Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 21 (2023) 376-383

Original

A nature-inspired algorithm to find community structure in complex networks

Bilal Saoud

Electrical Engineering Department, Sciences and Applied Sciences Faculty,
Bouira University, Bouira, Algeria

Received 09 23 2021; accepted 05 25 2022
Available 06 30 2023

Abstract: Complex networks are in general communities. These communities are especially important. Network communities represent sets of nodes, which are very connected. In this research, we developed a new method to find the community structure in networks. Our method is based on flower pollination algorithm (FPA) which is used in the split-ting process. The splitting of networks in our method maximizes a function of quality called modularity. We provide a general framework for implementing our new method to find community structure in networks. We present the effectiveness of our method by comparison with some known methods on computer-generated and real-world networks.

Keywords: Community detection, networks, flower pollination algorithm, normalized mutual information, modularity

[*] Corresponding author.
E-mail address: bilal340@gmail.com (Bilal Saoud).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

Many systems can be represented by network or a graph, which makes them very powerful structure. A network G is defined by two sets (Newman, 2010). The first set is node set V (node set) and the second is edge set E. Nodes share relationships between them. Relationships are represented by edges. In general, the number of nodes is $|V| = n$ and edges is $|E| = m$. Euler's solution of the seven bridges of Königsberg problem is the first use of networks to represent systems (Hopkins & Wilson, 2004). Today networks are used to illustrate several systems. For instance, in social network, which is an interaction between entities (persons, groups of persons, organizations, web sites, …), can be represented by a network with two sets V and E. Nodes stand for entities (for example persons) and edges stand for relationships between entities (for examples between persons). Analyzing and understanding a network leads to better understanding the system. Among features that can help to understand the structure of a network, we can find the community structure.

Community structure exists in networks, and it gives more information about the network. For instance, we can understand very well the system, which is represented by a network, by finding its community structure and the relationship between communities. In addition, networks can represent many systems like social networks, electric networks, biological networks, etc. It is vital to develop new methods to find network communities. When we analyze networks by studying relationships between nodes, we can get extra information about networks and systems. In general, nodes in the same community have common properties or insure similar tasks in network. A network has parts that are more densely connected than other parts. In other words, the nodes in these parts share many edges between them. These parts of nodes and edges are called communities (clusters). Finally, many studies have been done around networks and how to find community structure.

Many community structure detection methods have been developed (Fortunato, 2010). According to the type of network, we can find methods for unipartite/bipartite networks, weighted/unweighted networks, and directed/undirected networks. Furthermore, methods can be classified into different classes such as hierarchical methods (merging or splitting), methods that are based on maximization of an objective function. Some methods find disjoint communities, where intersection between communities is empty. However, other methods were designed to find overlapping communities, for instance the method in (Chen et al., 2019), where the intersection between communities is not empty.

In this paper, we address the problem of finding community structure in networks. We present a new method to discover community structure in unweighted and undirected networks. Our method is based on nature-inspired metaheuristics algorithm. We have developed our method based on the pollination process of flowers (Yang, 2012). Our method is a hierarchical one. It is based on the splitting of a given network $G(V, E)$, which models a system. Splitting step in our method is done by the flower pollination algorithm (FPA) (Yang, 2012) to optimize the function of quality called modularity Q. The process of splitting will be stopped when the graph G has been disconnected, which means that each node of G represents a community. Finally, our method builds a dendrogram and finds the most optimal community structure $\pi = \{c_1, \cdots, c_k\}$, such as $\bigcup_{i=1}^{k} c_i = V$ and $c_i \neq \emptyset, c_i \cap c_j = \emptyset$ (for $i, j = 1 : k$).

The paper is organized as follows. The concept of FPA is presented in Section 2. Our approach is detailed in Section 3. Experimental results and discussions are given in Section 4. Finally, Section 5 concludes the paper.

