Comparative Analysis for *k*-Means Algorithms in Network Community Detection

Jian Liu

LMAM and School of Mathematical Sciences, Peking University,
Beijing 100871, P.R. China
dugujian@pku.edu.cn

Abstract. Detecting the community structure exhibited by real networks is a crucial step toward an understanding of complex systems beyond the local organization of their constituents. Many algorithms proposed so far, especially the group of methods in the k-means formulation, can lead to a high degree of efficiency and accuracy. Here we test three k-means methods, based on optimal prediction, diffusion distance and dissimilarity index, respectively, on two artificial networks, including the widely known ad hoc network with same community size and a recently introduced LFR benchmark graphs with heterogeneous distributions of degree and community size. All of them display an excellent performance, with the additional advantage of low computational complexity, which enables the analysis of large systems. Moreover, successful applications to several real world networks confirm the capability of the methods.

Keywords: Community structure, *k*-means, Optimal prediction, Diffusion distance, Dissimilarity index.

1 Introduction

In recent years we have seen an explosive growth of interest and activity on the structure and dynamics of complex networks [1, 2]. This is partly due to the influx of new ideas, particularly ideas from statistical mechanics, to the subject, and partly due to the emergence of interesting and challenging new examples of complex networks such as the internet and wireless communication networks. Network models have also become popular tools in social science, economics, the design of transportation and communication systems, banking systems, powergrid, etc, due to our increased capability of analyzing these models. Since these networks are typically very complex, it is of great interest to see whether they can be reduced to much simpler systems. In particular, much effort has gone into partitioning the network into a small number of clusters [3–18], which are constructed from different viewing angles comparing different proposals in the literature. On a related but different front, recent advances in computer vision and data mining have also relied heavily on the idea of viewing a data set or an image as a graph or a network, in order to extract information about the important features of the images or more generally, the data sets [19, 20].

Z. Cai et al. (Eds.): ISICA 2010, LNCS 6382, pp. 158–169, 2010.

In a previous paper [13], an approach to partition the networks based on optimal prediction theory proposed by Chorin and coworkers [21, 22] is derived. The basic idea is to associate the network with the random walker Markovian dynamics [23], then introduce a metric on the space of Markov chains (stochastic matrices), and optimally reduce the chain under this metric. The final minimization problem is solved by an analogy to the traditional k-means algorithm in clustering analysis [24]. Another work [7] is also along the lines of random walker Markovian dynamics, then introduce the diffusion distance on the space of nodes and identify the geometric centroid in the same framework. This proximity reflects the connectivity of nodes in a diffusion process. Under the same framework [6], a dissimilarity index for each pair of nodes is proposed, which one can measure the extent of proximity between nodes of a network and signify to what extent two nodes would like to be in the same community. They can motivate us to solve the partitioning problem also by k-means algorithms [24] under these two measures [16].

We will compare the above three algorithms in k-means formulation based on optimal prediction, diffusion distance and dissimilarity distance, respectively. From the numerical performance to the artificial networks: the ad hoc network with 128 nodes and the LFR benchmark with 1000 nodes, we can see that the three k-means methods identify the community structure during with a high degree of accuracy, while they also produce little different. Moreover, applications to four real word social networks, including the karate club network, the dolphins network, the political books network and the SFI collaboration network, confirm the differences among them.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the three type of k-means algorithms and corresponding framework. In Section 3, we apply the algorithms to six representative examples mentioned before. Finally we make the conclusion in Section 4.

2 The Framework of k-Means Algorithms for Network Partition

2.1 The k-Means Based on Optimal Prediction

In [13], a new strategy for reducing the random walker Markovian dynamics based on optimal prediction theory [21, 22] is proposed. Let $G(S, E)$ be a network with n nodes and m edges, where S is the nodes set, $E = \{e(x, y)\}_{x,y \in S}$ is the weight matrix and $e(x, y)$ is the weight for the edge connecting the nodes x and y. We can relate this network to a discrete-time Markov chain with stochastic matrix p with entries $p(x, y)$ given by

$$p(x,y) = \frac{e(x,y)}{d(x)}, \qquad d(x) = \sum_{z \in S} e(x,z), \tag{1}$$

where $d(x)$ is the degree of the node x [7, 23, 25]. This Markov chain has stationary distribution

$$\mu(x) = \frac{d(x)}{\sum_{z\in S} d(z)} \tag{2}$$

and it satisfies the detailed balance condition [13].

