22. FOUR COMMUNITY DETECTION ALGORITHMS FOR DIRECT AND INDIRECT GRAPHS

by *Giulia Contu*[*], *Luca Frigau* and *Maurizio Romano*

Abstract

In the years, the researches have posed their attention on the study of networks analysing the nature, the composition, the relationship developed inside them, their construction and growth. Recently, particular attention has been posed on the analysis of complex networks and on the community detection analysis. The complex network has been defined as a network […] open, value-laden, directed, multilevel, multicomponent, reconfigurable systems of systems, and placed within unstable and changing environments (Boccaletti et al., 2014, p. 6). In this complexity, particular interesting is the identification of communities inside the network. It is possible to define the communities as subgroups of nodes with a density of internal connections larger than the density of external links. The aim of the community detection analysis is to identify the community structure inside the network in order to define the modular decomposition of the network.

In literature many community detection algorithms are identified. To evaluate the capacity of these algorithms to identify the community, the algorithms have been applied on two different networks: the Zachary's karate club network and Friendship network of a UK university faculty. The first is an indirect graph, as it is characterised to have edges that are not directed. On the contrary, the second is a direct graph, as it is composed by directed edges.

The first findings evidence how each algorithm identifies specific communities inside the networks. These communities are composed in the most of cases by different nodes. Only in few cases, similarities have been detected. Moreover, it has been identified some problematic in the analysis of direct graph.

Keywords: Network, Complex network, Community detection algorithms, Zackary networks.

[*] Corresponding author, giulia.contu@unica.it.

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

Quattro algoritmi di community detection per reti dirette e indirette

Negli anni, i ricercatori hanno posto la propria attenzione nello studio dei network analizzando in particolare la loro natura, la loro composizione, le relazioni sviluppate al loro interno, la loro nascita e il loro sviluppo. Recentemente i ricercatori hanno focalizzato i loro studi su due principali tematiche: le reti complesse e l'identificazione delle comunità all'interno dei network. Le reti complesse sono stata definite come sistemi […] aperti, carichi di valore, diretti, multilivello, multicomponente, riconfigurabili e collocati in ambienti instabili e mutevoli (Boccaletti et al., 2014, p. 6). In questa complessità, è particolarmente interessante l'identificazione dei gruppi, o comunità, all'interno delle reti. Le comunità possono essere definite come sottogruppi di nodi che presentano una connessione interna forte e superiore alla connessione creata con gli altri nodi della rete. L'obiettivo di identificare tali comunità è quella di comprendere la struttura della rete e le relazioni create tra i nodi di una comunità, e fra i nodi delle diverse comunità. In letteratura sono stati identificati molti algoritmi capaci di raggiungere tale obiettivo. Si è deciso di applicare alcuni degli algoritmi sviluppati per comprendere il loro funzionamento. Si è scelto come rete di riferimento il network creato dai dipendenti di una facoltà inglese e quello generato dal club di Karate Zachary. La prima rete è definita diretta, le informazioni si muovono seguendo una direzione specifica da un nodo ad un altro, al contrario la seconda è definita indiretta, le informazioni si muovono senza seguire una direzione specifica.

I primi risultati dimostrano che ciascun algoritmo identifica una specifica struttura all'interno della rete. Ogni struttura si caratterizza per un diverso numero di comunità e per la presenza di relazioni specifiche fra i membri della rete. Tuttavia, in alcuni casi è possibile identificare somiglianze nei risultati e riconoscere l'esistenza di relazioni forti fra specifici membri della comunità. Inoltre, sono state rilevate alcune problematiche legate all'analisi delle reti dirette.

Parole-chiave: reti, reti complesse, algoritmi di community detection, rete Zackary

1. Introduction

The study of network originated from the necessity to represent phenomena based on single units linked between them thorough specific relationships. In literature, the network is described as a collection of objects in which some pairs of these objects are connected by links (Easley et al., 2010; Kolascyk, 2013); moreover, as an ensemble of nodes and edges (Agarwal et al., 2008; Leicht et al., 2008; Porter et al., 2009; Borgatti et al., 2011); finally, as a graph composed by points joined together by connections (Berg et al., 2002; Easley et al., 2010).

