2015 International Symposium on Nonlinear Theory and its Applications
NOLTA2015, Kowloon, Hong Kong, China, December 1-4, 2015

Pseudo Clique Community Analysis for Social Network

Atsushi Tanaka†

†Graduate School of Science and Engineering, Yamagata University
4-3-16 Jonan, Yonezawa, 992-8510 Japan
Email: tanaka@yamagata-u.ac.jp

Abstract—In this paper, a new community analysis method with overlapping nodes suitable especially for social networks is proposed. It is a natural extension of well-known Clique Percolation Method(CPM). The essence of our method is reducing the conditions of CPM to find missing potential communities. To determine the systematic and optimal parameter values for our method still remains for future work.

1. Introduction

Thanks to the power of statistical physics, the progress of study of complex networks is remarkable these ten years. Study of networks based on the relationship has its roots in graph theory by L. Euler. After that in sociology many studies of human activity and social structure based on their relationships have been carried out. Since friendship between two people is invisible generally, its analysis has been performed by the method with high indeterminacy like questionnaire data. While recent diffusion of social media enabled us to observe and obtain relationships among people and the stream of information propagation easily, so their networks have been studied energetically. As social media has developed rapidly, their observed network data became enormous and the development of computers have enabled us to analyze them.

In this paper, among several network analysis we focus on community analysis and aim to pick up hidden information from the network. In that process we compare our proposed community analysis method and existing ones and discuss the validity and possibility of our method.

2. Community Analysis

Community analysis researches have been carried out for several purposes like the analysis of the strength or robustness of network structure and the extraction of potential communities and so on. In traditional sociology, several methods like dendrogram, K-means method etc. have been mainly used for that. From scale-free property, which one of the most important keywords in complex network, each node is not uniform in most case, and a community analysis method based of the difference of importance of nodes and edges has developed[2]. The essence of this method is betweenness and its accuracy though, it costs a lot of computational time and some improved methods have been developed and widely used these years[3, 4, 6].

However all above methods divide network definitely into some groups, so they are not proper for social network where each node can belong to different communities at the same time. Thus some overlapping community analysis methods have been studied. The most representative one is Clique Percolation Method(CPM)[5] and it has many application. This method is based on k-clique and two communities are fused if $(k-1)$ nodes are shared among them. As we continue this process as possible as we can, we can obtain widely spread communities.

3. Improvement of CPM

Though CPM is rather simpled and effective method, its condition is sometimes too strict to detect proper communities. Thus we have proposed ACPM which we loosed a little its condition[1]. In that method we make the fusion condition variable, and observe widely the dependence of the parameter to determine the proper decomposition. If we select the condition as $k-1$, it is exactly the same as CPM, so we can consider our method as the natural extension of CPM.

Even in this method the base of units is still k-clique, so all nodes that cannot form any cliques are excluded from the beginning. We can say it is rather strict condition. Thus we extend our method to use not k-clique but pseudo k-clique. Pseudo means we do not constrain complete clique and permit a few lacks of edges. That leads to ACPM with alleviated initial conditions. For a example, let us consider a graph with 8 nodes. If it composes a complete clique, the total number of edges is 28. Even if two of them are missing, the occupation rate of edges is $26/28 \doteq 0.93$. It might be reasonable that we regard it as a pseudo clique in the ordinal meaning.

4. Community analysis of SNS network

Our method is suitable to detect communities in social network, where a lot of nodes, typically people belong to some regional, educational, business, hobby and other communities at the same time. Since in human network, real relationships seldom become open, it is almost impossible to assemble their network data. On the contrary, it is relatively easy to obtain network data on social media

like SNS in recent years. Thus many researches of network structures and communities on them have been proceeded.

In order to investigate the network dynamics in SNS, we have constructed SNS site a.k.a. "tomocom.jp", and analyzed its network data. This SNS is only for college students and completely invitation-based system, that is different from other popular SNS systems. Several college teachers all over Japan invited their seminar students to this SNS for their activities. It possesses higher reliability thanks to its guardian system. This site was opened in 2009, and captured more than 400 users by the end of 2010, and many students in several colleges have been still making use of it. Though the interaction within their own seminars is basic, the connections between different seminars have been created by writing and browsing their blogs. Since activation event of the site over several times were held, the connections became denser. The network structure of this SNS at the end of 2010 is shown in Fig.1(a).

5. Results and Discussion

According to geometrical and social metrics,connections inside the same seminar are so strong that communities based on them tend to be created preferentially. Thus the number of seminars is one of the most important and basic measure for the estimation of our method. That is, it provides a case in which there exists a correct solution of the number of communities.

The parameters in our simulation are as follows.

k : Degree of cliques.

α : Allowance parameter for pseudo cliques.