2. FPA Presentation

Flower pollination is an interesting phenomenon in nature. Based on the studying flower pollination process, a new algorithm of optimization was designed by Yang (2012). The algorithm has been named flower pollination algorithm (FPA). In nature pollination can be abiotic form or biotic form. In general, 90% of flowers have biotic pollination where the pollen is transferred by animals (pollinator) like insects. Biotic pollination by bees for instance can be done over long distances.

FPA has three steps (Yang, 2012) described in the following:

- In the first step, the algorithm initializes its parameters and generates the initial population. The best solution is found also in the first step.
- The second step, flowers in population start doing pollination in d-dimensional search (solution space). Flowers can choose a local or global pollination at every iteration in the search space. The algorithm switch between local pollination and global pollination based on probability $p \in [0,1]$. Flowers' location represents the vector of solutions vector and the value of objective function for every solution estimated. According to the value of objective function the new solution is evaluated and updated at every iteration and the best solution will be improved.
- In the final step, the algorithm stops after some iterations and the best solution will be selected.

FPA can converge very fast and can escape the problem of local minima because it makes the long distances movement based on levy flight (Emary et al., 2019). FPA can be used to solve different optimization problems in various fields such as computer science (cloud computing, data clustering, wireless sensor networks, etc.), bioinformatics, operation research, image processing and engineering (Zhou et al., 2018; Abdel-Basset & Shawky, 2019).

3. A new method to find communities in networks

In this section, we present our hierarchical method to discover community structure in networks. Hierarchical methods can be divisive or agglomerative. Our method is hierarchical divisive method. Network is divided by our method based on the maximization of the function of quality called modularity (Clauset et al., 2004). Our method is designed to find community structure in networks with only a single type of node and undirected, unweighted edge.

We can measure the strength of a community structure by the function of quality called modularity (Clauset et al., 2004). Modularity function Q is based on the observed edges fraction $e(c_i)$ within communities and the expected edges fraction $a(c_i)$ within the same communities, $Q = \sum_{c_i} e(c_i) - a(c_i)^2$. Modularity can be estimated for undirected and unweighted graph $G(V,E)$ as:

$$Q = \frac{1}{2m}\sum_i \sum_j (A[i,j] - P[i,j])\delta(c_i, c_j) \qquad (1)$$

where n is the number of nodes in G $(n = |V|)$, m is the number of edges in G $(m = |E|)$ and the community structure is $\pi = \{c_1, \cdots, c_k\}$. $A_{n,n}$ represents the adjacency matrix of $G(V,E)$. For any node $i \in V$, d_i is the degree of node i and c_i its community. The matrix A takes two values 1 or 0 if there is an edge between node i and j then $A[i,j] = 1$ or $A[i,j] = 0$ if there is not a connection between i and j. $P_{n,n}$ represents the adjacency matrix that corresponds null model. In the null model the probability of an existing edge between nodes i and j is $P_{i,j} = \frac{d_i \times d_j}{2m}$. Finally, δ function is given as follows:

$$\delta(c_i, c_j) = \begin{cases} 1 \ if \ c_i = c_j \\ 0 \ otherwise \end{cases} \qquad (2)$$

Values of Q are between 0 and 1. Q closer to 1 indicates stronger community structures. According to (Clauset et al., 2004), a value above about 0.3 is a good indicator of significant community structure in a network.

Let $G(V,E)$ be an undirected and unweighted network, where $V = (v_1, \cdots, v_n)$ is the set of nodes, $E = (e_1, \cdots, e_n)$ is the set of edges. The goal of our community detection method is to partition the network G into k communities (groups): $\pi = (c_1, \cdots, c_n)$, where $c_i \neq \emptyset$, $c_i \cap c_j = \emptyset$ (for $i, j = 1 : k$) and $V = \bigcup_{i=1}^{k} c_i$. In addition, our method finds the community structure π of the network G with the greatest value of modularity Q. To reach this goal, we used an FPA. Our method splits $G(V,E)$ into two new networks G_1 and G_2. Nodes of each new network represent a community. Nodes of G_1 represent a community c_1 and nodes of G_2 represent a community c_2. The splitting is based on FPA to maximize the value of modularity function Q. Then, G_1 and G_2 will be split until the network G has been disconnected. At the end of our method each node in $G(V,E)$ represents a community. Finally, we get a dendrogram for our method and the community structure will be chosen based on value of modularity Q or the number of communities.

