A Novel Community Detection Method Based on Cluster Density Peaks

Donglei Liu[1], Yipeng SU[1], Xudong Li[1], and Zhendong Niu[1]⋆

School of Comupter Science, Beijing Institute of Technology,
5 South Zhongguancun Street, Haidian District, Beijing, China
{liudonglei,su_yipeng,lixudong,zniu}@bit.edu.cn

Abstract. Community structure is the basic structure of a social network. Nodes of a social network can naturally form communities. More specifically, nodes are densely connected with each other within the same community while sparsely between different communities. Community detection is an important task in understanding the features of networks and graph analysis. At present there exist many community detection methods which aim to reveal the latent community structure of a social network, such as graph-based methods and heuristic-information-based methods. However, the approaches based on graph theory are complex and with high computing expensive. In this paper, we extend the density concept and propose a density peaks based community detection method. This method firstly computes two metrics-the local density ρ and minimum climb distance δ -for each node in a network, then identify the nodes with both higher ρ and δ in local fields as each community center. Finally, rest nodes are assigned with corresponding community labels. The complete process of this method is simple but efficient. We test our approach on four classic baseline datasets. Experimental results demonstrate that the proposed method based on density peaks is more accurate and with low computational complexity.

Keywords: social network, community detection, density peak

1 Introduction

Social activities and social relations of people in real life constitute a network of relationships which is called social network. Each node in the network represents a person, and the edges represent some kinds of social relations between nodes, such as friendship, family relations. With the development of Internet technology and web services application, more and more people migrate their real social relationships to the online social network sites (OSN), such as the famous Facebook, Twitter, WeChat and Sina weibo site.

Online social network sites could map and extend the social network of people. On OSN sites people can easily post their daily activities, pictures, comments, repost their friends' posts, follow new friends, and manage their friends

⋆ (corresponding author, Zhendong Niu)

groups. Online social network sites greatly facilitate people's social activities, and makes it easier to build new friendships and to manage their existing relationships between friends.

Community structure is the basic structure of a social network, that is to say, nodes of a social network will spontaneously form groups of nodes,which are called communities. There exist many community detection methods based on graph theory and heuristic information, such as Infomap [18][17], Fastgreedy [5] and Louvain [2], the state-of-the-art greedy method.

However, the graph-based methods are complex and have high computational complexity. In this paper, we extend the density concept and propose a density peaks based community detetion method. This method firstly computes two metrics:the local density ρ and minimum climb distance δ for each node in a network, then treats the nodes with both higher ρ and δ as a local extreme point and a community center. Finally, Each node is assigned to the community label which the closest community center belongs to.The complete process of method is simple but efficient. We tested our method on four classic baseline datasets and a large dataset. Experimental results demonstrate that the proposed method based on density peaks is relatively accurate and with low computational complexity.

To address the above challenges, we propose a density-peaks based community detection approach, and summarize our technical contributions as follows:

- We extend the original density peak based cluster method to find the communities in social network.

The rest of this paper is organized as follows. Section 2 reviews the current community detection work. Section 3 presents our proposed algorithm based on density peaks. Section 4 presents the experiments and results analysis. Finally we conclude in Section 5.

2 Related Work

2.1 Social Network Definition

In order to describe the community detection problem accurately, we first introduce notations used in this paper. A social network, in mathematical context, can be formulated as a graph $G = (V, E)$ consisting of a set of nodes V representing the users, and a set of edges E denoting the relationships between users (e.g. followees or followers). $|V| = N$ denotes the number of nodes, and $(i, j) \in E$ denotes the edge from node i to node j $(i, j \in V)$, where A is the adjacency matrix of the network and $A_{ij} = 1$ represents the existence of edge (i, j) and $A_{ij} = 0$, otherwise.

2.2 Classical Community Detection Methods

Community detection methods aim to reveal the latent community structure in the social network. There exist a wide range of different community detection algorithms which follow different strategies in the literature [9].

