1-1-2022

A Systematic Analysis of Community Detection in Complex Networks

Haji Gul
City University of Science and Information Technology

Feras Al-Obeidat
Zayed University

Adnan Amin
Institute of Management Sciences Peshawar

Muhammad Tahir
Saudi Electronic University

Fernando Moreira
Universidade Portucalense

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 201 (2022) 343–350

Procedia Computer Science

www.elsevier.com/locate/procedia

The 13th International Conference on Ambient Systems, Networks and Technologies (ANT)
March 22 - 25, 2022, Porto, Portugal

A Systematic Analysis of Community Detection in Complex Networks

Haji Gul[a], Feras Al-Obeidat[b], Adnan Amin[c,*], Muhammad Tahir[d], Fernando Moreira[e]

[a]City University of Science and Information Technology, Peshawar 25000, Pakistan
[b]Zayed University,Abu Dhabi 51133,UAE
[c]Center for Excellence in Information Technology, Institute of Management Sciences, Peshawar 25000, Pakistan
[d]College of Computing and Informatics,Saudi Electronic University,Riyadh 11673,Saudi Arabia
[e]REMIT, IJP, Universidade Portucalense, Porto, Portugal & IEETA, Universidade de Aveiro, Aveiro, Portugal

Abstract

Numerous techniques have been proposed by researchers to uncover the hidden patterns of real-world complex networks. Finding a hidden community is one of the crucial tasks for community detection in complex networks. Despite the presence of multiple methods for community detection, identification of the best performing method over different complex networks is still an open research question. In this article, we analyzed eight state-of-the-art community detection algorithms on nine complex networks of varying sizes covering various domains including animal, biomedical, terrorist, social, and human contacts. The objective of this article is to identify the best performing algorithm for community detection in real-world complex networks of various sizes and from different domains. The obtained results over 100 iterations demonstrated that the multi-scale method has outperformed the other techniques in terms of accuracy. Multi-scale method achieved 0.458 average value of modularity metric whereas multiple screening resolution, unfolding fast, greedy, multi-resolution, local fitness optimization, sparse Geosocial community detection algorithm, and spectral clustering, respectively obtained the modularity values 0.455, 0.441, 0.436, 0.421, 0.368, 0.341, and 0.340..

Keywords: Community Detection; Graph Clustering; Graph Analysis; Complex Networks; Prediction and Recommendation;

1. Introduction

Complex networks are utilized to graphically express relationships or construction in numerous real-world systems including but not limited to terrorist, social, biomedical, technological, human, animal, or co-authorship. Understanding the interactions of organizational arrangements or considering certain frameworks demonstrating specific con-

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: adnan.amin@imsciences.edu.pk