The general algorithm of our method to find community structure is as follows:

Algorithm 1: Our method algorithm

```
        Data: $G(V,E)$
        Result: dendrogram
1       $\pi = FPA()$, find a partition $\pi$ based on FPA;
2       Divide $G$ based on $\pi$, $G = G_1 + G_2$;
3       Update the matrix of merge $M$ for a final dendrogram;
4       Go to Steps 1 for each graph $G_1$ and $G_2$;
5       Return the final dendrogram;
```

Figure 1 shows the result of splitting Zachary's karate club network (Zachary, 1977) by our method. Numbers from 1 to 34 stand for nodes. Our method gave a community structure with two communities $\pi = \{c_1, c_2\}$ such as:
$c_1 = \{1, 2, 18, 20, 22, 3, 4, 14, 13, 12, 8, 5, 6, 7, 11, 17\}$,
$c_2 = \{8, 9, 15, 16, 19, 21, 23, 31, 33, 34, 27, 30, 24, 25, 26, 28, 29, 32\}$, which were separated by vertical lines in Figure 1. The community structure that was found by our method on the same network is also represented in Figure 2. In this figure, communities' nodes have different colors and shapes.

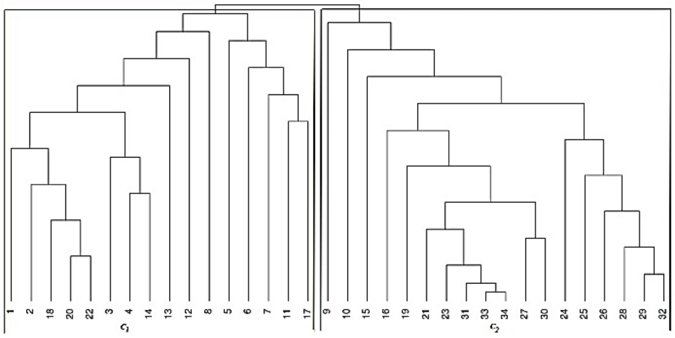

Figure 1. The dendrogram of Zachary's karate club network was created by our method.

4. Experiments and results

To evaluate our method to find community structures in networks, we have tested it on computer-generated and several real networks (Zackary's karate club (Zachary, 1977), American college football (Girvan & Newman, 2002), dolphins (Lusseau et al., 2003), books about US politics (Krebs, 2021), jazz musicians (Gleiser & Danon, 2003), word adjacencies (Newman, 2006) and

Les Misérables (Knuth, 1993). We have compared our method with some well-known methods: fast greedy method (Clauset et al., 2004), label propagation method (Raghavan et al., 2007), and infomap method (Rosvall & Bergstrom, 2008).

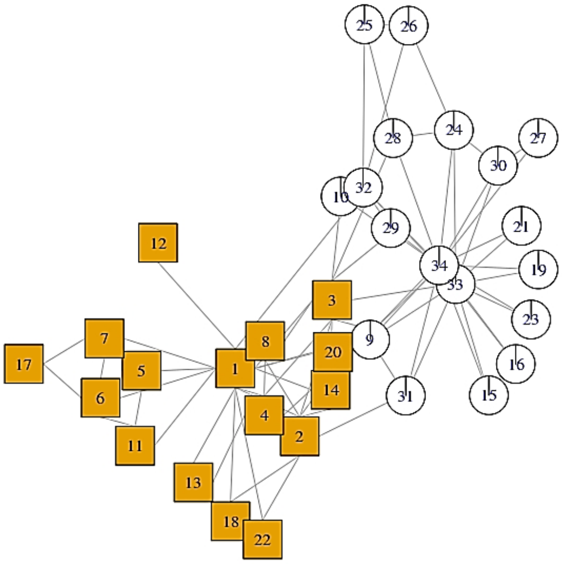

Figure 2. Zachary's karate club network community structure is detected by our method.

