, clustering and community structure. The problem becomes important because complex system's dynamics is actually determined by the interaction of many components and the topological properties of the network will affect the dynamics in a very fundamental way. A vast number of overlapping community detection methods have been developed, especially in the last few years. These include modularity based methods [7][15][22], spectral based methods [9][13][14][20] and matrix factorization based methods [10][24][28]. Matrix factorization methods such as Non-Negative Matrix Factorization (NMF) [16], can be used to classify nodes into corresponding communities. For example, Wang et al. [28] propose various NMF frameworks that can be used in overlapping community detection. Also, Zarei et al. [32], proposed a NMF-based method to detect overlapping communities using Laplacian matrix of a given network. NMF can also be used to detect communities on large networks [30]. However, the paramount drawback of such methods is, the number of communities must be known in advance, which is often not feasible.

To overcome the above mentioned challenge, several NMF-based method like, Bayesian NMF [24], Bounded Non-Negative Matrix Tri-Factorization[33] and Binary matrix factorization [19][34] have been proposed. Nodes in Bayesian NMF are classified into corresponding communities using Bayesian NMF and the number of communities present in the network is defined as the inner rank of network relation graph. Bounded Non-Negative Matrix Tri-factorization [33], uses the stated method to detect overlapped communities. Binary matrix factorization, such Symmetric Binary Matrix Factorization (SBMF) (19,20) uses optimized NMF methods on binary matrices to detect communities in the network. For instance, Zhang et al. [34] proposed an overlapping community detection method using SBMF. In SBMF [34], partition density [1] is used to compute the number of communities present in the network. Although these methods can be extended in link communities [5][12], they are still characterized by limited resolution and high computation complexity.

One of the fastest algorithms proposed to date is the label propagation algorithm (LPA) of Raghavan et al[25] well as its near-linear time complexity (for sparse networks), it is very simple and has no parameters. However, like most community detection algorithms, it can detect only disjoint communities. In this paper, we propose an algorithm that generalizes the LPA based on Time-sequence to find overlapping communities. It takes a parameter, *r*, which controls the potential degree of overlap between communities. The LPA is essentially a special case of the proposed algorithm with *r*=1.The section 3 describes the COPRA_TS algorithm.

RELATED WORKS

Basic Concepts

Suppose that $G = (V, E)$ is an undirected network, where $V = \{v_1, v_2 ..., v_n\}$ is a non-empty set of n vertices, E, is the set of edges $e_{ij} \in$ E, such that each edge connects vertices v_i and v_j . The value of $n = |V|$ and $m = |E|$ is the total number of vertices and edges respectively, that are present in a network.

Detecting Communities By Label Propagation

The LPA algorithm can be described very simply. Each vertex is associated with a label, which is an identifier such as an integer.

1. To initialize, every vertex is given a unique label.

2. Then, repeatedly, each vertex x updates its label by replacing it by the label used by the greatest number of neighbors. If more than one label is used by the same maximum number of neighbors, one of them is chosen randomly. After several iterations, the same label tends to become associated with all members of a community.

3. All vertices with the same label are added to one community.

The propagation phase does not always converge to a state in which all vertices have the same label in successive iterations. To ensure that the propagation phase terminates, Raghavan *et al* propose the use of "asynchronous" updating, whereby vertex labels are updated according to the previous label of some neighbors and the updated label of others. Vertices are placed in some random order. *x*'s new label in the *t*th iteration is based on the labels of the neighbors that precede *x* in the *t*th iteration and the labels of its neighbors that follow x in the (*t*-1)th iteration. The algorithm terminates when every vertex has a label that is one of those that are used by a maximum number of neighbors.

The algorithm produces groups that contain all vertices sharing the same label. These groups are not necessarily connected, in the sense that there is a path between every pair of vertices in the group passing only through vertices in the same group. Since communities are generally required to be connected, Raghavan *et al* propose a final phase that splits the groups into one or more connected communities.