The basic idea in [13] is to introduce a metric for the stochastic matrix $p(x,y)$

$$\|p\|_\mu^2 = \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} |p(x,y)|^2 \tag{3}$$

and find the reduced Markov chain $\tilde{p}$ by minimizing the distance $\|\tilde{p}-p\|_\mu$. For a given partition of S as $S = \cup_{k=1}^N S_k$ with $S_k \cap S_l = \emptyset$ if $k \neq l$, let $\hat{p}_{kl}$ be the coarse grained transition probability from S_k to S_l on the state space $\mathbb{S} = \{S_1, \ldots, S_N\}$. This matrix can be naturally lifted to the space of stochastic matrices on the original state space S via

$$\tilde{p}(x,y) = \sum_{k,l=1}^N \mathbf{1}_{S_k}(x)\hat{p}_{kl}\mu_l(y), \tag{4}$$

where $\mathbf{1}_{S_k}(x) = 1$ if $x \in S_k$ and $\mathbf{1}_{S_k}(x) = 0$ otherwise, and

$$\mu_k(x) = \frac{\mu(x)\mathbf{1}_{S_k}(x)}{\hat{\mu}_k}, \qquad \hat{\mu}_k = \sum_{z\in S_k} \mu(z). \tag{5}$$

Based upon this formulation, we can find the optimal $\hat{p}_{kl}$ for any fixed partition. With this optimal form $\hat{p}_{kl}$, we further search for the best partition $\{S_1, \cdots, S_N\}$ with the given number of communities N by minimizing the optimal prediction error

$$\begin{aligned}\min_{\{S_1,\cdots,S_N\},\hat{p}_{kl}} J = \big\|\tilde{p}-p\big\|_\mu^2 &= \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} \Big[\tilde{p}(x,y) - p(x,y)\Big]^2 \\ &= \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} p^2(x,y) - \sum_{k,l=1}^N \frac{\hat{\mu}_k}{\hat{\mu}_l}\hat{p}_{kl}^2. \end{aligned} \tag{6}$$

A direct calculation shows that the minimizer of (6) is one in which $\hat{p}$ is given by

$$\hat{p}_{kl} = \sum_{x\in S_k, y\in S_l} \mu_k(x)p(x,y). \tag{7}$$

It can be checked that (7) is a stochastic matrix and $\hat{\mu}$ in (5) is an equilibrium distribution for the Markov chain on $\mathbb{S}$ with transition matrix (7). Furthermore, it is easy to see that (7) satisfies a detailed balance condition with respect to $\hat{\mu}$. A variant of k-means algorithm can be used to handle (6). Given an initial partition $\{S_k^{(0)}\}_{k=1}^N$, for $n \geq 0$, use

$$S_k^{(n+1)} = \Big\{x : k = \arg\min_l D(x, S_l^{(n)})\Big\} \tag{8}$$

to update the new state, where

$$D(x, S_k) = \sum_{l=1}^{N} \sum_{y \in S_l} \mu(x)\mu(y) \left(\frac{p(x,y)}{\mu(y)} - \frac{\hat{p}_{kl}}{\hat{\mu}_l} \right)^2 . \tag{9}$$

This is the theoretical basis for constructing the k-means algorithm for the community structure of complex networks in [13], which is considered to address this optimization issue which guarantees convergence towards a local minimum.

2.2 The k-means Based on Diffusion Distance

The main idea of [7, 16] is to define a system of coordinates with an explicit metric that reflects the connectivity of nodes in a given network and the construction is also based on a Markov random walk on networks. The diffusion distance $D(x, y)$ between x and y is defined as the weighted L^2 distance

$$D^2(x,y) = \sum_{z \in S} \frac{(p(x,z) - p(y,z))^2}{\mu(z)}, \tag{10}$$

where the weight $\mu(z)^{-1}$ penalize discrepancies on domains of low density more than those of high density. This notion of proximity of nodes reflects the intrinsic geometry of the set in terms of connectivity of the nodes in a diffusion process. The transition matrix p has a set of left and right eigenvectors and a set of eigenvalues $1 = \lambda_0 \geq |\lambda_1| \geq \cdots \geq |\lambda_{n-1}| \geq 0$

$$p\varphi_i = \lambda_i \varphi_i, \quad \psi_i^T p = \lambda_i \psi_i^T, \quad i = 0, 1, \cdots, n-1. \tag{11}$$