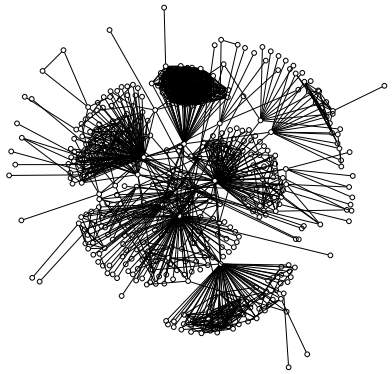
$$= \frac{\text{Number of edges in pseudo cliques}}{k(k-1)/2}$$

σ : Number of shared nodes to fuse cliques.

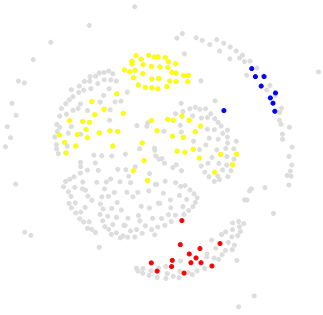
In case of $k = 6$ and $\alpha = 0.95$, we demonstrate the whole network and the change of detected communities for $\sigma = 1, 3, 5$ in Fig.1.

In that figure, yellow, red and blue circles are indicated the first, second and third largest communities respectively. We can find the larger the fusing parameter σ becomes, the smaller the size of communities becomes. It is very important and still open that what are the optimal values of k, αandσ in each network and how we find to decide them in generally.

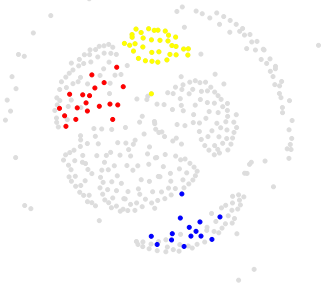
If we fix the degree of clique k, we can vary the parameter σ from 1 to $k-1$. For each k and σ, we can calculate the number of communities using ACPM as shown in Table 1. As you can see, the larger k becomes, the smaller the number of communities becomes. However in large k, the number is small and almost constant. Moreover if we fix k, the larger σ becomes, the larger the number becomes. That is because the increase of σ means tightening the condition of the fusion of cliques, and the fusion becomes difficult. For pseudo cliques, similar characteristics are observed as



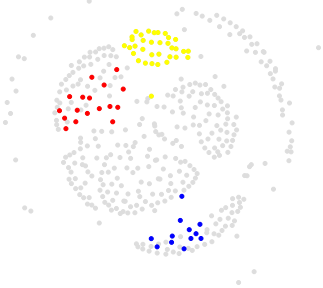
(a) Whole network of tomocom.jp



(b) $k = 6, \alpha = 0.95, \sigma = 1$

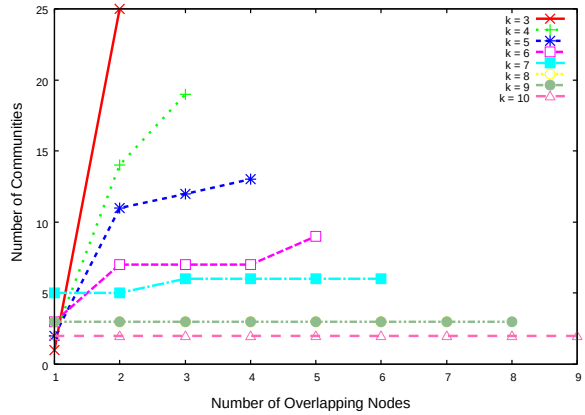


(c) $k = 6, \alpha = 0.95, \sigma = 3$

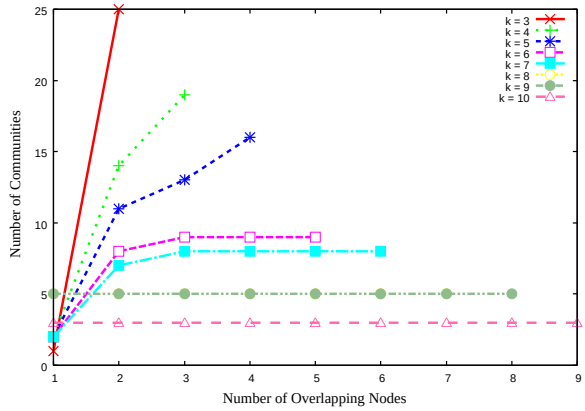


(d) $k = 6, \alpha = 0.95, \sigma = 5$

Figure 1: Whole network structure and community detection in tomocom.jp using Alternative CPM with pseudo cliques.



(a)



(b)

Figure 2: Change of number of communities in tomocom.jp using (a) ACPM and (b) pseudo ACPM.

shown in Table 2 though, the numbers of communities become a little large. These tables gives us some important aspects of our community analysis.

On the way we decrease k, the marginal region of such states can be found. The number of communities there is likely to be optimal. Namely no changes imply obviously different communities, and the hypothesis that the marginal region provides the optimal solution is considered to be natural.

