Multi-Label Propagation for Overlapping Community Detection Based on Connecting Degree

Xiaolan Wu[1] and Chengzhi Zhang[2]

[1]*wuxiaolananhui@163.com,* [2]*zhangchz@istic.ac.cn*
Dept. of Information Management, Nanjing University of Science and Technology, Nanjing 210094 (China)

Introduction

With the growth of social media, social network analysis draws a great attention and becomes a hot research topic in the field of complex network, web mining, information retrieval, etc. An important aspect of social networks analysis is community structure (Newman, 2003).

In general, community detection methods are classified into two categories: overlapping methods (and non-overlapping methods (Hofman & Wiggins, 2008)). The former allows communities overlap, while the latter assumes that a network only contains disjoint communities. In this paper, we focus on the overlapping community detection. To find overlapping community, researchers use a wide variety of techniques, such as Clique Percolation Method, COPRA (Gregory, 2010), etc. COPRA is very fast, but the result of COPRA is nondeterministic, so we propose an improved COPRA with high determinacy in this paper.

An Improved COPRA Algorithm Based on Connecting Degree

To eliminate the nondeterministic of COPRA, we use Connecting Degree as definition 1.

Definition 1: Let v be a node on the undirected Graph $G(V;E)$, C is the set of overlapped communities on Graph, the connecting degree between node v and community $c(c \in C)$, denoted $C(v,c)$, be computed by the following formula (Duanbing, Mingsheng, Xia, 2013).:

$$C(v,c) = \frac{\sum_{u \in c} w_{vu}}{k_v} \qquad (1)$$

Where k_v is the degree of node v , w_{vu}=1 if there is an edge between node v and node u , zero otherwise.*

Connecting Degree can reflect the community tendency for a node to its neighbour communities, so we proposed a COPRA Based on Connecting Degree, named COPRA-CD. COPRA-CD works as follows: 1) To start, all nodes are initialized with a unique community identifier and a belonging coefficent setting to 1; 2) Each node updates its community identifier by the union of its neighbours labels, the corresponding belonging coefficient is obtained by normalizing the sum of the belonging coefficients of the communities over all neighbours. Then, comparing all the belonging coefficients and the parameter v, if all the belonging coefficients are less than v, calculating the connecting degree between node and its neighbour community, then only retain neighbour community with greatest connecting degree, else keeping these belonging coefficients that are more than v, then renormalize these belonging coefficients of remaining communities so that they sum to 1. After several iterations, if the stop criteria proposed by Gregory is satisfied, the propagation procedure stops; 3) Remove communities that are totally contained by others; 4) Split disconnected communities.

Experimental Results and Discussion

Test networks

At first, we do experiments on four real-world networks, whose information are shown in Table 1.

Table 1. General information of real networks

Networks	Description	Node&Edge
Karate	Zachary's karate club (Zachary, 1977)	34 &78
Dolphin	Lusseau's Dolphins (Lusseau, 2003)	62 & 159
Books	Books about US politics	105 & 441
Football	American College football union (Girvan, Newman, 2002)	115 & 616

Then we also test the performance of COPRA-CD on six LFR synthetic networks with various mixing parameter μ ranging from 0.1 to 0.6, the other standard configuration of LFR synthetic network used in this experiment is: n=1000, t_1=2, t_2=1, k=10, $\max k$=30, $\min c$=10, $\max c$=50, O_n=100, O_m=2.

Test metrics

To measure overlapping communities detection, Q_{ov} was be proposed by Nicosia et al (2009). The formulation of Q_{ov} as following:

$$Q_{ov} = \frac{1}{m} \sum_{c \in C} \sum_{i,j \in V} \left[\beta_{l(i,j),c} A_{ij} - \frac{\beta^{out}_{l(i,j),c} k_i^{out} \beta^{in}_{l(i,j),c} k_j^{in}}{m} \right] \qquad (2)$$

Where A_{ij} is the adjacency matrix of Direct Graph $G(E,V)$, C is the set of overlapped

* Corr. author: C. Zhang, Tel: +86-25-84315963.

communities, $l(i,j)$ is a link which starts at node i and ends at node j . $\beta_{l(i,j),c}$ is the belonging coefficient of $l(i,j)$ for community c , $\beta^{out}_{l(i,j),c}$ is the expected belonging coefficient of any possible link $l(i,j)$ starting from a node into community c , $\beta^{in}_{l(i,j),c}$ is the expected belonging coefficient of any link $l(i,j)$ pointing to a node going into community c . k_i^{out} is the out degree of node i , while k_j^{in} is the in degree of node j .