Note that $\psi_0 = \mu$ and $\varphi_0 \equiv 1$. We also have $\psi_i(x) = \varphi_i(x)\mu(x)$. Let q be the largest index i such that $|\lambda_i| > \delta|\lambda_1|$ and if we introduce the diffusion map

$$\Psi : x \longmapsto \begin{pmatrix} \lambda_1 \varphi_1(x) \\ \vdots \\ \lambda_q \varphi_q(x) \end{pmatrix}, \tag{12}$$

then the diffusion distance $D(x, y)$ can be approximated to relative precision δ using the first q non-trivial eigenvectors and eigenvalues

$$D^2(x,y) \simeq \sum_{i=1}^{q} \lambda_i^2 \Big(\varphi_i(x) - \varphi_i(y) \Big)^2 = \|\Psi(x) - \Psi(y)\|^2. \tag{13}$$

The geometric centroid $c(S_k)$ of community S_k is defined as

$$c(S_k) = \sum_{x \in S_k} \frac{\mu(x)}{\hat{\mu}(S_k)} \Psi(x), \quad k = 1, \cdots, N, \tag{14}$$

where $\hat{\mu}(S_k) = \sum_{x \in S_k} \mu(x)$ [7]. Here $c(S_k)$ may not belong to the set $\{\Psi(x)\}_{x \in S}$. In order to obtain representative centers of the communities that belong to the node set S, we introduce the diffusion center $m^D(S_k)$ by

$$m^D(S_k) = \arg \min_{x \in S_k} \|\Psi(x) - c(S_k)\|^2, \quad k = 1, \cdots, N. \tag{15}$$

2.3 The *k*-Means Based on Dissimilarity Index

In [6, 16], a dissimilarity index between pairs of nodes is defined, which one can measure the extent of proximity between nodes of a network. Suppose the random walker is located at node x. The mean first passage time $t(x, y)$ is the average number of steps it takes before it reaches node y for the first time, which is given by

$$t(x,y) = p(x,y) + \sum_{j=1}^{+\infty}(j+1) \cdot \sum_{z_1,\cdots,z_j \neq y} p(x,z_1)p(z_1,z_2)\cdots p(z_j,y). \tag{16}$$

It has been shown that $t(x, y)$ is the solution of the linear equation

$$[I - B(y)] \begin{pmatrix} t(1,y) \\ \vdots \\ t(n,y) \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{17}$$

where $B(y)$ is the matrix formed by replacing the y-th column of matrix P with a column of zeros [6]. The difference in the perspectives of nodes x and y about the network can be quantitatively measured. The dissimilarity index is defined by the following expression

$$\Lambda(x,y) = \frac{1}{n-2}\left(\sum_{z \in S, z \neq x,y} \Big(t(x,z) - t(y,z)\Big)^2 \right)^{\frac{1}{2}}. \tag{18}$$

If two nodes x and y belong to the same community, then the average distance $t(x, z)$ will be quite similar to $t(y, z)$, therefore the network's two perspectives will be quite similar. Consequently, $\Lambda(x, y)$ will be small if x and y belong to the same community and large if they belong to different communities. The center $m(S_k)$ of community S_k can be defined as

$$m(S_k) = \arg\min_{x \in S_k} \frac{1}{|S_k|} \sum_{y \in S_k, y \neq x} \Lambda(x,y), \quad k = 1, \cdots, N, \tag{19}$$

where $|S_k|$ is the number of nodes in community S_k. This is an intuitive and reasonable idea for us to choose the node reached others in the same community with the minimum average dissimilarity index as the center of S_k.

3 Experimental Results

3.1 Ad Hoc Network with 128 Nodes

We apply our methods to the ad hoc network with 128 nodes. The ad hoc network is a typical benchmark problem considered in many papers [4, 6, 13–18]. Suppose we choose $n = 128$ nodes, split into 4 communities containing 32 nodes each. Assume pairs of nodes belonging to the same communities are linked with probability

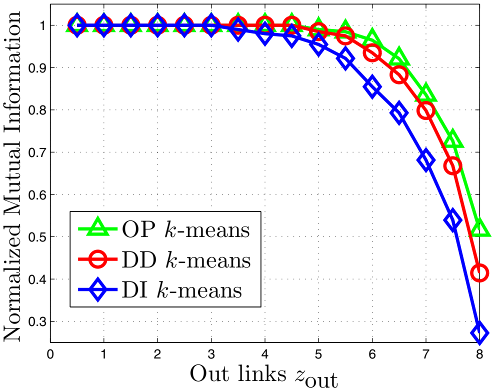


Fig. 1. Test of the three k-means algorithms on the ad hoc network with respect to the normalized mutual information defined in (21). Each point corresponds to an average over 20 graph realizations.

