Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 151 (2019) 295–302

Procedia
Computer Science

www.elsevier.com/locate/procedia

The 10th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 29 – May 2, 2019, Leuven, Belgium

A comprehensive literature review on community detection: Approaches and applications

Mohamed EL-MOUSSAOUI*, Tarik AGOUTI, Abdessadek TIKNIOUINE, Mohamed EL ADNANI

Computer Science Department, Faculty of Sciences, Cadi Ayyad University, Marrakech, Morocco

Abstract

Community detection has been designed as an axial field in Complex Network Analysis (CNA), since it allows to reveal cohesive and meaningful sub-graphs, recognize the features, functions, structure and dynamic of such complex networks. In this sense, various methods and approaches have been developed over the years to provide appropriate solutions to complex network paradigms, especially to community detection problems. Meanwhile, identifying communities in a given complex network is a big challenge for scientists, which needs a significant amount of literature and survey. In this paper, we detail literatures on community detection for complex networks, because of the need for researchers to perform reviews on the main papers related to identification of communities in complex networks, and in order to point out their principal strengths and limitations. We believe that this literature contribution can be a valuable source of information in particular for practitioners in the field of community detection, and do not include all existing contributions. Therefore, we have been interested in the value of the contribution of the selected approaches more than the chronological order of the publications.

Keywords: Complex Network, Social Network, Community Detection, Network Analysis, comparative analysis.

* Corresponding author.
E-mail address: mohamed.el-moussaoui@ced.uca.ma

1877-0509 © 2019 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2019.04.042

1. Introduction

Complex Network Analysis (CNA) constitutes recently an important field of research, it is widely involved in social science, neuroscience and biology, metabolic networks and food webs and many others. Network analysis is considered as one of basic pillars of the discrete mathematics, while it represents a form of mathematical graph theory [1, 2]. Since networks become immeasurably huge, and voluminous data could be modelled as complex networks, Complex Network Analysis (CNA) gained attention of scientist communities to bring valuable explanations to complex networks behaviours, functions and models. For example, the observation of similarities of individuals in a given social network (e.g., Facebook, LinkedIn, Twitter, …etc.) represents their global characteristics which reflect specific behaviours and particular organization. Hence, we identify those groups of individuals as community or cluster. Their identification could help to understand how complex networks work and how they are structured.

At this organization level, a community which represents a group of nodes sharing common and similar properties, is an important tool for network analysts to understand interacting behaviors, cohesive sub-groups of the network. Thus, a growing number of community detection approaches and methods have been published since the last years. They differ from one to the other in defining criteria of identification of communities. Many of this approaches have been applied successfully in different domains of applications, we provide later a detailed comparative summary review table where we distinguish some variables used for the comparison.

In this paper, we provide a comprehensive review and categorization of different approaches and methods on community detection, in order to point out their principal strengths and limitations. In addition, review researches dealing with application on different domain. Finally, we focus on journal papers written in English, containing valuable information for practitioners interested in community detection problem, and providing the latest results, and finally which are accessible via online databases. The rest of the paper is structured as follows: In Section 2, we provide a brief review of the fundamental concepts of network and community structure. Section 3 presents the literature review. In Section 4, we draw a classification summary of different papers with its application domains. Finally, in Section 5 we discuss obtained result and draw our conclusions and a number of perspective future research.

2. Fundamental concepts

In this section, we provide some primer concepts and definitions that will be used throughout this review.

2.1. Complex Network

A complex network is a representation of complex system representing virtually any discrete system from real life, the basic displayed properties for those kind of system are as follow:

- Emergence of behaviors of the complex network components.
- Self-organization of the network.
- Nonlinearity of interactions among components.
- Evolution of components.

It is practically an application area of graph theory, which is related to a multidisciplinary field of research.

2.2. Network measurements

In order to distinguish a particular node in a given complex network (for example social network) composed of very large number of nodes, a number of measurements are provided as characteristics that represent a development tools to gain greater insights into networks and provide deep analysis for single or community of nodes.