1877-0509
10.1016/j.procs.2022.03.046

structions can reveal some information about infections and dispersion. Similarly, the logical investigation of complex network systems is a multidisciplinary field spanning over physical, social, and computer science. A graph or complex network system shows entities and their interactions with each other. Entities in the network are expressed by vertices or nodes whereas the interactions are represented by edges or links.Community recognition in a real-world network is, nonetheless, a crucial issue because of the shortfall of a widespread meaning of the item for discovery, i.e., "local area or community". Because of the vagueness, there is a dispersion of problematic writing in this field. Logical assessments got advanced from the old style of looking at a cluster as a thick sub-network graph to the cutting edge technology. It is worth noting that the old style graph were used to display a level of self-sufficiency in the organization due to the high edge thickness between vertices inside the graph rather than outside whereas the new technology relies on the calculation of probabilities of edge arrangements among vertices. The old style relied more on the degree of distribution of the vertices to decide about graph clusters leading to the concepts of solid networks and powerless networks that rely on the external and internal degree of the vertices under consideration [1]. In complex network analysis, community detection is an important and innovative problem. It could be, for the most part, expected that an organization is a 'community organization' if a portion of its vertices have a higher likelihood of being connected than vertices of different gatherings in the complex network. Regular divisions in complex networks differ in the density of associated vertices from vertices of different gatherings to which they are more averse of being associated. A number of philosophies and calculations have been fostered approaching the issue from different perspectives, such as statistical science [2], optimization [3], spectral [4], Bayesian and regularized [5] based approaches. Many researchers have used dynamical network procedures like diffusion [6], traditional greedy [4] and divide-and-conquer [7].Recently, some methods have been proposed and focused on avoiding the need to require the number of clusters [8].
Benchmark datasets are of utmost significance that can be used for estimating the effectiveness of the existing and newly proposed algorithms in the field. As reported in [9, 10, 11], each dataset has different network structures and properties such as entropy, average degree, cluster coefficient, and cliques, etc. The Grivan and Newman benchmark dataset has been broadly used to verify the performance of an algorithm and validate whether its produced results are aligned with the literature. Though, several characteristics are dissimilar to real-world complex networks, for instance, all vertices of the real-world networks have the same likelihood degree and different networks are of similar sizes [12]. Lancichinetti et al. have adapted a benchmark (LFR) that produces synthetic complex networks with already built communities. The LFR benchmark is broadly utilized for evaluating the nature of new local area discovery models. It is more sensible with its suspicion of force law dispersion for the degree of complex networks and community approximations, and it takes into consideration covering networks. LFR employs the blending boundary *mu* to show a small portion of a vertice's connections or edges that are outside of its allocated bunch i.e. the segment's nature. For lower upsides of μ, groups are better isolated from one another and are simpler to identify. As μ expands, it turns out harder progressively to recognize the cluster and allocate vertices to their legitimate networks.

2. Related Work

Recently, a number of algorithms have been proposed by the research community for community detection in various fields such as transportation, terrorism, human interaction, routers, and computer as well as real-world social complex networks [13]. Simultaneously, a survey paper looks to be the single community detection in social networks using node information [14]. Besides [15], a very detailed paper on the community identification topic was almost completed. It typically has a short survey of preceding methods and an attempt to classify them. We observed that many authors are just partly aware of the relevant literature and this sometimes leads to repetitions in approaches. Furthermore, multiple classifications are mainly incomplete and even contradictory.

2.1. Multiple Screening Resolution Community Detection Method

The author analyzed various modules at different resolution levels and increased the accuracy of community detection in complex networks, [16]. Mathematically, the multi-resolution formula can be expressed as given in equation 1:

$$Q_r = \sum_{s=1}^{m} \left(\frac{2w_{ss} + n_s r}{2w + Nr} - \left(\frac{w_s + n_s r}{2w + Nr} \right)^2 \right) \tag{1}$$

For the negative value of r the author achieved super-structure or high accuracy of community detection. In the given equation, s is the staring node of the network, W is the weight, and N is the set of nodes between start and end. While $r = 0$ relates to the topological scale established by maximising Q at which the detection of the modular structure has been tackled, which controls the path length n.

2.2. Greedy Community Detection Method

The Greedy community detection method was introduced in [17], which improved the accuracy of community detection for large-scale networks. In the given equation 2 $N(v_x)$ and $N(v_y)$ computation for the total nodes who are adjacent to v_x and v_y correspondingly. N is the set of nodes of x and y. While $W(v_x, v_y)$ calculates the status of the association inside the community, another parameter expressed by η is the revolving constraint, where the range is $[0, 1]$. Broadly speaking, the value of the η depends on the structure of the complex network. If the given complex network is higher or less sparse, then the value of η approaches 1 or 0. Finally, l calculates and regulates the length of the path between nodes x and y.

$$W(v_x, v_y) = \eta + \left(\frac{1}{N(v_x)} + \frac{1}{N(v_y)}\right)\ell(v_x, v_y) \tag{2}$$

2.3. Multi-Resolution Community Detection Method

S^C_{our} and S^C_{in} the given equations 4,5,6 and 7, expresses the similarities outside the community and inside the community. S^C_{out} is equal to the similarity score of S^C_{out}–S^C_{in} while only S^C_{in} is equal to twice the internal community similarities. where S^C_{in} can be calculated from the given equation 3. C is the sum of double similarity of between pairs of nodes inside the community. While, S^C_{our} can be computed from the given equation 3 where C is the external similarity of the community.