Girvan-Newman method The Girvan-Newman method is a seminal method used to detect communities in complex systems [11]. This method is based on edge's betweenness which identifies number of shortest-path passing this edge. The idea is that the betweenness of the edges connecting two communities is typically high, as many of the shortest paths between nodes in separate communities go through them. By removing edges owning maximum betweenness iteratively, the complex network is divided into many reasonable communities, and the result is a dendrogram. The computational complexity of GN method is $O(m^2n)$, where m is the number of edges and n is the number of nodes.

Modularity-based methods The family of methods based on maximizing modularity is the biggest one in community detection algorithms. Modularity scores high those partitions containing communities with an internal edge density larger than that expected in a given graph model, which is almost always an ER model [14][15].

Several strategies have been proposed for modularity optimization, such as agglomerative greedy [5], Fast Newman [13], CNM method [4]. A multilevel approach Louvain has been proposed which scales to graphs with hundreds of millions of objects [2]. However, it has been reported that modularity has resolution limits [10]. Modularity is unable to detect small and well defined communities when the graph is large, and its maximization delivers sets with a tree-like structure, which cannot be considered communities.

2.3 Density Cluster-Based Method

There are several density cluster-based methods, such as SCAN(A Structural Clustering Algorithm for Networks, SCAN) [20] and DENGRAPH-ho(Density-based hierarchical community detection for explorative visual network analysis) [19].

The SCAN algorithm, derived from DBSCAN(Density-based Spatial Clustering of Applications with Noise, DBSCAN) [7], is capable of discovering communities, hubs, and outliers in a network. A community is grown from a group of centralized nodes which all satisfy a given neighborhood size. A user-defined threshold ε is introduced to define the neighborhood of a node. SCAN uses the ε-neighborhood of a node and groups it with those who share a common set of neighbors. A structural similarity measure is used to calculate the similarity between two nodes.

DENGRAPH-ho algorithm uses its own DENGRAPH(DENGRAPH: A Density-based Community Detection Algorithm) [8] density cluster method which derived from DBSCAN to update the current community structure of a network from a previously detected structure and its changes over time. This method can discover overlapping communities, by allowing each node to inherit multiple community labels instead of one. Also, to define a density-based neighborhood of a node, DENGRAPH uses the distance between two nodes, while SCAN uses neighborhood similarity.

3 Community Detection by Cluster Density Peaks

DensityPeak [16] is a new density based method proposed by Rodriguez and Laio. The basic idea of this method is that cluster centers are characterized by a higher density than their neighbors' and by a relatively large distance from points with higher densities. This method is simple and can recognize clusters regardless of their shape and the dimensionality of the space in which they are embedded. However, this method is developed for data in Euclidean space, and not fit for network data. Based on the similar idea, we propose a novel community detection method based on density peaks. We firstly check whether the social network addresses density peak phenomenon–the density peaks which with both higher ρ and δ will be a community center, then explore which definition of the metric ρ and δ can better reflect this phenomenon.

3.1 Definitions of Local Density ρ and Minimum Climb Distance δ

Rodriguez's new density peaks-based clustering method defines two metrics for each node. Because the definitons of the two metrics are about vector space, and cannot be adapted to network data straightforwardly, we extend the two metrics. We call the two metrics as the local density ρ and minimum climb distance δ separately according to their semantic meanings.

$$\rho_i = \frac{1}{2} \sum_{k,j \in \{i\} \cup F_i} A_{kj} \ . \tag{1}$$

where A is the adjacency matrix of the network defined in Section 2.1, and F_i is the set of nodes which directly connect with node i. So the ρ_i means the number of edge of the subgraph formed by node i and its neighbors.

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \ . \tag{2}$$

where d_{ij} denotes the length of shortest path between node i and j. For the node with highest density $\delta_i = \max_j(d_{ij})$.

For each node i, local density ρ_i is the number of edge of the subgraph formed by node i and its neighbors, and minimum climb distance δ_i is the minimum distance to nodes whose local density is higher than the node i. We meaningfully call this distance as climb distance, because this distance δ_i is the minimum distance from the node i up to any node with higher density than node i.