In general, for complex networks analysis, and especially for social network, the following class of measurements are defined:

- Connectivity inside the network structures: Network connection, transitivity, Multiplicity, Reciprocity, …etc.
- Network distribution: PageRank, betweenness centrality, degree centrality, shortest path, …etc.
- Network community or clustering: k-core, hierarchical clustering, cliques, …etc.

Acknowledgment of these measures and classes of network, allow to describe precisely networks.

2.3. Network Structure

We recognize two major components for a given network: (i) Nodes representing actors of the network, and (ii) edges representing links or relationships between nodes.

2.4. Community structures

This concept was proposed firstly in social sciences [3]. In social networks it is a sub-graph for which the nodes are cohesive, densely connected internally than the rest of the network and sparsely connected with others. For instance, pages with a similar content in web page network and groups of individuals in social networks represent communities of a complex network.

3. Literature review

A wide research study in the recent years focused on community detection in complex systems [4], most of them focus on undirected networks to enhance the efficiency of identifying communities in understanding complex networks. For instances, Fortunato et al [3] based his proposed approach on statistical inference perspectives, Schaeffer et al [5], proposed their approach for clustering problem as an unsupervised learning task based on similarity measure over the data of the network, Girvan and Newman based their community detection proposal on betweenness calculation to find out community boundaries where modularity measure is the overall quality of the graph partitioning [6, 7]. The weight used by Newman and Girvan [7] aims to be the betweenness measure of the edge, representing the number of shortest paths connecting any pair of nodes passing through.

However, community detection problem has been studied mainly in case of undirected networks, various solutions was proposed in this context, motivating many disciplines to deal with the issue. Interestingly, Fortunato et al [3] mentioned the few possibilities for extending techniques from undirected to directed case, where the edge directedness is not the only complication that could face the clustering problem. Nevertheless, diverse graph data in many real-world applications are by nature directed, thus interesting to save available information behinds the edge directionality. Malliaros et al [8], revealed in their survey that the most common way for researcher community to deal with the problem of clustering is to ignore the directionality of the graph, then proceed to clustering with a wide range of proposed tools. Therefore, most of community detection proposals can not be used directly on weighted directed graphs, where the number of communities not always known in advance and the communities present different granularity scale. Since the problem of community detection in complex network analysis acquires more attention, many researchers have been interested into structural information and topological networks metrics [1, 3, 4, 6, 7, 8, 9]. In [10], S. Ahajjam based the proposed community detection algorithms on a new scalable approach using leader nodes characteristics through two steps: (i) Identification of potential leaders in the network, and (ii) exploration of nodes similarities around leaders to build communities. Therefore, recent works start focusing on both topological and topical aspects [9, 11, 12] to overpass limited performances of topology-based community detection approaches.

Topic-based community detection started gaining attention through different works for community detection in complex network [9, 13, 14]. The essence behind the approach is to similarly detect nodes with same properties, which are not necessarily real connections between nodes of the network, in which actors communicate on topics of mutual interest [14] to determine the communities which are topically similar.

4. Community detection approaches

In this chapter, we provide an overview of different approaches through selective papers covering most of technical aspects related to community detection algorithms.

An important number of papers contributed with different proposals to bring resolution to the problematic of community detection as in [2, 3, 4, 12, 15, 16, 17, 18, 19, 20, 21, 22]. Fortunato [3] described in his precious work more that 50 methods and more than 250 algorithms [23] without being exhaustive. Meanwhile, some contributions represent the reference for researchers in terms of community detection algorithms [6, 7, 18, 24]. Nowadays, the field of community detection still in development, and the categorization of an exhaustive list of methods and approaches seems not yet fixed. We summarized some famous approaches through Table 2 referring to category of applied algorithms, nature of network, directionality of the network, type of target community (static/dynamic), network size, and so on. The most suitable classification found through examination of all referred papers is as follow:

- Approach based static non-overlapping communities;
- Approach based static overlapping communities;
- Approach based static hierarchical communities;
- Approach based dynamic communities.