The fast greedy method was proposed by Clauset et al. (2004). It is an improvement of the method of Newman (Newman, 2004). It is based on the greedy optimization of modularity, and it is a hierarchical agglomeration algorithm to detect community structure. The method label propagation was proposed by Raghavan et al. (2007). It is based on label propagation. Initially, every node in the graph is initialized with a unique label and at every step of the method each node takes the label that most of its neighbors currently have. The iterative process converges when labels cannot be changed. Then, nodes having identical labels form a community. Finally, the method that was proposed by Rosvall and Bergstrom (2008), which is known as infomap, uses the concept of random walks and entropy communities to find the community structure in network.

4.1. Normalized mutual Information

The comparison of our method with other methods is based on the normalized mutual information (NMI) function (Danon et al., 2005). The NMI is a powerful function to compare a community structure that was found by methods with the real community structure. The value of NMI is based on defining a confusion matrix N, where the rows represent the real communities, and the columns represent the found communities. N_{ij} is the number of nodes in the real community that appears in the found community j. For two partitions A and B, the partition A represents the real partition with c_A communities and B represents the found partition with c_B communities, The normalized mutual information (NMI) is estimated as follows:

$$NMI(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij}\log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{c_A} N_{i.}\log(\frac{N_{i.}}{N})+\sum_{j=1}^{c_B} N_{.j}\log(\frac{N_{.j}}{N})} \tag{3}$$

NMI values are in the range $[0, 1]$. Partitions A and B are identical if $NMI(A, B) = 1$.

4.2. Dataset based on computer-generated networks

Our method is evaluated on computer-generated networks benchmark proposed by Lancichinetti et al. 2008. The benchmark parameters are the number of nodes N, the exponents γ and β of the degree and community size distribution respectively (both distributions are power laws), the number of average degree $\langle k \rangle$, number of communities N_c, and the mixing parameter μ. Each node shares a fraction $(1 - \mu)$ of its links with other nodes of its community and a fraction μ with the other nodes of the network.

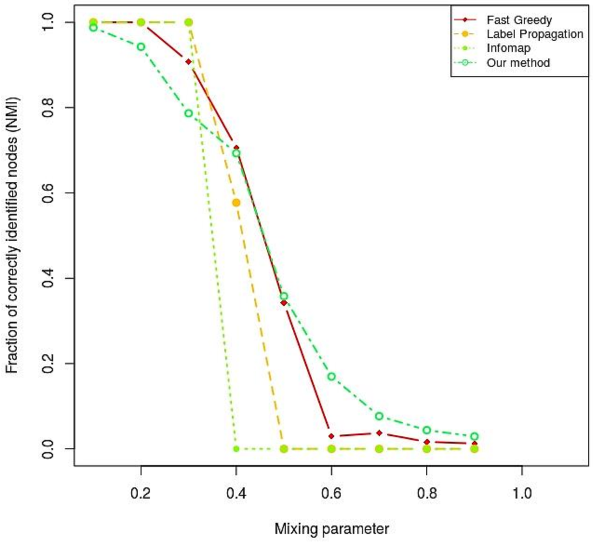

Figure 3. NMI vs. mixing parameter μ.

Figure 3 shows the variation of the NMI obtained by our method, fast greedy method, label propagation method and infomap method on the benchmark networks, with the parameters: mixing parameter μ between 0.1 and 0.9, $\langle k \rangle = 16, \gamma = 3, \beta = 2, N = 128$ and $N_c = 4$. The value of NMI obtained by our method is high when μ changes from 0 to 0.5 and the same thing with other methods. At this range, nodes share many edges with nodes of its community that makes the community structure clear and easy to find. Methods could group the most nodes in the correct communities when the mixing parameter μ is in $[0, 0.5]$. When μ is in $[0.5 - 0.9]$ range, it is difficult for all methods to find the true community structure. At this range nodes share few edges with nodes of its community and many edges with nodes from other communities, which make the community structure unclear and difficult to find. However, our method is still more accurate than the other methods. Our method evaluates the community structure at each step of splitting process and at the end our method selects the best community structure based on modularity value. From Figure 3, we see that our method can discover community structure better than fast greedy, label propagation method and infomap method when μ is greater than 0.5.