The most important problems to be solved are which σ, the parameter of overlapping nodes, is optimal , and what extent conditions can be loosened in pseudo cliques. Of course we have not solved that problem yet, and is one of the most important issues. However, as a promising possibility, we pay attention to the change of communities along with the change of the degree of cliques and the parameter σ. As shown in Table 1 and 2, if we change these parameters, the number of communities also change. In large k, the number of communities is almost constant for all σ. On the other hand, in small k, it changes significantly.

For each degree k in Fig.2(a), the number of communities is increasing according to the increase of overlapping nodes σ. For large k, the change becomes small though, the transition is not clear. On the contrary, for the pseudo cliques, we can find the critical value of the transition in Fig.2(b). Therefore the method with pseudo cliques is promising to find optimal parameters for this community analysis.

From the computational time point of view, the cost of finding cliques or pseudo cliques is very high and dominant in this community extraction algorithm. We have to deal with the combination of k nodes from all the nodes, then it causes the explosion of the cost. In fact, our computational time of our method is very high. However to solve this difficulty, Zero-suppressed binary Decision Diagram(ZDD) for combinatorial problem proposed by Minato[7] is very promising.

6. Summary

We proposed new community analysis method with overlapping nodes suitable especially for social networks. It is a natural extension of well-known Clique Percolation Method(CPM). If we start with pseudo cliques, not normal cliques, we can find new community decomposition. By using pseudo cliques we can discriminate the critical value for community analysis. To determine the systematic and optimal parameter values for our method still remains for future work.

References

[1] A. Tanaka, “Proposal of Alleviative Method of Community Analysis with Overlapping Nodes”, Proc. The Seventh IEEE Int. Conf. on Social Computing and Networking, C.1.30-5, 2014.

[2] M. Girvan, M. and M. E. J. Newman, “Community structure in social and biological networks.” *Proc. Natl. Acad. Sci. U.S.A.*, Vol.99, No.12, pp.7821–7826, 2002.

[3] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Physical Review E*, Vol.69, 066133, 2004.

[4] A. Clauset, M. E. J. Newman and C. Moore, “Finding community structure in very large networks,” *Physical Review E*, Vol.70, 066111, 2004.

[5] G. Palla, I. Derényi, I. Farkas and T. Vicsek, “Uncovering the overlapping modular structure of protein interaction networks.” *FEBS JOURNAL* 272: pp.434–434 Suppl. 1, 2005.

[6] S. Fortunato, “Community detection in graphs”, *Physics Report*, Vol. 486, Issues 3-5, pp.75–174, 2010.

[7] Shin-ichi Minato, “Zero-suppressed BDDs for set manipulation in combinatorial problems”, DAC ’93: Proceedings of the 30th international conference on Design automation, 1993.

Table 1: Number of communities in tomocom.jp using ACPM. k, σ are order of initial clique and minimum overlapping nodes respectively.

σ \ k	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	2	2	3	5	3	3	2	2	1	1	1	1	1	1	1	1	1
2	25	14	11	7	5	3	3	2	2	1	1	1	1	1	1	1	1	1
3		19	12	7	6	3	3	2	2	1	1	1	1	1	1	1	1	1
4			13	7	6	3	3	2	2	1	1	1	1	1	1	1	1	1
5				9	6	3	3	2	2	1	1	1	1	1	1	1	1	1
6					6	3	3	2	2	1	1	1	1	1	1	1	1	1
7						3	3	2	2	1	1	1	1	1	1	1	1	1
8							3	2	2	1	1	1	1	1	1	1	1	1
9								2	2	1	1	1	1	1	1	1	1	1
10									2	1	1	1	1	1	1	1	1	1
11										1	1	1	1	1	1	1	1	1
12											1	1	1	1	1	1	1	1
13												1	1	1	1	1	1	1
14													1	1	1	1	1	1
15														1	1	1	1	1
16															1	1	1	1
17																1	1	1
18																	1	1
19																		1

Table 2: Number of communities in tomocom.jp using pseudo-ACPM. k, σ are order of initial clique and minimum overlapping nodes respectively.

σ \ k	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	2	2	2	2	5	5	3	3	2	1	1	1	1	1	1	1	1
2	25	14	11	8	7	5	5	3	3	2	1	1	1	1	1	1	1	1
3		19	13	9	8	5	5	3	3	2	1	1	1	1	1	1	1	1
4			16	9	9	5	5	3	3	2	1	1	1	1	1	1	1	1
5				9	9	5	5	3	3	2	1	1	1	1	1	1	1	1
6					9	5	5	3	3	2	1	1	1	1	1	1	1	1
7						5	5	3	3	2	1	1	1	1	1	1	1	1
8							5	3	3	2	1	1	1	1	1	1	1	1
9								3	3	2	1	1	1	1	1	1	1	1
10									3	2	1	1	1	1	1	1	1	1
11										2	1	1	1	1	1	1	1	1
12											1	1	1	1	1	1	1	1
13												1	1	1	1	1	1	1
14													1	1	1	1	1	1
15														1	1	1	1	1
16															1	1	1	1
17																1	1	1
18																	1	1
19																		1