5. Classification summary

In this section, we performed a comparison of different works basing on characteristics that differ from approach to another. The selected papers do not represent the exhaustive list of research on community detection field. The most known approaches for community detection was discussed and described within advantages of each one [23], with respect to previous survey on the subject, the comparison seems to be a difficult task for the main reason that each approach is based on different metrics and structures. We choose to base our comparison on the following characteristics:

- **Approach**: Used methodology for community detection finding;
- **Technical principal of approach**: We describe in this column the technical principal or algorithm of the applied approach;
- **Network type**: This attribute includes “Weighted” in case edges linking nodes of the studied network are weighted, and “Unweighted” in case edges do not have weights;
- **Directionality of the network**: Indicates “Directed” in case hyperlinks between nodes of the network are directed, and “Undirected” if directionality of the hyperlink is ignored. “All” is mentioned if the approach can support both structures;
- **Network nature**: This attribute describes either the network is static or dynamic;
- **Network size**: The network size is an important metric for computation performance of the approach. We provide in this column the size of the supported network;
- **Implementation datasets**: We provide the common abbreviation according to Table 1, of the used network datasets for each approach experimentation.

Table 1. Implementation network datasets used in the referred papers of the main comparison (table 2).

Network Dataset	Common Abbreviation	Description
LFR Benchmark	LFR	Lincoln Fire & Rescue (LFR) Benchmark.
Zachary’s Karate Club Friendship Network	ZKC	Contains the network of friendships between 34 members of a karate club at a US university.
Dolphin Social Network	DLP	Contains an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand.

American College Football Network	ACF	Dataset that contains the network of American football games between Division IA colleges during regular season Fall 2000.
Books Network	BN	Network of books from the online bookseller Amazon.com.
Computer-Generated Graphs	CGG	A large set of artificial graph. Each graph is constructed by 128 vertices.
Collaboration Network of Scientists	CN	A network of scientists in residence at the Santa Fe Institute, with 271 vertices.
Scientific Communication Network	SCN	A dataset which traces communication between scientists. Contains 6128 journals and around 6,5M citations tracing the scientific activity.
Food Web Network	FWN	Contains 33 nodes describing the ecosystems of marine organisms.
Jazz musician collaborations Network	JMC	Dataset that describes associations between 198 Jazz musicians. Relationships are created once musician played together in a cancer.
Physics E-print Archive	PEA	A dataset of electronic archive and distribution server for research.
Primate Network of Mankeys	PNM	Dataset recording 3 months of interactions between 20 monkeys.
P2P Networks	P2P	Peer-to-peer networks datasets, suited to distributed data mining (DDM).
Co-Authorships Network	CAN	A co-authorship dataset with a total of 1589 scientists.
Belgian Mobile Phone network	BMP	A network of communications between users of Belgian phone operator. With 2.6M of vertices and cumulative calls duration as edges weights.
Web Graph	WG	A dataset of billions of web pages and billions of hyperlinks.
Protein-protein Interactions	PPI	Datasets of proteins interacting with hub proteins.
Email Network (EM)	EM	Email dataset, tracing incoming and outgoing email.
Amazon datasets	AM	A two million nodes Amazon co-purchasing dataset.
High School Networks	HS1, HS2	Social networks of high schools, self-reported by students and conducted in a project by the National Institute of Child Health and Human Development (NICHHD).
Word Associations of Cognitive Sciences	WACS	The largest free association cognitive database collected in USA, more than 6K participants.
WebBase Network	WBN	Stanford WebBase crawler datasets, with more than 39M and 783M edges.

The above Table 1 describes a list of synthetic networks used for benchmarking community detection algorithms and approaches to obtain good performance metrics, and to assess the performances of the proposed approaches. Practically, the comparison between obtained results of a proposed approach against existing ones according to the same network datasets, allows to realize how efficient is the proposed work, and insights researchers to the achieved results. Hence, testing community detection algorithms have been considered as a standard practice. In addition, the choice of experimental graph datasets represents a necessary step and provides real-world networks with realistic challenges.

We describe through the below comparison table, examples of used implementation datasets (referring to Table 1) in each mentioned reference, in addition to diverse characteristics as follow:

Table 2. Community detection approaches comparison.