The time complexity of the algorithm is almost linear in the network size. Initialization takes time $O(n)$, each iteration takes time $O(m)$, and the time for processing disconnected communities is $O(m+n)$. The number of iterations required is harder to predict, but Raghavan *et al* claim that five iterations is sufficient to classify 95% of vertices correctly.

Leung *etal* [17] have analysed the LPA algorithm in more detail. They compare asynchronous with synchronous updating, whereby the new label of each vertex in the *i*th iteration is always based on the labels of its neighbors in the (*i*-1)th iteration. They found that synchronous updating requires more iterations than asynchronous updating, but is "much more stable". They also propose restraining the propagation of labels to limit the size of communities, and a similar technique to allow detection of hierarchical communities. Both Refs. [10] and [16] hint at the possibility of detecting overlapping communities, but neither extends the algorithm to find them. COPRA [11] modified the classic LPA [17] such that each node can retain multiple labels in order to find overlapped community structure. But it imposes the number of communities a node participates in as a restriction, which is not the case in real network [29]. Furthermore, the method is deterministic i.e., the results are not dependent on the sequence in which the nodes are considered. This is also a problem in [3][6][8][11][15].We do this in the next section.

NEIGHBORHOOD TOPOLOGY METHOD

Problem Definition

For each node $v_i \in V$, $N(v_i)$ is a set of all vertices adjacent to v_i. In other words, $N(v_i)$ is the neighborhood set of vertex v_i. Or, $N(v_i) = \{ v_k | (v_i, v_k) \in E\}$. The value of $\delta(v_i)$ denotes the degree of the vertex v_i. Adjacency matrix, *A*, of a graph, *G*, represents a relation between nodes where, $A_{ij} = 1$, if there is an edge between v_i and v_j and $A_{ij} = 0$ otherwise. And then let $N_i(v_j)$ be the within community neighborhood of node v_j defined for community $S_i \in S(v_j)$ as follows: $N_i(v_j) = \{ v_k | (v_j, v_k) \in E \wedge v_k \in S_i \}$. Furthermore, to measure the importance of its community $S_i,$ neighborhood connectedness is defined by FOCS[4] for a node v_j as the ratio of the size of its within community neighborhood to the size of its (overall) neighborhood. $\xi_j^i = |N_i(v_j)| / |N(v_j)|$ This score emphasizes on the fraction of neighborhood of node v_i that is present within the community S_i.

In the COPRA[11] method, initially a vertex label identifies a single community to which the vertex belongs. And then it extends the label and propagation step to include information about more than one community: each vertex can belong to up to *v* communities, where *v* is the parameter of the algorithm. Alternatively, each vertex *x* is labeled with a set of pairs (*c*, *b*), where *c* is a community identifier and *b* is a belonging coefficient, indicating the strength of *x*'s membership of community *c*, such that all belonging coefficients for *x* sum to 1.

The driving principle for this paper is that communities are initiated by the interest of individuals, and influenced by their neighbors and neighboring communities. Those that find enough common interest may choose to stay and have more connectivity. The communities then expand further as the process is iterated by the newly added ones.

Neighborhood Overlapped Community Detection Algorithm Based on Label Time-sequence

Initial Communities

Initially every node v_i, $\forall$ $i\in\{1,2,\ldots,|V|\}$, that has at least K neighbors, builds a community S_i with its neighbors. The number of communities thus is equal to the number of nodes with degree greater than K. In this way each node becomes a part of the communities initiated by itself and by its neighbors as well, allowing overlap between the communities at the initiation. This approach further helps a node participating in multiple communities to selectively stay in more than one community based on high connectedness scores (or leave the rest), simultaneously.

Label Time-Sequence

By adopting the aforementioned labels of node v_i along each iteration $L_i = \{l_1, l_2, \ldots, l_t\}$, we can comprehensively use the information in the entire network. Moreover, a weight value is assigned to each node as follows. A node will choose to add its label by calculating the longest common subsequence in a social community topology. For example, as shown in the Figure 1, all of nodes around the core one have the longest common subsequence (2, 5). Therefore, all of the members of this community add a new label “7” to their label sequence and reorder them by its time sequence appeared in the algorithm.