$$S^C_{in} = \sum_{uv \in C, uv \in E} S(u, v), \qquad S^C_{out} = \sum_{u \in C, v \in N, uv \in E} S(u, v) \tag{3}$$

For more details, see [18]. The given parameter, α, is used to control the proportionality changes of a candidate node a outside of the community.

$$\tau^{\alpha}_{C}(a) = \frac{S^C_{out}}{S^C_{in}} - \frac{\Delta S^C_{out}}{\Delta S^C_{in}} \tag{4}$$

$$\Delta S^C_{out} = S^C_{out} - S^C_{in} \tag{5}$$

$$\Delta S^C_{in} = 2S^C_{in} \tag{6}$$

$$\tau^{\alpha}_{C}(a) = \frac{S^C_{out}}{S^C_{in}} - \frac{\alpha S^a_{out} - S^a_{in}}{2S^a_{in}} \tag{7}$$

2.4. Local Fitness Optimization Community Detection Method

The fundamental suspicion behind this method is that complex networks have local organization, as [19], with vertices having a place with the actual modules, as well as a long locality of them.

$$f_{\mathcal{G}} = \frac{\delta^{\mathcal{G}}_{in}}{\left(\delta^{\mathcal{G}}_{in} + \delta^{\mathcal{G}}_{out}\right)^a} \tag{8}$$

In the given optimization equation 8, $\delta^{\mathcal{G}}_{in}$ express internal while $\delta^{\mathcal{G}}_{out}$ denoted external degree of the vertices inside module $\mathcal{G}$ finally α parameter used to work with the size of communities. These local fitness functions have the property of being able to be linearly integrated to generate the quality function of interest, and the quality function is optimised through node co-evolution. The difference between the ratio of real edge number and total degrees of nodes in the community with and without this node was defined as the local fitness function of each node.

2.5. Unfolding Fast Community Detection for Large Complex Networks

A higher accuracy algorithm given in equation 9, based on fast unfolding criteria proposed in [20]. The mathematical equation is given below:

$$\Delta Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)\right] \tag{9}$$

The given unfolding community detection equation ΔQ, where $\sum_{in}$ denoted the total sum of edges inside the community C, $\sum_{tot}$ express total edges weight occurrence to the vertices in community C, K_i express the count the weight of vertex i, m used for the total density of links in the network while $k_{i,in}$ represent the weight of link from i to community C. This algorithm initiates with the weighted network containing multiple nodes. Initially, it assigns each node to different communities. So, each and every has a community label. Next, for the starting node, assume the ending node. Then, for each starting node i we assess its neighbours j and calculate the increase in modularity that would result from shifting i from its community to the community of j. The node I is then assigned to the community with the highest gain, but only if the gain is favorable.

2.6. Multi Scale Community Detection Method

This algorithm works based on multi-scale community detection criteria [21], it figuring the modification between the previous and new community detection quality. The change in community detection i (previous) and new j (new community) modularity can be expressed by the given equation 10.

$$\Delta Q_{Mij} = 2(e_{ij} - a_i a_j) \tag{10}$$

During each step, the best partition result has been kept in Q_s given in equation 11, and then it is updated as $Q_s = Q'_s$. For computation, the time range starts from 0 to t. This algorithm is very efficient for community detection. In the given equation, e_{ij} is the total edges, also called the universal set of edges. While a_i, a_j are the specific start and end nodes, they finally multiply with 2. Based on equation 10, the community has been predicted given in equation 11 and updated with respect to time.

$$Q'_S = min_{0\le t\le \tau}\,(Qs_t + \Delta Qs_t) \tag{11}$$

2.7. Spectral Community Detection Method

Spectral community detection is given in equation 12, where $\delta(\sigma_i, \sigma_j)$ control the path length among i and j. While k_i and k_j express the total degrees of each spin condition. Finally, the parameter A_{ij} is the connected matrix among starting node i and destination node j.