Ref	Approach	Technical principal of the approach	Network type	Network Nature	Network Direction	Network size	Implementation datasets
[1]	Divisive	Local Modularity Optimization	Unweighted	Static	Undirected	Large	LFR
[25]	Divisive	Modularity Maximization	Unweighted	Static	Undirected	Medium	ZKC, DLP, ACF, BN
[26]	Random walk	Network Structure (Node Neighbors & Spread Capability)	Weighted	Static	All	Large	LFR
[6]	Divisive	Network Structure	Weighted	Static	Undirected	Large	CGG, ZKC, ACF, CN,

		(Centrality)					FWN
[7]	Divisive	Modularity	All	Static	All	Large	ZKC, ACF, JMC, PEA
[27]	Hierarchical	Network Structure (Centrality)	Unweighted	Static	Undirected	Large	ZKC, ACF, FWN, PNM
[28]	Divisive	Modularity	Unweighted	Static	Undirected	Medium	CAN, BMP
[29]	Agglomerative	Modularity Optimization	Unweighted	-	Directed	Large	BMP, WG
[30]	Random walk	Hierarchical agglomerative algorithm	Weighted	Static	Undirected	Large	ZKC, ACF, PPI, SCN, WG
[31]	Label Propagation	Network Structure (Neighbors)	Unweighted	Static	Undirected	Large	ZKC, ACF, CAN, CN, PPI, WG
[19], [35]	Label Propagation	Network structure	Unweighted	Static	Undirected	Large	LFR, ZKC, ACF, EM, AM, HS1, HS2, DLP, P2P
[32]	m-clique	-	Unweighted	Static	Undirected	Large	CN, WACS, PPI
[33]	Hierarchical	Hierarchical agglomerative algorithm	Unweighted	-	Directed	Medium	AM
[29]	Cohesive Agglomerative	Modularity Optimization	Weighted	-	Undirected	Large	ZKC, WG, BMP, WBN
[34]	Random walk	Information Flow	Weighted	All	Directed	Large	SCN

The bellow comparison provides an overview of most used approaches that concentrate around divisive, Hierarchical, Random walk, Agglomerative, Cohesive agglomerative, m-clique and Label propagation. All of these approaches demonstrate their robustness through different implementations [23], and provide different result distinguished by their complexity, nature of identified communities and time of execution.

Regardless the importance of each one of these approaches, the objective to identify communities in a given network, bring the question to what kind of community to look for? In what kind of network? What kind of metrics to consider? And what network datasets to use for experimentation?

In this context, this paper aims through the provided comparison review to help researchers in answering those questions, in order to point out the principal strengths of each approach, and to identify the appropriate directives for their future community detection contributions. We believe that the provided literature review does not include all existing contributions in the field of community detection.

6. Discussion

Applications of community detection in real life became an emerging area, especially in social networks but also in several domains as transport, communication, medicine and so on, it simplifies the perception for the word complexity. For example, a very large network could be reduced to smaller ones by introducing the community structure concept. For this reason and all, we invite new researchers through the above overview to enhance their perception to complexity by getting ideas of pioneers of the community detection field.

Most of reviewed works share the statement that methods are made generally for specific conditions related to target network environment. For instance, a dynamic approach will achieve the community detection problem as far as it can be achieved by a static approach, especially when taking into consideration the dynamic characteristics.

We observe through the provided comparison (Table 2) that papers which are based on divisive approach represent the most entertained in academic literatures. Static and undirected network metrics represent the best preference for divisive approach. Directed and Dynamic network metrics seem to be avoided due to complexity that encumber. Nowadays, static approaches refer to approaches making static communities easy to find, and dynamic approaches referring to dynamic communities. Meanwhile, some works propose to cohabitate many methods to manage the community detection problem, based on the nature of the target communities: hierarchical, overlapped communities, cliques, communities based on hubs which enrich the existing contributions.