Figure 4 illustrates the result of our method on network generated by computer with mixing parameter $\mu = 0.8$. Figure 4 shows the different communities that were found by our method. On this network with a mixing parameter $\mu = 0.8$, our method found a community structure (π) with eight communities ($\pi = \{c_1, c_2, \cdots, c_8\}$). Dendrogram labels stand for nodes. In this example, we have a network with 128 nodes. We mention that the community structure can be found by breaking the dendrogram (Figure 4) at various levels (Abonyi & Feil, 2007). In our case, we have chosen to break the dendrogram at the level which maximizes the modularity function.

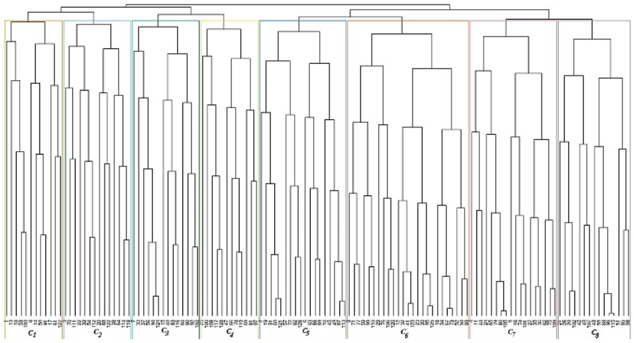

Figure 4. The dendrogram and community structure by our method on computer generated network with mixing parameter $\mu = 0.8$.

4.3. Dataset based on real networks

In this section, we give the simulation results of our method, fast greedy, label propagation and infomap on real networks. We considered some real networks drawn from disparate fields (Zachary 1977), dolphins (Lusseau et al., 2003), football (Girvan & Newman, 2002) and books about US politics (Krebs, 2021), where the community structure is known, which made them suitable to evaluate community detection methods.

1. Zachary's club network (Zachary, 1977) is a real network that corresponds to a social network of friendships between 34 members of a karate club at a university in the United States in the 1970 ($n = 34$ and $m = 78$). The network has two clusters.
2. Dolphins Network (Lusseau et al., 2003) is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand. This network ($n = 62$ and $m = 159$) has two communities.
3. College football network (Girvan & Newman, 2002) represents the schedule of Division I Games for the year 2000 season. This network is made of 115 teams (nodes) and 613 edges. It is divided into 12 groups.
4. Books about US politics network (Krebs, 2021) is 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 purchasing of books by the same buyers. Compiled by Valdis Krebs. Books network has three communities.

Table 1 gives obtained results on networks. In this table, for each network we have estimated the value of modularity function according to equation 1, NMI values (according to equation 3) and we have mentioned the number of communities. As can be seen from Table 1, methods find community structure with different number of communities. According to NMI values, our method can regroup the most nodes in the correct communities on Zachary's karate club, dolphin social network, American college football and books about US politics respectively. The value of modularity by our method on these networks was above 0.3.

Figures 5 and 6 show the community structure that was found by our method on dolphins' network and Books about US politics network. Each label represents a node and edges stand for the relationship between nodes. The community structure that was found by our method was represented by different shapes and colors. Nodes of the same community are represented by the same color and shape. From these Figures 5 and 6, we can see that nodes in the same community are more connected between them and have a few connections with nodes from other communities.

Table 1. Performance results on real networks with known community structure.