$$Q = \frac{1}{2M}\sum_{i\neq j}\left(A_{ij} - \frac{k_i k_j}{2M}\right)\delta(\sigma_i \sigma_j\text{`}) \tag{12}$$

2.8. Sparse Community detection for Geo-social Data

Sparse scale is a higher accuracy community detection algorithm [22], the mathematical equation 13 is given below.

$$W_{i,j} = \alpha S_{i,j} + (1-\alpha)e^{\frac{-d^2_{i,j}}{\sigma^2}} \tag{13}$$

Here, S_{ij} the similarities among the initial and ending nodes, while α used for the distance. Its range is between 0 and 1. $\delta_{i,j}$ is the normal Euclidean path length between the average halt positions of each i and j, while σ denotes the range of standard deviation which is greater than the mean distance. Sparse community detection is an accurate and popular algorithm due to its simplicity. In the given equation 13, α controls the length of the path among the source and destination nodes i and j. The variable alpha multiplies the value with the similarity variable S. The alpha range is $0-1$. Additionally, the σ is used to check the error rate of the algorithm for community detection. Finally, the end result of the similarity store is in W.

3. Proposed Systematic Analysis Framework

In this paper, we introduce a systematic framework to address the problem of community detection. The complete systematic methodology is shown in figure 1. First, the dataset is imported and then irrelevant information has been removed from it, such as loops, weights, and directions of the edges. This is extra information in our case, and it increases the experiment throughput time. There are a total of nine datasets taken from different domains such as animals, humans, transportation, and terrorists. Next, the adjacency matrix has been created, which is the representation of a graph. The entries in the matrix denoted as 1 mean that the links exist if $e = (x, y)$ whereas the entries of 0s indicate the absence of links $e \neq (x, y)$. Then, all the algorithms under consideration have been applied for detecting communities in the networks.
In order to evaluate the performance of the algorithms, modularity evaluation metric has been utilized to check the efficiency. The modularity metric was calculated over 100 iterations and the average is produced as a final result, which is shown in table 1. The obtained results clearly show that the multi-scale algorithm performs best as compared to others. Multiple screening resolutions provide the second-highest modularity. A total of eight state-of-the-art algorithms have been analyzed over nine datasets.

3.1. Evaluation Procedure, Modularity

Modularity is an evaluation method that aims to check the quality of a community in a complex network [23]. It is the most important and relevant accuracy measurement metric used in different fields to check the quality of the community [24, 25, 26].In addition, researchers in the literature assumed modularity to assess a community's quality or reliability. Modularity is produced by the links between the nodes of an organization assuming only the absence or presence of links in a network. The mathematical representation of modularity is given by Equation 14.

$$Q = \frac{1}{2M} \sum_{k=1}^{C} \left[\alpha_{x,y} - \frac{\sigma_x \sigma_y}{2M} \right] .\delta(C_x, C_y) \tag{14}$$

A_{xy} represents the connection between the source node x and the destination node y. A link between x and y can also be expressed by $e = (x, y)$. If there is no connection between $x$and y, then $A_{xy} = 0$. In Equation 14, M denotes the total number of links in a given complex network, σ_x and σ_y show the degrees of nodes x and y, respectively, and C_x denotes the community where vertex x exists and C_y expresses the community where vertex y exists.

3.2. Data Preprocessing

The development of the community detection procedure has begun with the creation of adjacency matrix from the complex network where the matrix is denoted as A_{xy}. The entries of the adjacency matrix are binary values that represent the presence or absence of a link between two given nodes or vertices. The presence of a link between two nodes can be mathematically expressed as $e = (x, y)$ where x is the source node and y is the destination node. In a real-world complex network dataset, a link may be an interaction between proteins, a contact between humans, a connection between routers, a relationship between animals, and a friendship between people over social media. All the datasets utilized in this work belong to real-world complex organizations where most of the complex network datasets are obtained from [1] [2]. There are nine complex network datasets that have been used, and a total of eight algorithms are evaluated over these data-sets. The datasets are:

- Karate •Dolphin •Kangaroo •Human Contact •MISC •Train Bombing •Foot Ball •Contiguous USA •Zebra