p_{in}, and pairs belonging to different communities with probability p_{out}. These values are chosen so that the average node degree, d, is fixed at $d = 16$. In other words p_{in} and p_{out} are related as

$$31p_{\text{in}} + 96p_{\text{out}} = 16. \tag{20}$$

Here we naturally choose the nodes group $S_1 = \{1 : 32\}, S_2 = \{33 : 64\}, S_3 = \{65 : 96\}, S_4 = \{97 : 128\}$.

Testing an algorithm on any graph with built-in community structure also implies defining a quantitative criterion to estimate the goodness of the answer given by the algorithm as compared to the real answer that is expected. For reviews of similarity measures see [4, 9–18]. In the first tests of community detection algorithms, one used a measure called fraction of correctly identified nodes [4, 13–18], but it is not well defined in some cases when a detected community is a merger of two or more real communities, so a more precise measure, which is called the normalized mutual information, is more appropriate [9–12]. It is based on defining a confusion matrix M, where the rows correspond to the real communities, and the columns correspond to the found communities. The member of M, M_{kl} is simply the number of nodes in the real community k that appear in the found community l. The number of real communities is denoted N_r and the number of found communities is denoted N_f, the sum over row k of matrix M_{kl} is denoted M_k and the sum over column l is denoted M_l . A measure of similarity between the partitions, based on information theory, is then

$$NMI(\mathbb{S}_r, \mathbb{S}_f) = \frac{-2\sum_{k=1}^{N_r}\sum_{l=1}^{N_f} M_{kl}\log(\frac{nM_{kl}}{M_k M_l})}{\sum_{k=1}^{N_r} M_k \log(\frac{M_k}{n}) + \sum_{l=1}^{N_f} M_l \log(\frac{M_l}{n})}. \tag{21}$$

We change z_{out} from 0.5 to 8 and look into the corresponding normalized mutual information produced by the three methods, which is shown in Figure 1. It seems that OP k-means performs better than the two others, especially for the more diffusive cases when z_{out} is large.

3.2 The LFR Benchmark

The LFR benchmark [10–12] is a special case of the planted partition model, in which groups are of different sizes and nodes have different degrees. The node degrees are distributed according to a power law with exponent γ; the community sizes also obey a power law distribution, with exponent β. In the construction of the benchmark graphs, each node receives its degree once and for all and keeps it fixed until the end. In this way, the two parameters p_{in} and p_{out} of the planted partition model in this case are not independent. Once the value of pin is set one obtains the value of pout and vice versa. It is more practical to choose as independent parameter the mixing parameter μ, which expresses the ratio between the external degree of a node with respect to its community and the total degree of the node. Of course, in general one may take different values for the mixing parameter for different nodes, but we will assume, for simplicity, that μ is the same for all nodes, consistently with the standard hypotheses of the planted partition model. A realization of the LFR benchmark, with 500 nodes and parameters $\mu = 0.1, \gamma = 2, \beta = 1, \langle k \rangle = 20$, corresponding to 11 communities represent by different colors is shown in Figure 2(a).

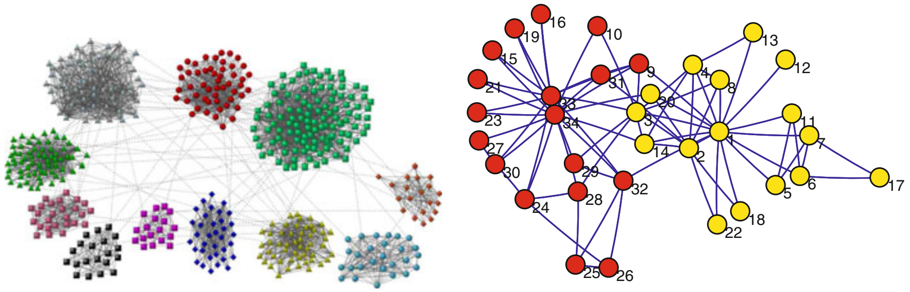
In Figure 3, we show what happens if one operates the three k-means methods on the benchmark, for $n = 1000$ and the average degree $\langle k \rangle = 20$. The four panels correspond to four pairs for the exponents (γ, β) = (2,1), (2,2), (3,1), (3,2). We have chosen combinations of the extremes of the exponents' ranges in order to explore the widest spectrum of network structures. Each curve shows the variation of the normalized mutual information with the mixing parameter μ. In general, we can infer that the k-means type methods give good results.