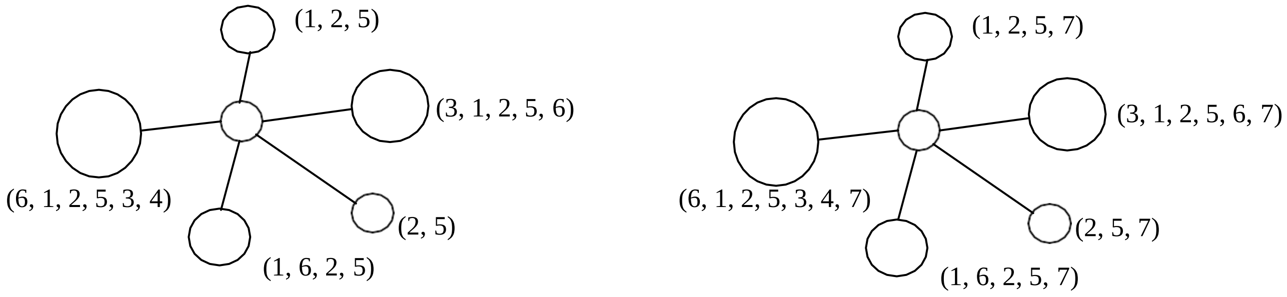

Figure 1. The procedure of label time-sequence

Model

```
Algorithm Neighborhood Overlapped Community Detection Based on Label Time-sequence
Input: $G = (V, E)$: input graph, $k$: maximum common sequence allowed overlap between communities
Output: $S = \{S_i|S_i \subseteq V$ and $S_i$ is a community$\}$
Auxiliary Variables: $n = |V|$, $N(v)$ = neighbors of node $v$, $Added_i$ = Nodes added to community $S_i$ in last round
1. For each vertex x:
      old.$x$ ← {($x$,1)}.
2. For each vertex $x$:
      Propagate($x$,old,new).
3. If id(old) = id(new);
      min ←mc(min,count(new)).
   Else:
      min ←count(new).
4. If min ≠ oldmin:
      Old ← new.
      Oldmin ←min.
      Repeat from step 2.
5. For each vertex $x$:
      Ids ← id(old $x$).
      For each c in ids:
         If, for some m, ($c$,$m$) is in coms, ($c$,$i$) in sub;
            coms ← coms - {($c$,$m$)} ∪ {($c$,$m$)} ∪ {$x$}}.
            LCS(coms,subs).
         Else:
            coms ←coms ∪ {$c$,{$x$}}.
            sub ←sub ∪ {(c,ids)}.
6. For each ($c$,$v$) in sub:
      If $i$ ≠ { }: coms ← coms - {$c$,$m$}.
```

```
7. Split disconnected communities in coms.
```

Figure 2. The COPRA_TS algorithm

Experiments And Comparisons

In this section, we apply the algorithm COPRA_TS to two real-world complex networks, namely, Zachary's karate club dataset [31] and the Dolphin social network [18]. Girvan and Newman [21] proposed the concept of modularity, which is mainly based on the assumption that a community structure is not found in random graphs. However, modularity has some trouble dealing with overlapping community structures, so Chen et al. [26] extended and then redefined it as follows:

$$EQ = \frac{1}{2m}\sum_{i}\sum_{v\in c_i, w\in c_i}\frac{1}{O_v O_w}[A_{vw} - \frac{k_v k_w}{2m}] \qquad (1)$$

In Eq.(1), A represents the adjacency matrix, k_u and k_v are the degrees of nodes u and v respectively, c is the set of all communities and m is the total number of nodes in the networks.

Experiment On Zachary's Karate Club Dataset

Zachary's karate club dataset is the social network of a karate club at an American university that reflects the relationship among its 34 members. The graph has 34 nodes and 78 edges. Each node represents two members of the club that frequently join activities together.

The experiment result on Zachary's karate club dataset is shown in Table 1. We found that the community detection result derived from our proposed algorithm is the same as the COPRAalgorithm. However, the speed of detecting community is much lower than the CPM.