[1] https://networkrepository.com/
[2] https://snap.stanford.edu/data/

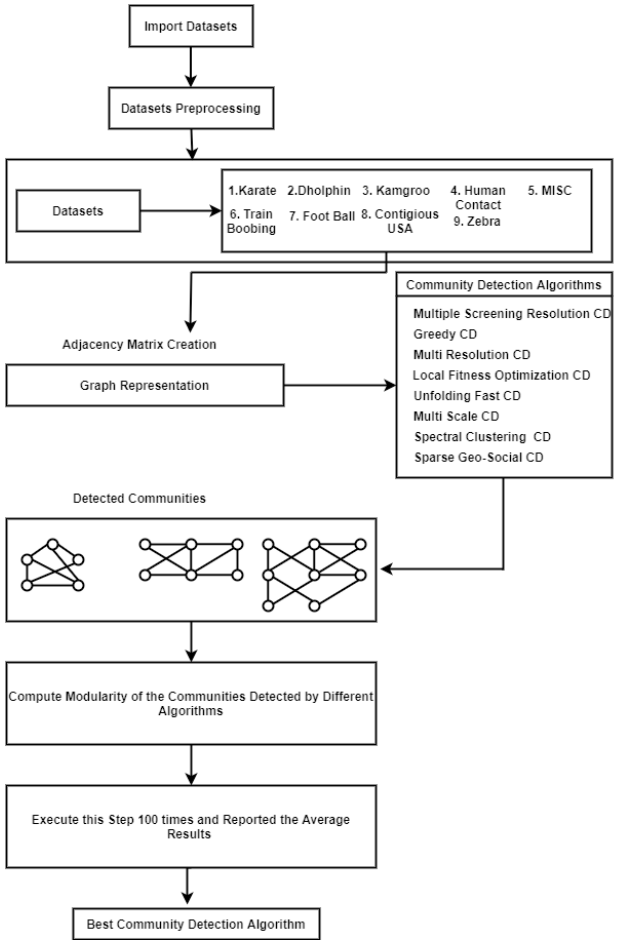

Fig. 1. Methodology

4. Results

Table 1 highlights the obtained results after applying the algorithms on a range of different datasets for community detection. As evident from the obtained results, the multi-scale community detection algorithm demonstrated the highest performance over four datasets, namely karate, dolphin, misc, and the Contagious USA. The multi-screening resolution community detection algorithm is the second best performer over three datasets: kangaroo, terrorist, and Zebra whereas the third best performer is unfolding fast community detection algorithm that performed best over one dataset i.e. human contacts. Figure 2 shows the results of each algorithm for each dataset where each algorithm is represented by different color to indicate its respective performance for a given dataset. It can be observed that the performance of the algorithms is not good for Zebra and Human contact datasets.

The average performance of each algorithm over 100 iterations is depicted in Figure 3 where the performance of multi-scaled community detection algorithm is observed to be superior compared to other algorithms in the same settings.

Table 1. The average result of each algorithm stated in this table after the execution of 100 times modularity evaluation metric.

Algorithms	Karate	Dolphin	Kangaroo	Human Cont	MISC	Train Bomb	Foot Ball	Con USA	Zebra
Multiple Screening Resolution CD	0.4149	0.51228	**0.59836**	0.25415	0.5566	**0.44646**	0.39989	0.58604	**0.2768**
Greedy CD	0.39549	0.49363	0.57836	0.23814	0.51971	0.42054	0.35614	0.5691	0.27656
Multi Resolution CD	0.4154	0.48842	0.4125	0.25403	0.53615	0.42145	0.38521	0.57046	0.27023
Local Fitness Optimization CD	0.38998	0.40585	0.29827	0.13435	0.51701	0.39557	0.39011	0.53982	0.27023
Unfolding Fast CD	0.39917	0.4904	0.57827	**0.27433**	0.53554	0.44208	0.32989	0.55473	0.2558
Multi Scale CD	**0.42428**	**0.52223**	0.59816	0.25746	**0.55785**	0.41235	**0.44547**	**0.58713**	0.2634
Spectral Clustering CD	0.23468	0.41815	0.59716	0.17508	0.50886	0.35669	0.37985	0.55446	0.2726
Sparse Geo-Social CD	0.31967	0.38703	0.28493	0.18168	0.43304	0.38659	0.31989	0.50098	0.23786