3.3 The Karate Club Network

This network was constructed by Wayne Zachary after he observed social interactions between members of a karate club at an American university [26]. Soon after, a dispute arose between the clubs administrator and main teacher and the club split into two smaller clubs. It has been used in several papers to test the algorithms for finding community structure in networks [3–6, 13–18]. The partitioning results are shown in Figure 2(b). The three kinds of k-means algorithms produce the same results, which seem consistent with the original structure of the network.

3.4 The Dolphins Network

The dolphins network is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand [27]. The network was compiled from the studies of the dolphins, with ties between dolphin pairs being established by observation of statistically significant frequent



(a) (b)

Fig. 2. (a)A realization of the LFR benchmark, with 500 nodes, corresponding to 11 communities represent by different colors. Here $\mu = 0.1, \gamma = 2, \beta = 1, \langle k \rangle = 20$. (b)The community structure for the karate club network. The three kinds of k-means algorithms produce the same results.

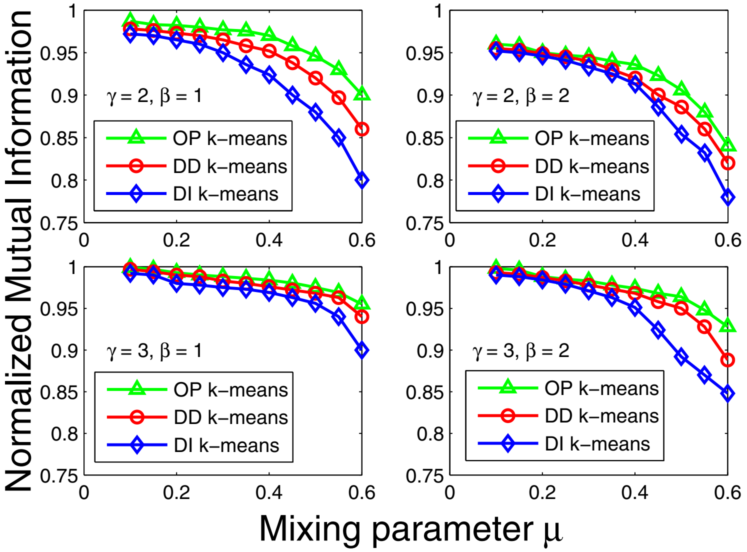
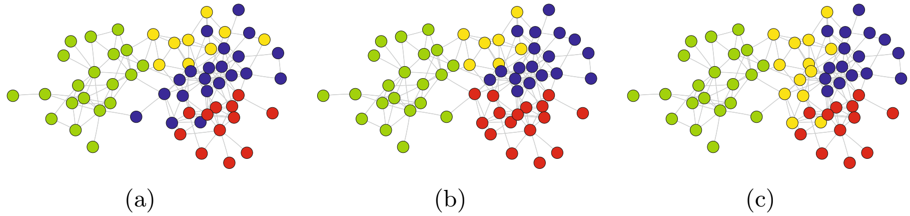


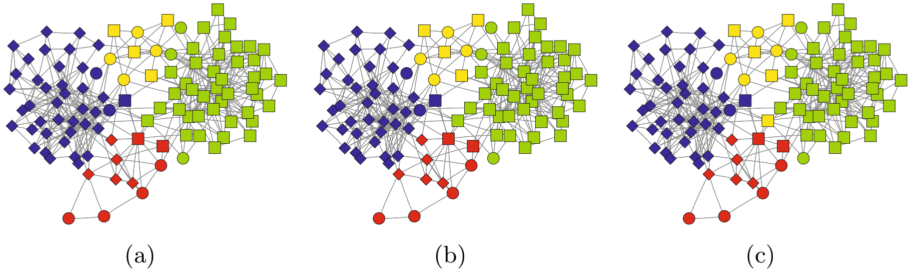
Fig. 3. Test of the three k-means methods on the LFR benchmark. The number of nodes $n = 1000$ and the average degree $\langle k \rangle = 20$. The results clearly depend on all parameters of the benchmark, from the exponents γ and β to the mixing parameter μ. Each point corresponds to an average over 20 graph realizations.

association [4]. The results obtained by our methods are shown in Figure 4. According to the results, the network seems splitting into two large communities by the green part and the larger one and the larger one keeps splitting into a few smaller communities, represent by different colors. The split into two groups appears to correspond to a known division of the dolphin community [28]. The



(a) (b) (c)

Fig. 4. The community structure for the dolphins network, corresponding 4 clusters represent by different colors. (a)OP k-means; (b)DD k-means; (c)DI k-means. The three kinds of k-means algorithms produce a little different.