Rodriguez found the phenomenon that each cluster center point has higher local density and longer minimum distance, like peaks of the mountain, and other nodes have lower local density.

In our approach, we use the length of shortest path between nodes as the distance between every two nodes, which satisfies the triangle inequality condition required by Rodriguez's method. For the local density metric, we examine two different measurements, such as the node degree and the number of edges of the subgraph formed by node and its neighbors, and we find that the latter definition is better.

3.2 Procedure of the Community Detection by Cluster Density Peaks

We propose the community detection method by Density Peaks (DP-D). We firstly computes two metrics-the local density ρ and minimum climb distance δ for each node in a network, then treats the nodes with both higher ρ and δ as a local extreme point and a community center. Finally, each node is assigned with the label found by the density weighted major voting label propagation process.

Thus the general form of our community structure finding algorithm is as follows:

1. For each node i, computing the number of edges of the subgraph formed by node i and its neighbors as the local density ρ_i for node i.
2. For each node i, computing the minimum distance to nodes whose local density is higher than the node i as the minimum climb distance δ_i for node i. To decrease the computing time, we can search the k-hop neighbors of node i and get the first node j with higher ρ_j than node i, then $\delta_i = k$.
3. Ploting the 2-D decision graph by using the ρ as the x-axis and δ as y-axis. For example, Figure 2 shows the decision graph for the Zarchary's Karate Club network whose network structure is plotted as Figure 1. The nodes in the upper right region of Figure 2(a) have higher local density ρ and longer minimum distance δ, such as the two nodes in Figure 2. Using these nodes as community centers. To easily select the centers, we also plot the γ-rank figure as Figure 2(b) where $\gamma_i = \rho_i * \delta_i$.
4. Every rest node is assigned to the label found by the density weighted major voting label propagation process. That is to say, we firstly count the sum of density of each community label of the labeled neighbors of node i, and then assign the community label with maximum sum of density to node i.

Complexity. The time complexity of step 1 described above is $O(m+n)$, where n is the number of nodes and m is the number of edges. The time complexity of step 2 is $O(n)$. The time complexity of step 4 is $O(n)$. The total time complexity of our methods is $O(m + n)$.

4 Experiments

In this section we provide an overview of the datasets and methods which we will use in our experiments.

4.1 Datasets

An overview of the networks we consider in our experimetns is given in Table 1.

Zarchary's Karate Club Network [21]. The well-known Zarchary's karate network is a classic society network. This network contains 34 nodes and 78 edges which represent members in a karate network and connections between them.