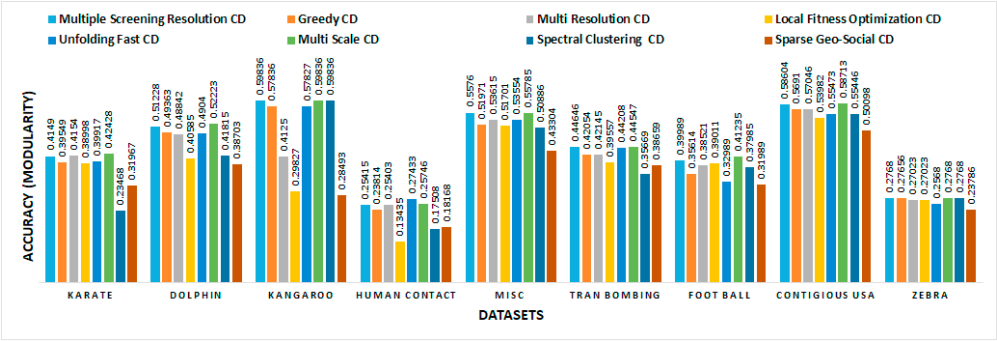

Fig. 2. Results of Each Algorithm for Every Dataset

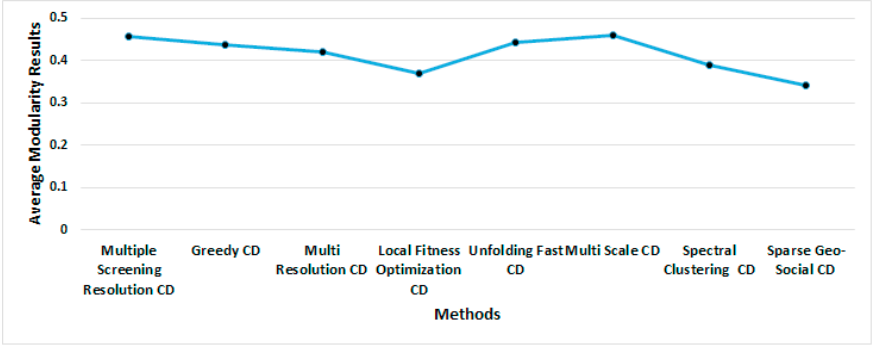

Fig. 3. It stated the average modularity results of each community detection algorithm. Y-axis express algorithms while X-axis denoted modularity results.

5. Conclusion

In this paper, we have compared eight community detection algorithms over nine real-world complex network datasets, which belong to a variety of domains including human, animal, contagious, USA, MISC, terrorist, and transportation. Community structure prediction aims at grouping the most similar or related nodes in a given network. The modularity evaluation metric has been used as a performance measure to evaluate the performance of each algorithm over 100 iterations for community structure detection. The experimental results have shown the significance of multi-scale method that outperformed other techniques under consideration. In the future, we intend to apply the existing algorithms to extra-large dynamic and complex networks to evaluate their performance. Further, we plan to develop a novel algorithm for more generalized performance over a range of networks. Similarly, development of a stable com-

munity detection algorithm will also be considered as a future plan. We also intend to minimize the computational complexity of the existing algorithms.

Acknowledgements

The authors are grateful to the anonymous reviewers for their insightful comments and suggestions, which helped improve the quality of this paper. Dr. Feras Al-Obeidat of Zayed University supported this work.