Method	Karate			Dolphins			Football			Books		
	N_c	NMI	Q	N_c	NMI	Q	N_c	NMI	Q	N_c	NMI	Q
Fast greedy	3	0.69	0.38	4	0.55	0.49	6	0.70	0.54	4	0.53	0.50
Label propagation	4	0.70	0.41	3	0.76	0.48	11	0.85	0.58	3	0.50	0.47
Infomap	3	0.50	0.40	5	0.53	0.52	12	0.91	0.60	6	0.49	0.52
Our method	2	1	0.37	2	0.81	0.38	10	0.78	0.51	2	0.52	0.43

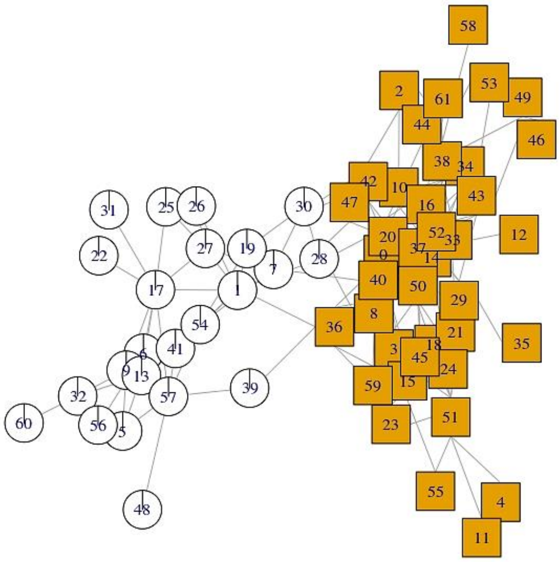

Figure 5. The community structure of dolphins' network detected by our method and represented by different colors and shapes.

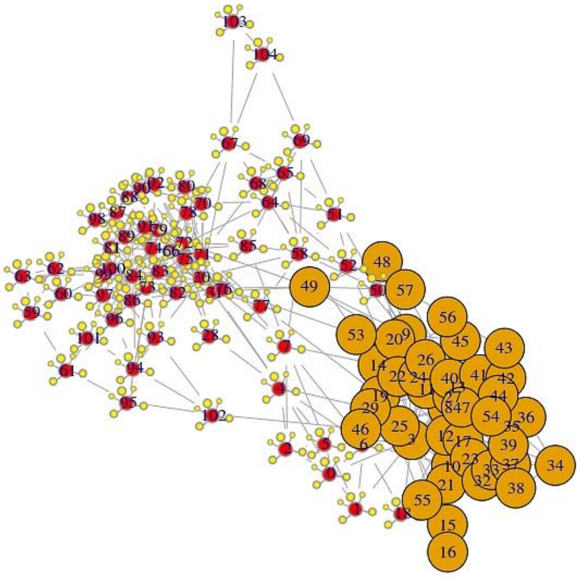

Figure 6. Community structure of books about US politics network detected by our method and represented by different colors and shapes.

We evaluated the performance of our method with other different real networks without a known community structure. A brief description of these networks is given below.

- Jazz network is a collaborative network (Gleiser & Danon, 2003), which represents the association between jazz musicians. Jazz musicians are represented by nodes and edge existing between nodes just if two musicians played together. The network has $n = 198$ nodes and $m = 2742$ edges.

- Word adjacencies network represents the adjacency network of common adjectives and nouns in the novel David Copperfield by Charles Dickens (Newman, 2006). It has $n = 112$ nodes and $m = 425$ edges.

- Les Misérables network is co-appearance network of characters in the novel Les Misérables (Knuth, 1993). The network has $n = 77$ nodes and $m = 254$ edges.

Table 2 gives results of our method, fast greedy, label propagation and Infomap. The number of communities and the estimated value of modularity were mentioned in Table 2. From Table 2, we can see that our method finds community structures with a high value of modularity. It is difficult to compare methods between them because we do not have a reference (a known community structure).

Table 2. Performance results on real networks with unknown community structure.