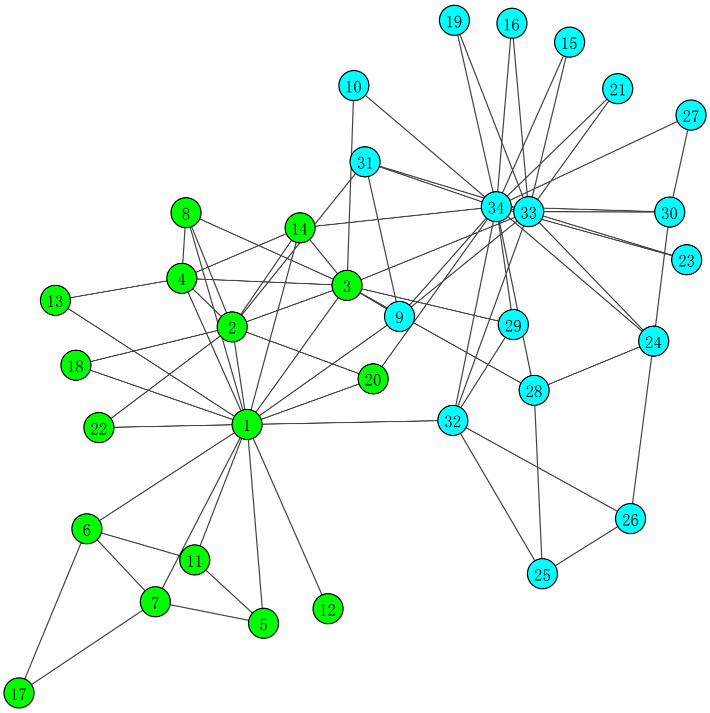
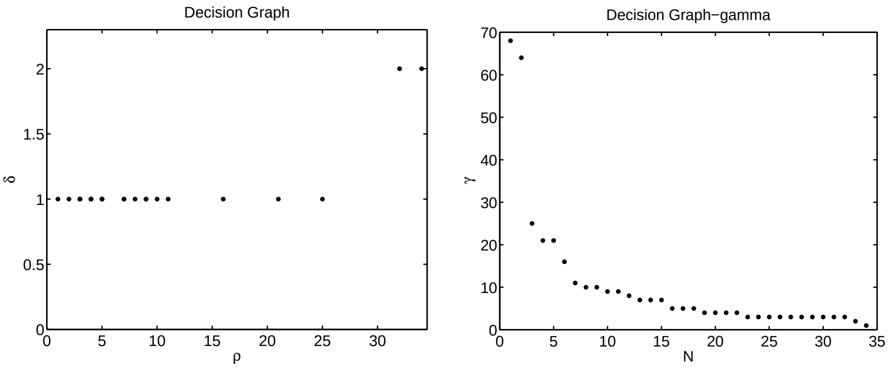


Fig. 1. Zarchary's Karate Club Network. the two communities are denoted by different colors



(a) Decision Graph I (b) Decision Graph II

Fig. 2. Decision Graph for the Zarchary's Karate Club Network. (a) $\delta - \rho$ (b) $\gamma - rank$ The top two nodes are community centers

Table 1. classic social networks

Datasets	Nodes	Edges	Communities	Q-value
karate	34	78	2	0.3715
dolphins	62	159	2	0.3787
polbooks	105	441	3	0.4149
polblogs	1222	19089	2	0.4052

In real world, members of the club are separated into two communities because of the dispute between club administrator (node 1) and principal karate teacher (node 33). Figure 1 shows the whole network.

Dolphin Social Network [12]. The Dolphins network is an undirected social network of frequent associations between 62 bottlenose dolphins in a community living in Doubtful Sound, New Zealand. The network was compiled from seven years of field studies of the dolphins, with ties between dolphin pairs being established by observation of statistically significant frequent association. The network splits naturally into two groups. The split into two groups appears to correspond to a known division of the dolphin community [12].

Books about US politics. A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent copurchasing of books by the same buyers. The network was compiled by V. Krebs and can be found in this websit. [1]

Political Blogs [1]. Those blogs form two communities according to their political attitude. Links between blogs were automatically extracted from a crawl of the front page of the blog. In addition, the authors drew on various sources (blog directories, and incoming and outgoing links and posts around the time of the 2004 presidential election) and classified the first 758 blogs as left-leaning and the remaining 732 as right-leaning. In our experiments, we remove the isolated nodes and just focus on the maximum component which has 1222 nodes and 19089 edges.

4.2 Evaluation Metrics

Modularity We use the classic modularity measure as the metric for our experiments. Modularity is the most used measure in evaluating the quality of communities found by community detection algorithms. The modularity of a partition is a scalar value between -1 and 1 that measures the density of links inside communities compared to links between communities. Modularity [14] [15] is simply defined as:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\phi(c_i, c_j) \qquad (3)$$

[1] http://www-personal.umich.edu/ mejn/netdata/

where A is the adjacency matrix of the network, k_i is the degree of node i, and $\phi(c_i, c_j)$ is 1 if nodes i and j have the same community membership, and 0 otherwise, and $m = \frac{1}{2}\sum_{i,j} A_{ij}$. This definition indicates that for each node pair (i, j) which shares communities, its contribution to modularity is positive if i, j are linked and is negative otherwise. It matches our intuition that nodes inside one community tends to build links with each other.