Table 1. Communitydetection result on Zachary's karate club dataset

	CPM	COPRA	COPRA_TS
EQ	0.265	0.459	0.462
Time(s)	0.098	0.168	0.147

Experiment On The Dolphin Social Network

The Dolphin social network refers to the relationship formed by a group of bottlenose dolphins that live in Doubtful Sound Gulf, New Zealand. The dolphin group consists of two families. A total of 62 nodes and 159 edges are present in the network.

The experiment result on the Dolphin social network is shown in Table 2. In table 2, the new algorithm is found to improve the quality of the community. When the network is more complex, the superiority is more obvious.

Table 2. Modular degree of EQ in dataset

datasets	LAP	COPRA_TS	EQ-Increasing/%
Zachary's	0.3653	0.3762	2.2
the Dolphin	0.4770	0.6139	4.5

CONCLUSIONS

We have presented an algorithm, COPRA_TS, to detect overlapping communities in networks by label propagation. It is based on time-sequence of the labels propagated in every iterations condition that permits "synchronous updating". COPRA_TS is guaranteed to terminate, and usually terminate with a good solution especially on giant networks. COPRA_TS inherits some theoretical drawbacks that the original COPRA has. We note that many the recent improvements to the LPA and COPRA may also be applicable to COPAR_TS. And it can compromise the idea of Hausdorff distance and LCS in trajectory classification in the near future.

REFERENCES

[1] Ahn, Y.-Y., Bagrow, J.P., Lehmann, S. (2010) 'Link communities reveal multiscale complexity in networks', *Nature*, Vol. 466, No. 7307, pp. 761–764.

[2] Albert, R. & Barabasi, A.-L. (2002) 'Statistical mechanics of complex networks'. *Rev. Mod. Phys*, Vol. 74, pp. 47-97.

[3] Alvari, H., Hashemi, S. & Hamzeh, A., (2013) 'Discovering overlapping communities in social networks: A novel game-theoreticapproach', *AI Communications*, Vol. 26, No. 2, pp. 161–177.

[4] Bandyopadhyay, S., Chowdhary, G., Sengupta, D. 'FOCS: Fast overlapped community search', *IEEE Transactions on Knowledge and Data Engineering*, p. 1-1.

[5] Banerjee, A., Dhillon, I., Ghosh, J., Sra, S. (2003) 'Generative model-based clustering of directional data', in *Proceedings*

of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'03, ACM, New York, NY, USA, pp. 19–28.

[6] Baumes, J., Goldberg,M. & Magdon-Ismail, M. (2005) 'Efficient identification of overlapping communities', *Intelligence and Security Informatics*. Vol. , Springer, 2005, pp. 27–36.

[7] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008) 'Fast unfolding of communities in large networks', *J. Stat. Mech. Theory Exp'* No.10, 10008.

[8] Chen, W., Liu, Z., Sun, X. & Wang, Y. (2010) 'A game-theoretic framework to identify overlapping communities in social networks', *Data Mining and Knowledge Discovery,* Vol. 21, No. 2, pp. 224–240,.

[9] Donetti, L., Munoz, M.A. (2004) 'Detecting network communities: Anew systematic and efficient algorithm', *J. Stat. Mech. Theory Exp,* 2004, No. 10, 10012.

[10] Gauvin, L., Panisson, A., Cattuto, C. (2014) 'Detecting the community structure and activity patterns of temporal networks: Anon-negative tensor factorization approach', *PLoS One,* Vol.9, No. 1, e86028.

[11] Gregory, S. (2010) 'Finding overlapping communities in networks by label propagation', *New Journal of Physics,* Vol.12, No. 10, 103018,.

[12] He, D., Jin, D., Baquero, C., Liu, D. (2014) 'Link community detection using generative model and nonnegative matrix factorization', *PLoS One,* Vol. 9, No. 1, e86899.

[13] Inoue, K., Li, W., Kurata, H. (2010) 'Diffusion model based spectral clustering for protein–protein interaction networks', *PLoS One,* Vol. 5, No. 9, e12623.