(a) (b) (c)

Fig. 5. The community structure for the dolphins network, corresponding 4 clusters represent by different colors. (a)OP k-means; (b)DD k-means; (c)DI k-means. The three kinds of k-means algorithms produce nearly the same.

subgroupings within the larger half of the network also seem to correspond to real divisions among the animals that the largest subpart consists almost of entirely of females and the others almost entirely of males.

3.5 The Political Books Network

We consider the network of books on politics, which are assigned based on a reading of the descriptions and reviews of the books posted on Amazon [5]. In this network the nodes represent 105 recent books on American politics bought from the on-line bookseller Amazon.com, and the edges join pairs of books that are frequently purchased by the same buyer, as indicated by the feature that customers who bought this book also bought these other books. As shown in Figure 5, nodes have been given whether they are conservative(box) or liberal(diamond), except for a small number of books which are neutral(ellipse). The results are shown in Figure 5. We find four communities denoted by different colors. It seems that one of these communities consists almost entirely of liberal books and one almost entirely of conservative books. Most of the neutral books fall in the two remaining communities. Thus these books appear to form communities of copurchasing that align closely with political views.

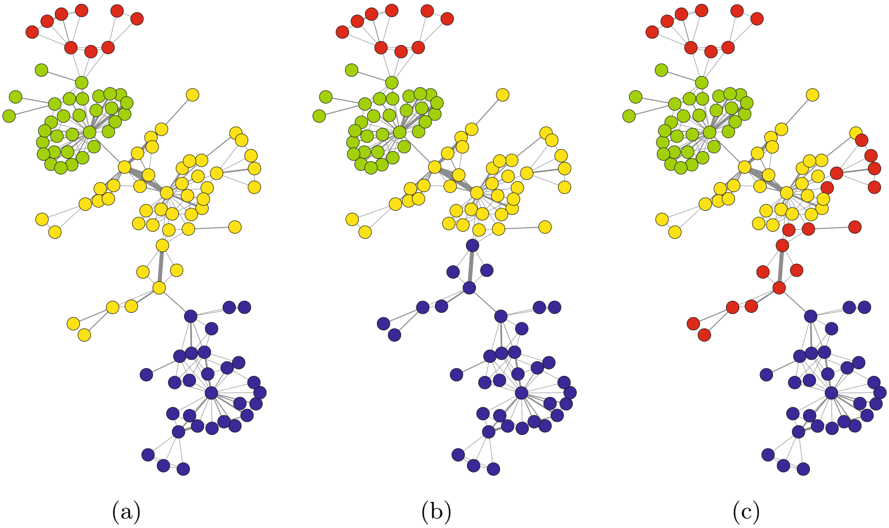


Fig. 6. The community structure for the dolphins network, corresponding 4 clusters represent by different colors. (a)OP k-means; (b)DD k-means; (c)DI k-means. The three kinds of k-means algorithms produce a little different, while the result obtained by OP k-means seems most reasonable.

3.6 The SFI Collaboration Network

We have applied the three k-means method to a collaboration network of scientists at the Santa Fe Institute, an interdisciplinary research center in Santa Fe, New Mexico [3, 6]. The 271 vertices in this network represent scientists in residence at the Santa Fe Institute during any part of calendar year 1999 or 2000, and their collaborators. A weighted edge is drawn between a pair of scientists if they coauthored one or more articles during the same time period. In Figure 6, we illustrate the results from the application of our algorithm to the largest component of the collaboration graph (which consists of 118 scientists). We find that the algorithms split the network into a few strong communities, with the divisions running principally along disciplinary lines. The community at the top of the figure (red) is the least well defined, and represents a group of scientists using agent-based models to study problems in economics and traffic flow. The next community (green) represents a group of scientists working on mathematical models in ecology, and forms a fairly cohesive structure The largest community (yellow) is a group working primarily in statistical physics, and seems sub-divided into several smaller groups. In this case, each sub-community seems to revolve around the research interests of one dominant member. The final community at the bottom of the figure (blue) is a group working primarily on the structure of RNA. It too can be divided further into smaller subcommunities, centered once again around the interests of leading members.