Modularity has been used to compare the quality of the partitions obtained by different community detection methods, but also as an objective function to optimize [13]. Unfortunately, it is computationally expensive to search all such partitions for finding the optimal value of modularity since modularity optimization is NP-hard problem [3]. However, many approximation methods were introduced to find high-modularity partitions to deal with large network in a reasonable time, such as the Louvain method [2].

NMI The normalized mutual information (NMI) [6] is an information-theoretic-based measurement. It is currently widely used in measuring the performance of clustering algorithms. Formally, the measurement metric NMI can be defined as

$$NMI(A, B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij} log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{C_A} N_{i.} log(\frac{N_{i.}}{N}) + \sum_{j=1}^{C_B} N_{j.} log(\frac{N_{.j}}{N})} \tag{4}$$

where N is the confusion matrix, where the rows correspond to the real communities, and the columns correspond to the found communities. N_{ij} is the number of nodes in the real community i that appear in the found community j. The number of real communities is denoted C_A and the number of found communities is denoted C_B, the sum over row i of matrix N_{ij} is denoted $N_{i.}$ and the sum over column j is denoted $N_{.j}$. If the found partitions are identical to the real communities, then NMI takes its maximum value of 1. If the partition found by the algorithm is totally independent of the real partition, for example when the entire network is found to be one community, NMI = 0. In our experiments, we use the NMI metric to measure the difference between the communities found by methods and the ground-truth.

4.3 Comparison Methods

We select three methods to compare with our method:

1. Fastgreedy method [5]: The Fastgreedy community detection algorithm is an algorithm based on the greedy optimization of modularity. This algorithm merges individual nodes into communities in a way that greedily maximizes the modularity score of the graph. It can be proven that if no merge can increase the current modularity score, the algorithm can be stopped since no further increase can be achieved.
2. Louvain method [2]: The Louvain community detection algorithm is an algorithm for performing community detection in networks by maximizing a modularity function which uses local moving strategy to greedily maximize

the modularity of the structure after processed by the Louvain method. It starts with all vertices in clusters by themselves. Then, for each vertex, it tries to reassign the vertex to the cluster of its neighbor which increases the modularity value the most. If reassigning to a neighbor's cluster does not increase the modularity value, it stays with its current cluster. This process repeats until no vertices can find a better cluster to be reassigned to. The algorithm then contracts each cluster into a supervertex, keeping track of the number of multiple edges between the clusters as the edge weight. The self loops are also kept. The whole process is then repeated on this new graph until the contraction does not reduce the number of nodes. The Louvain algorithm is fast and produces good solutions in practice. In this paper, we use the freely available C++ implementation of the method written by E. Lefebvre [2] to conduct the comparision experiments.

3. Infomap method [18][17]: The Infomap algorithm is based on the principles of information theory. Infomap characterizes the problem of finding the optimal clustering of a graph as the problem of finding a description of minimum information of a random walk on the graph. The algorithm maximizes an objective function called the Minimum Description Length, and in practice an acceptable approximation to the optimal solution can be found quickly. In this paper, we use the freely available Python implementation of the Python-igraph package [3] to conduct the comparision experiments.

4.4 Experiment Results and Analysis

Table 2. Experiment Results NMI-value

Datasets	Fastgreedy	Infomap	Louvain	DP-D
karate	0.6925	0.6995	0.5866	**1.0**
dolphins	0.5727	0.5662	0.6647	**1.0**
polbooks	0.5308	0.4935	0.5125	**0.6012**
polblogs	0.6461	0.4872	0.6440	**0.6633**

Table 2 and Table 3 show the Q and NMI values of communities found by our method and baseline methods on four datasets.

From Table 2, we can see that communities found by our proposed method is identical to the real communities on karate and dolphins datasets (NMI=1.0). For the two large datasets polbooks and polblogs, our method also get higher NMI than other methods. The modularity values of our method are lower than those found by other methods, because those method are specifically designed to maximize the modularity value but our method is not. Our method mainly reveals the real community structure whose modularity value maybe not high.