Method	Jazz		Word adjacencies		Misérables	
	N_c	Q	N_c	Q	N_c	Q
Fast greedy	4	0.438	7	0.294	5	0.500
Label propagation	2	0.281	1	0	4	0.475
Infomap	7	0.280	2	0.009	9	0.546
Our method	3	0.346	6	0.264	7	0.505

5. Conclusion and future work

A new hierarchical method to discover the community structure for unweighted and undirected networks was presented in this paper. Our new method was developed based on maximization of function of modularity by FPA. Results obtained on computer-generated networks and real benchmark networks prove the efficiency of our method in terms of finding community structures with high values of modularity and accuracy.

Our method can be tested on large scale networks. We can develop it to find community structure in weighted or directed network. It can be extended to find overlapping communities.

Conflict of interest

The author has no conflict of interest to declare.

Funding

The author received no specific funding for this work.

References

Abdel-Basset, M., & Shawky, L. A. (2019). Flower pollination algorithm: a comprehensive review. *Artificial Intelligence Review*, *52*, 2533-2557
https://doi.org/10.1007/s10462-018-9624-4

Abonyi, J., & Feil, B. (2007). *Cluster analysis for data mining and system identification*. Springer Science & Business Media.
https://doi.org/10.1007/978-3-7643-7988-9

Chen, J., Liu, M., & Liu, X. (2019). Research on of overlapping community detection algorithm based on tag influence. *Cluster Computing*, *22*, 6669-6679.
https://doi.org/10.1007/s10586-018-2402-x

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.
https://doi.org/10.1103/PhysRevE.70.066111

Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment*, *2005*(09), P09008.
https://doi.org/10.1088/1742-5468/2005/09/P09008

Emary, E., Zawbaa, H. M., & Sharawi, M. (2019). Impact of Lèvy flight on modern meta-heuristic optimizers. *Applied Soft Computing*, *75*, 775-789.
https://doi.org/10.1016/j.asoc.2018.11.033

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3-5), 75-174.
https://doi.org/10.1016/j.physrep.2009.11.002

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821-7826.
https://doi.org/10.1073/pnas.122653799

Gleiser, P. M., & Danon, L. (2003). Community structure in jazz. *Advances in complex systems*, *6*(04), 565-573.
https://doi.org/10.1142/S0219525903001067

Hopkins, B., & Wilson, R. J. (2004). The truth about Königsberg. *The College Mathematics Journal*, *35*(3), 198-207.
https://doi.org/10.1080/07468342.2004.11922073

Knuth, D. E. (1993). The Stanford GraphBase: a platform for combinatorial computing (Vol. 1). New York: AcM Press.

Krebs, V. (2021). Social & Organizational Network Analysis software & services for organizations, communities, and their consultants.
http://www.orgnet.com/divided.html

Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical review E*, *78*(4), 046110.
https://doi.org/10.1103/PhysRevE.78.046110

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: can geographic isolation explain this unique trait?. *Behavioral Ecology and Sociobiology*, *54*, 396-405.
https://doi.org/10.1007/s00265-003-0651-y

Newman, M. (2010) Networks: An Introduction.
Oxford University Press, Oxford.
http://dx.doi.org/10.1093/acprof:oso/9780199206650.001.0001

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E*, *69*(6), 066133.
https://doi.org/10.1103/PhysRevE.69.066133

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, *74*(3), 036104.
https://doi.org/10.1103/PhysRevE.74.036104

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, *76*(3), 036106.
https://doi.org/10.1103/PhysRevE.76.036106

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, *105*(4), 1118-1123.
https://doi.org/10.1073/pnas.0706851105

Yang, X. S. (2012). Flower pollination algorithm for global optimization. In *Unconventional Computation and Natural Computation: 11th International Conference, UCNC 2012, Orléan, France, September 3-7, 2012. Proceedings 11* (pp. 240-249). Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-32894-7_27

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, *33*(4), 452-473.
https://doi.org/10.1086/jar.33.4.3629752

Zhou, G., Wang, R., & Zhou, Y. (2018). Flower pollination algorithm with runway balance strategy for the aircraft landing scheduling problem. *Cluster Computing*, *21*, 1543-1560.
https://doi.org/10.1007/s10586-018-2051-0