4 Conclusions

In this paper, we test three k-means methods, based on optimal prediction, diffusion distance and dissimilarity index, respectively, on two artificial networks, including the widely known ad hoc network with same community size and a recently introduced LFR benchmark graphs with heterogeneous distributions of degree and community size. All of them have an excellent performance, with the additional advantage of low computational complexity, which enables one to analyze large systems. They identify the community structure during iterations with a high degree of accuracy, with producing little different. Moreover, successful application to several real world networks confirm the capability among them and the differences and limits of them are revealed obviously.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant 10871010 and the National Basic Research Program of China under Grant 2005CB321704. The author thanks Professor S. Fortunato for providing the softwares on generating the LFR benchmarks, Professor M.E.J. Newman for providing the data of the karate club network, the dolphins network and the political books network, Professor Haijun Zhou and Professor Junhua Zhang for providing the data of SFI collaboration network.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74(1), 47–97 (2002)
2. Newman, M., Barabási, A.L., Watts, D.J.: The structure and dynamics of networks. Princeton University Press, Princeton (2005)
3. Girvan, M., Newman, M.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99(12), 7821–7826 (2002)
4. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69(2), 026113 (2004)
5. Newman, M.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA 103(23), 8577–8582 (2006)
6. Zhou, H.: Distance, dissimilarity index, and network community structure. Phys. Rev. E 67(6), 061901 (2003)
7. Lafon, S., Lee, A.: Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization. IEEE Trans. Pattern. Anal. Mach. Intel. 28, 1393–1403 (2006)
8. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Phys. Rev. E 72, 027104 (2005)
9. Danon, L., Diaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. J. Stat. Mech. 9, P09008 (2005)
10. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E 78(4), 046110 (2008)
11. Weinan, E., Li, T., Vanden-Eijnden, E.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E 80(1), 016118 (2009)

12. Li, T., Liu, J., Weinan, E.: Community detection algorithms: a comparative analysis. Phys. Rev. E 80(5), 056117 (2009)
13. Weinan, E., Li, T., Vanden-Eijnden, E.: Optimal partition and effective dynamics of complex networks. Proc. Natl. Acad. Sci. USA 105(23), 7907–7912 (2008)
14. Li, T., Liu, J., Weinan, E.: Probabilistic Framework for Network Partition. Phys. Rev. E 80, 026106 (2009)
15. Liu, J.: Detecting the fuzzy clusters of complex networks. Patten Recognition 43, 1334–1345 (2010)
16. Liu, J., Liu, T.: Detecting community structure in complex networks using simulated annealing with k-means algorithms. Physica A 389, 2300–2309 (2010)
17. Liu, J.: An extended validity index for identifying community structure in networks. In: Zhang, L., Lu, B.-L., Kwok, J. (eds.) ISNN 2010. LNCS, vol. 6064, pp. 258–267. Springer, Heidelberg (2010)
18. Liu, J.: Finding and evaluating fuzzy clusters in networks. In: Tan, Y., Shi, Y., Tan, K.C. (eds.) Advances in Swarm Intelligence. LNCS, vol. 6146, pp. 17–26. Springer, Heidelberg (2010)
19. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intel. 22(8), 888–905 (2000)
20. Meilă, M., Shi, J.: A random walks view of spectral segmentation. In: Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, pp. 92–97 (2001)
21. Chorin, A.J., Kast, A.P., Kupferman, R.: Unresolved computation and optimal predictions. Comm. Pure Appl. Math. 52(10), 1231–1254 (1999)
22. Chorin, A.J.: Conditional expectations and renormalization. Multi. Model. Simul. 1, 105–118 (2003)
23. Lovasz, L.: Random walks on graphs: A survey. Combinatorics, Paul Erdos is Eighty 2, 1–46 (1993)
24. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2001)
25. Chung, F.: Spectral Graph Theory. American Mathematical Society, Rhode Island (1997)
26. Zachary, W.: An information flow model for conflict and fission in small groups. J. Anthrop. Res. 33(4), 452–473 (1977)
27. Lusseau, D.: The emergent properties of a dolphin social network. Proceedings of the Royal Society B: Biological Sciences 270, 186–188 (2003)
28. 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)