Moreover, the dynamic approaches dealing with dynamic communities are in high speed of development, because they perfectly match with the real-world evolving complex networks, and represent the lifecycle phenomena. A well as the availability of several dynamic network datasets, making possible for researchers to experiment their community detection algorithms and approaches, to check characteristics and explicitly explore how communities are formed? And how they evolve and die? Practically, some approaches were applied to detect dynamic communities without taking into account characteristics of dynamic networks. Therefore, the challenge of detecting dynamic communities is defined in handling six main operations: (i) The birth, (ii) The death, (iii) The growth, (iv) The contraction, (v) The fusion and (vi) The division. Thus, most of proposed dynamic approaches still handling above operations in a separate manner. Indeed, a proposal of literature review for dynamic approaches dealing with instability of dynamic environments can be object of the next work.

7. Conclusion

In this paper, we introduced comprehensive taxonomy of community detection approaches in complex networks. We presented through a synthetic study, the main approaches and methods making success in analysis performances and quality of obtained results. We compared between them via various characteristics (Table 2) employed as main indicators of this paper in the classification summary section. Thereafter, we observed an important number of algorithms and approaches applied to static networks instead of dynamic ones, which makes static approaches acquiring a high level of maturity for implementation and tests, reduced complexity calculation and improved real-world networks datasets to balance performances of developed approaches and algorithms. Afterwards, this dynamic activity in static approaches made a remarkable diversity of published papers from different topics and domains.

The focus of researcher communities on dynamic approaches has been started once the concept of time and order integrated networks attributes as important characteristics of nodes (vertices) and links (edges). Meanwhile, dynamic approaches still under exploration and promise good performances, taking into consideration changes of characteristics of real-world complex network. In fact, most real-world complex networks in nature are by default dynamic, which makes sense to develop more efficient dynamic approaches, through handling dynamic community detection problem.

As features of this work, an important challenge taking place for the dynamic approaches and attracting attention of researchers either in combining existing methods or in exploring limits of others. We focus on the next work on adaptation of existing methods to achieve community detection challenge for dynamic evolving networks.

References

[1] J. Xiang, Z.Z. Wang, H.J. Li, Y. Zhang, S. Chen, C.C. Liu, …, L.J. Guo. (2017) “Comparing local modularity optimization for detecting communities in networks.” *International Journal of Modern Physics C Vol. 28, No. 6*

[2] A. Lancichinetti, S. Fortunato, and F. Radicchi. (2008) “Benchmark graphs for testing community detection algorithms.” Physical Review E 78.

[3] S. Fortunato. (2010). “Community detection in graphs”. *Physics Reports, 486 (3-5), 75–174.*

[4] X.S. Zhang, R.S. Wang, Y. Wang, J. Wang, Y. Qiu, L. Wang and L. Chen. (2009) “Modularity optimization in community detection of complex networks. *EPL (Europhysics Letters), 87(3).*

[5] S.E. Schaeffer. (2007). “Graph clustering”. *Computer Science Review, 1(1), 27-64.*

[6] M. Girvan and M.E.J. Newman (2002) “Community structure in social and biological networks”', *Proceedings of the National Academy of Sciences, 99 (12), 7821-7826.*

[7] M.E.J. Newman and M. Girvan. (2003). “Mixing Patterns and Community Structure in Networks”. *Lecture Notes in Physics, 66-87.*

[8] F.D. Malliaros and M. Vazirgiannis (2013) “Clustering and community detection in directed networks: A survey.”, *Physics Reports, 533(4), 95-142.*

[9] J. D. Cruz, C. Bothorel, and F. Poulet. (2014) “Community detection and visualization in social networks: Integrating structural and semantic information”. *ACM Transactions on Intelligent Systems and Technology, 5(1), 1–26.*

[10] S. Ahajjam, M. El Haddad and H. Badir. (2018). “A new scalable leader-community detection approach for community detection in social networks”. *Social Networks, 54, 41-49.*

[11] M. Qin, D. Jin, K. Lei, B. Gabrys and K. Musial (2017) "Adaptive Community Detection Incorporating Topology and Content in Social Networks", *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 675–682.*

[12] Z. Xia and Z. Bu (2012) "Community detection based on a semantic network", *Knowledge-Based System, 26, 30-39.*