[2] http://perso.uclouvain.be/vincent.blondel/research/louvain.html

[3] python-igraph package

Table 3. Experiment Results Q-value

Datasets	Groundtruth	Fastgreedy	Infomap	Louvain	DP-D
karate	0.3715	0.3807	0.4020	**0.4188**	0.3715
dolphins	0.3787	0.4955	**0.5277**	0.5185	0.3787
polbooks	0.4149	0.5019	**0.5228**	0.5205	0.4495
polblogs	0.4052	0.4269	0.4227	**0.4270**	0.4200

Furthermore, our experiments show that the social network also have the phenomenon that each cluster center points have higher local density and longer minimum distance, like peaks of the mountain, and other nodes have lower local density. If we treat the local density of one node as the personal influence in his social network, this phenomenon may mean that the person prefers to attach connection with whom with higher influence.

5 Conclusions

In this paper, We proposed a simple but efficient community detection method based on cluster density peaks. Our method can mainly reveal the real community structure with high NMI. For the future work, we can extend this method to find overlapping communities where nodes may belong to many different communities.

References

1. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 u.s. election: Divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery. pp. 36–43. LinkKDD '05, ACM, New York, NY, USA (2005), `http://doi.acm.org/10.1145/1134271.1134277`
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
3. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Trans. on Knowl. and Data Eng. 20(2), 172–188 (Feb 2008), `http://dx.doi.org/10.1109/TKDE.2007.190689`
4. Clauset, A.: Finding local community structure in networks. Phys. Rev. E 72, 026132 (Aug 2005), `http://link.aps.org/doi/10.1103/PhysRevE.72.026132`
5. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70, 066111 (Dec 2004), `http://link.aps.org/doi/10.1103/PhysRevE.70.066111`
6. Danon, L., Díaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment 2005(09), P09008 (2005), `http://stacks.iop.org/1742-5468/2005/i=09/a=P09008`
7. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial

databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231. KDD'96, AAAI Press (1996), `http://dl.acm.org/citation.cfm?id=3001460.3001507`

8. Falkowski, T., Barth, A., Spiliopoulou, M.: Dengraph: A density-based community detection algorithm. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. pp. 112–115. WI '07, IEEE Computer Society, Washington, DC, USA (2007), `http://dx.doi.org/10.1109/WI.2007.43`
9. Fortunato, S.: Community detection in graphs. Physics Reports 486(3–5), 75 – 174 (2010), `http://www.sciencedirect.com/science/article/pii/S0370157309002841`
10. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences 104(1), 36–41 (2007)
11. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826 (2002), `http://www.pnas.org/content/99/12/7821.abstract`
12. Lusseau, D., Schneider, K., Boisseau, O., Haase, P., Slooten, E., Dawson, S.: The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology 54(4), 396–405 (2003), `http://dx.doi.org/10.1007/s00265-003-0651-y`
13. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 066133 (Jun 2004), `http://link.aps.org/doi/10.1103/PhysRevE.69.066133`
14. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (Feb 2004), `http://link.aps.org/doi/10.1103/PhysRevE.69.026113`
15. Newman, M.E.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)
16. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science 344(6191), 1492–1496 (2014), `http://www.sciencemag.org/content/344/6191/1492.abstract`
17. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. The European Physical Journal Special Topics 178(1), 13–23 (2009), `http://dx.doi.org/10.1140/epjst/e2010-01179-1`
18. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA p. 1118 (2008)
19. Schlitter, N., Falkowski, T., et al.: Dengraph-ho: Density-based hierarchical community detection for explorative visual network analysis. In: Research and Development in Intelligent Systems XXVIII, pp. 283–296. Springer (2011)
20. Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.J.: Scan: A structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 824–833. KDD '07, ACM, New York, NY, USA (2007), `http://doi.acm.org/10.1145/1281192.1281280`
21. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of anthropological research 33, 452–473 (1977)