References

[1] S. Fortunato, D. Hric, Community detection in networks: A user guide, Physics reports 659 (2016) 1–44.
[2] Y. R. Wang, P. J. Bickel, Likelihood-based model selection for stochastic block models, The Annals of Statistics 45 (2017) 500–528.
[3] A. Lancichinetti, F. Radicchi, J. J. Ramasco, S. Fortunato, Finding statistically significant communities in networks, PloS one 6 (2011) e18961.
[4] J. Sánchez-Oro, A. Duarte, Iterated greedy algorithm for performing community detection in social networks, Future Generation Computer Systems 88 (2018) 785–791.
[5] X. Yan, Bayesian model selection of stochastic block models, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, pp. 323–328.
[6] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: J. Graph Algorithms Appl, Citeseer, 2006.
[7] Z. Liu, Y. Ma, A divide and agglomerate algorithm for community detection in social networks, Information Sciences 482 (2019) 321–333.
[8] Z. Zhao, C. Li, X. Zhang, F. Chiclana, E. H. Viedma, An incremental method to detect communities in dynamic evolving social networks, Knowledge-Based Systems 163 (2019) 404–415.
[9] F. Aziz, H. Gul, I. Muhammad, I. Uddin, Link prediction using node information on local paths, Physica A: Statistical Mechanics and its Applications 557 (2020) 124980.
[10] F. Aziz, H. Gul, I. Uddin, G. V. Gkoutos, Path-based extensions of local link prediction methods for complex networks, Scientific reports 10 (2020) 1–11.
[11] H. Gul, A. Amin, F. Nasir, S. J. Ahmad, M. Wasim, Link prediction using double degree equation with mutual and popular nodes, in: World Conference on Information Systems and Technologies, Springer, 2021, pp. 328–337.
[12] R. Baggio, Network science and tourism–the state of the art, Tourism Review (2017).
[13] Z. Yang, R. Algesheimer, C. J. Tessone, Corrigendum: A comparative analysis of community detection algorithms on artificial networks, Scientific reports 7 (2017).
[14] P. Chunaev, Community detection in node-attributed social networks: a survey, Computer Science Review 37 (2020) 100286.
[15] C. Bothorel, J. D. Cruz, M. Magnani, B. Micenkova, Clustering attributed graphs: models, measures and methods, Network Science 3 (2015) 408–444.
[16] A. Arenas, A. Fernandez, S. Gomez, Analysis of the structure of complex networks at different resolution levels, New journal of physics 10 (2008) 053039.
[17] A. F. Al-Mukhtar, E. S. Al-Shamery, Greedy modularity graph clustering for community detection of large co-authorship network, Int. J. Eng. Technol 7 (2018) 857.
[18] J. Huang, H. Sun, Y. Liu, Q. Song, T. Weninger, Towards online multiresolution community detection in large-scale networks, PloS one 6 (2011) e23829.
[19] H.-W. Shen, Detecting the overlapping and hierarchical community structure in networks, in: Community Structure of Complex Networks, Springer, 2013, pp. 19–44.
[20] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of statistical mechanics: theory and experiment 2008 (2008) P10008.
[21] E. Le Martelot, C. Hankin, Multi-scale community detection using stability as optimisation criterion in a greedy algorithm., in: KDIR, 2011, pp. 216–225.
[22] Y. Van Gennip, B. Hunter, R. Ahn, P. Elliott, K. Luh, M. Halvorson, S. Reid, M. Valasik, J. Wo, G. E. Tita, et al., Community detection using spectral clustering on sparse geosocial data, SIAM Journal on Applied Mathematics 73 (2013) 67–83.
[23] M. E. Newman, Modularity and community structure in networks, Proceedings of the national academy of sciences 103 (2006) 8577–8582.
[24] A. J. Schwarz, A. Gozzi, A. Bifone, Community structure and modularity in networks of correlated brain activity, Magnetic resonance imaging 26 (2008) 914–920.
[25] D. Meunier, R. Lambiotte, A. Fornito, K. Ersche, E. T. Bullmore, Hierarchical modularity in human brain functional networks, Frontiers in neuroinformatics 3 (2009) 37.
[26] B. A. Siebert, C. L. Hall, J. P. Gleeson, M. Asllani, Role of modularity in self-organization dynamics in biological networks, Physical Review E 102 (2020) 052306.