[13] T. Nguyen, D. Phung, B. Adams, T. Tran and S. Venkatesh (2010) "Hyper-Community Detection in the Blogosphere", *Proceedings of Second ACM SIGMM Workshop on Social Media – WSM'10.*

[14] N. Pathak, C. DeLong, A. Banerjee and K. Erickson (2008) "Social Topic Models for Community Extraction", *In the 2nd SNA-KDD Workshop.*

[15] S. Jin, P.S. Yu, S. Li and S. Yang (2015) "A parallel community structure mining method in big social networks", *Mathematical Problems in Engineering, 2015, 1-13.*

[16] Y. Zhou, H. Cheng and J. X. Yu (2009) "Graph clustering based on structural/attribute similarities"', *Proceeding if the VLDB Endowment, 2(1), 718-729.*

[17] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi (2004) "Defining and identifying communities in networks", *Proceeding of National Academy of Sciences, 101(9), 2658-2663.*

[18] M.E.J. Newman and M. Girvan (2004) "Finding and evaluating community structure in networks.", *Physical Review E, 69(2).*

[19] J. Xie and B.K. Szymanski (2012) "Towards Linear Time Overlapping Community Detection in Social Networks", *16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Vol 2, 25-36.*

[20] G.P. Clemente and R. Grassi (2018) "Directed clustering in weighted networks: a new perspective", *Chaos, Solitons & Fractals, 107, 26-38.*

[21] L. Bai, J. Liang, H. Du, and Y. Guo. (2018). "A novel community detection algorithm based on simplification of complex networks". *Knowledge-Based Systems, 143, 58–64.*

[22] T. Wan, W. Liu and Z. Liu. (2014) "A Community Discovering Method Based on Event Network for Topic Detection." *In proceeding of the 16th International Conference on Advanced Communication Technology, (ICACT '14), pp. 1242–1246.*

[23] M. Coscian F. Giannotti and D. Pedreschi (2011) "A Classification for Community Discovery Methods in Complex Networks", *Statistical Analysis and Data Mining, 4(5), 512-546.*

[24] E.A. Leicht and M.E.J. Newman (2008). "Community structure in directed networks". *Physical Review Letters, 100(11).*

[25] B. Saoud and A. Moussaoui (2018) "Node Similarity and Modularity for Finding Communities in Networks", *Physica A: Statistical Mechanics and Its Applications, 492, 1958-1966.*

[26] M. Tasgin, H.O. Bingol (2018) "Community detection using preference networks", *Physica A: Statistical Mechanics and Its Applications, 495, 126-136.*

[27] S. Fortunato, V. Latora and M. Marchiori (2004) "Method to find community structures based on information centrality", *Physical Review E, 70 (5).*

[28] M.E.J. Newman (2016) "Finding community structure in networks using the eigenvectors of matrices", *Physical Review E, 74(3).*

[29] V.D. Blondel, J.L. Guillaume, R. Lambiotte and E. Lefebvre (2008) "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics; Theory and Experiment, 2008(10), P10008.*

[30] P. Pons and M. Latapy (2006) "Computing Communities in Large Networks Using Random Walks", *Journal of Graph Algorithms and Applications, 10(2), 191-218.*

[31] U.N. Raghavan, R. Albert and S. Kumara (2007) "Near linear time algorithm to detect community structures in large-scale networks", *Physical Review E, 76(3).*

[32] G. Palla, I. deényi, I. Farkas and T. Vicsek (2005) "Uncovering the overlapping community structure of complex networks in nature and society", *Nature, 435(7043), 814-818.*

[33] A. Clauset, M.E.J. Newman and C. Moore (2004) "Finding community structure in very large networks", *Physical Review E, 70(6).*

[34] M. Rosvall and C.T. Bergstrom (2008) "Maps of random walks on complex networks reveal community structure", *Proceeding of the National Academy of Sciences, 105(4), 1118-1123.*

[35] Q. Gui, R. Deng, P. Xue and X. Cheng (2018) "A community discovery algorithm based on boundary nodes and label propagation", *Pattern Recognition Letters, 109, 103-109.*