[14] Jiang, J.Q., Dress, A.W.M., Yang, G. (2009) 'A spectral clustering-based framework for detecting community structures in complex networks', *Appl. Math' Lett,* Vol. 22, No.9, pp. 1479–1482.

[15] Lancichinetti, A., Fortunato, S., Kertész, J. (2009) 'Detecting the overlapping and hierarchical community structure in complex networks', *New J. Phys,* Vol.11, No.3, 033015.

[16] Lee, D.D., Seung, H.S. (1999) 'Learning the parts of objects by non-negative matrix factorization', *Nature,* Vol. 401, No.6755, pp. 788–791.

[17] Leung, I. X. Y., Hui, P., Liò, P. & Crowcroft, J. (2009) 'Towards real-time community detection in large networks', *Phys. Rev. E,* Vol. 79, 066107.

[18] Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M. (2003) 'The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations', *Behav. Ecol. Sociobiol,* Vol. 54, No. 4, pp. 396-405.

[19] Ma, X., Gao, L., Yong, X., Fu, L. (2010) 'Semi-supervised clustering algorithm for community structure detection in complex networks', *Physica A,* Vol. 389, No.1, pp. 187–197.

[20] Nadakuditi, R.R., Newman, M.E.J. (2012) 'Graph spectra and the detectability of community structure in networks', *Phys. Rev. Lett,* Vol. 108, 188701.

[21] Newman, M. E. J. (2003) 'The structure and function of complex networks', *SIAM Revies,* Vol. 45, pp. 167-256.

[22] Newman, M.E.J., Girvan, M. (2004) 'Finding and evaluating community structure in networks', *Phys. Rev,* Vol. 69, No. 2, 026113.

[23] Palla, G., Derenyi, I., Farkas, I., Vicsek, T. (2005) 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature,* Vol. 435, No.7043, pp. 814-818.

[24] Psorakis, I., Roberts, S., Ebden, M., Sheldon, B. (2011) 'Overlapping community detection using bayesian non-negative matrix factorization', *Phys. Rev. E,* Vol. 83, No. 6, 066114.

[25] Raghavan, U. N., Albert, R. & Kumara, S. (2007) 'Near linear time algorithm to detect community structures in large-scale networks', *Phys. Rev. E,* Vol. 76, 036106.

[26] Shang, D. M., Lv, Z., Fu, Y. (2010) *Physica A,* Vol. 389 , 4177.

[27] Strogatz, S. H. (2001) 'Exploring complex networks', *Nature,* Vol. 410, pp. 268-276.

[28] Wang, F., Li, T., Wang, X., Zhu, S., Ding, C. (2011) 'Community discovery using nonnegative matrix factorization', *Data Min. Knowl. Discov,* Vol. 22, No. 3, pp. 493–521.

[29] Yang, J. & Leskovec, J. (2012) 'Structure and overlaps of communities in networks', *CORR,* Vol. Abs/1205.6228.

[30] Yang, J., Leskovec, J. (2013) 'Overlapping community detection at scale: A nonnegative matrix factorization approach', in *Proceedings of the Sixth ACMInternational Conference on Web Search and Data Mining*, WSDM'13, ACM, New York, NY, USA, pp. 587–596.

[31] Zachary, W.W. (1977) 'An information flow model for conflict and fission in small groups. J', *Anthropol,* Vol.33, No. 4, pp. 452-473.

[32] Zarei, M., Izadi, D., Samani, K.A. (2009) 'Detecting overlapping community structure of networks based on vertex–vertex correlations', *J. Stat. Mech. Theory Exp,* Vol. 2009, No. 11, 11013.

[33] Zhang, Y., Yeung, D.-Y. (2012) 'Overlapping community detection via bounded nonnegative matrix tri-factorization', in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'12, ACM, New York, NY, USA, pp. 606–614.

[34] Zhang, Z.-Y., Wang, Y., Ahn, Y.-Y., (2013) 'Overlapping community detection in complex networks using symmetric binary matrix factorization', *Phys. Rev. E,* Vol. 87, No. 6, 062803.