Test results and discussion

In order to show its performance, we compare three multi-label propagation algorithms, i.e., COPRA, COPRA-CD, and RC-COPRA. RC-COPRA stands for the version of COPRA with initialization using RC proposed by Wu et al. (2012). In our test, we run each algorithm 100 times on each network for the same value of parameter v . The average modularity result on real-world network was shown in Table 2, and the comparison performance on LFR synthetic networks was shown in Figure 1.

Table 2. Test Results on real-world Networks.

Networks	**COPRA** **(v =2)**	**COPRA-CD** **(v =2)**	**RC_COPRA** **(v =2)**
Karate	0.428	0.745	0.703
Dolphins	0.645	0.759	0.761
Books	0.826	0.815	0.830
Football	0.684	0.661	0.668
Networks	**COPRA** **(v =3)**	**COPRA-CD** **(v =3)**	**RC_COPRA** **(v =3)**
Karate	0.408	0.717	0.725
Dolphins	0.652	0.710	0.713
Books	0.830	0.822	0.827
Football	0.677	0.665	0.670

From Table 2, we find the modularity of CORPA is lower than that of other algorithms at the same v . At v =3, RC_COPRA algorithm gives better average modularity for every network, but at v=2, the modularity of RC_COPRA algorithm on Karate network is not better than that of COPRA-CD.

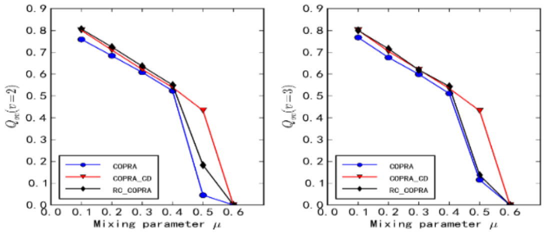

Figure 1. Experiment on synthetic networks.

As Figure 1 shows, when $\mu \le 0.4$, all three algorithms show good performance. When $\mu = 0.5$, LFR synthetic networks are very fuzzy, the overlapping community structure is not detected by COPRA and RC_COPRA, but detected by COPRA-CD, so we can conclude that for the given parameter, COPRA-CD is the most stable algorithm in these overlapping community detection algorithms.

Conclusions

In this paper, we propose COPRA-CD to uncover overlapping communities in social networks. Then we test it on four real-word networks and a group of synthetic networks. Experimental results show that both RC initialization and the connecting degree update strategy can bring improvements in quality, especially COPRA-CD has the best stability for fuzzy networks. In the future, COPRA-CD can be applied to analyze the community of co-author in paper.

Acknowledgments

This work is supported by the National Social Science Fund Project (grant number 14BTQ033).

References

Duanbing, C., Mingsheng, S., & Xia, L. (2013). Two-phase strategy on overlapping communities detection. *Computer Science, 40*(1), 225-228.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *PNAS, 99*(12), 7821-7826.

Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, *12*(10), 103018.

Hofman, J. M., & Wiggins, C. H. (2008). Bayesian approach to network modularity. *Physical review letters, 100*(25), 258701.

Krebs, V. (2008). A network of co-purchased books about US politics, www.orgnet.com.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270 (Suppl 2), S186-S188.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167-256.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E, 69*(6), 066133.

Nicosia, V., Mangioni, G., Carchiolo, V., & Malgeri, M. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment,* (03), P03024.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814-818.

Wu, Z.-H., Lin, Y.-F., Gregory, S., Wan, H.-Y., & Tian, S.-F. (2012). Balanced multi-label propagation for overlapping community detection in social networks. *Journal of Computer Science and Technology, 27*(3), 468-479.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research,* 452-473.