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

The study of networks started in the 1700s. The first study has been realized by Leonhard Euler in 1736 (Euler, 1741). In years, researchers have analysed networks taking into account their nature, composition, within relationships, construction and development. Recently, the development of always more large network, with high level of interaction and high inner complexity have determined the necessity for researchers to focus on the analysis of complex networks and networks communities.

The complex networks have been defined as «[…] graph with large and complex size» (Chen et al., 2007, p. 1317). Additionally, they have been described as «[…] open, value-laden, directed, multilevel, multicomponent, reconfigurable systems of systems, and placed within unstable and changing environments» (Boccaletti et al., 2014, p. 6). They are able «to change, evolve, transform through inner and outer dynamic interactions affecting the subsystems and components at both local and global scale» (Barabási et al., 1999; Schaeffer, 2007).

In a complex network, it is possible to identify communities composed by subgroups of nodes with a density of inner connections larger than the density of outer links (De Meo et al., 2013; Radicchi, 2014; Hric et al., 2014). The aim of the community detection analysis is to identify the community structure within the network to define its modular decomposition.

Several community detection methods have been developed. Many of these are developed referring to tools and techniques from different disciplines such as physics, biology, applied mathematics, computer and social sciences (Radicchi et al., 2004; Clementi et al., 2015; Sun et al., 2014).

The main aim of this paper is to review the literature related to community detection algorithms and to apply some of these algorithms on a real network. Three sections compose the remaining part of the paper. The next two sections are focused on the review of the literature. The last part is focused on the application of the community detection algorithms on two well-know networks: Zachary's karate club and Friendship of a UK university faculty. The last section summarizes some concluding remarks.

2. Community detection algorithms

Following the literature review realized by Fortunato (2010), the methods of community detection can be classified in *traditional methods, divisive algorithms, modularity-based methods, dynamic methods and other methods*.

Traditional methods can be casted in graph partitioning, hierarchical clustering and partitional clustering (Xu et al., 2007; Wang et al., 2015). In graph

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

partitioning, the graph is divided into groups with specific properties. The number of parts is generally specified *a priori* (Bedi et al., 2016). The aim of this method is to divide the vertices in groups of predefined size, such that the number of edges lying between the groups is minimal (Newman, 2010). In hierarchical clustering, different groups are identified choosing a similarity measure capable of computing the similarity for each pair of vertices (Newman, 2004). This method can be applied considering two different approaches: the agglomerative approach and the divisive one. In partitional clustering, the number of clusters is pre-assigned and the point is attributed to each cluster considering the distance in the metric space. Its aim is either to maximize or to minimize a given cost function based on distances between points or from points to centroids. The method starts considering the presence of centroids and it attributes each vertex to the nearest centroid.

Divisive algorithms compose the second group of methods. These methods identify communities through the detection of the edges that connect vertices of different communities and removing them (Paliouras et al., 2015). The aim is to disconnect the clusters from each other. In this method, it is fundamental to choose the edges and to split the network in communities, which are constructed removing edges progressively from the original graph (Girvan et al., 2002; Clauset et al., 2004; Hoffman et al., 2017). The most famous method has been theorized by Girvan and Newman (2002). It uses a new measure called *betweenness* that identifies the frequency of the participation of edges to a process. The model of Girvan and Newman (2002) begins with the computation of the betweenness and proceeds removing the edges with the highest betweenness. Initially, the nodes are considered in a single cluster and next they are split in components. Later, the betweenness for all edges affected by the removal is recalculated. Finally, the second phase is repeated until no edges remain. The main idea of the model is that edges that run between communities have higher betweenness values than those that lie within communities.

The third group of community detection methods is based on a specific measure: *the modularity*. It is a measure proposed by Newman and Girvan (2004) to estimate the goodness of the modules obtained from the community detection. The modularity is computed as $Q = \sum_i e_{ii} - a_i^2$, where e_{ii} is the fraction of the edges that connects vertices in community i, $a_i = \sum_j e_{ij}$ is the fraction of edges that connects to vertices in community i, and e_{ij} indicates the fraction of the edges connecting vertices in two different communities i and j (Clauset et al., 2004; Newman et al., 2004). The modularity defines the quality of a specific community division in a network. For this

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

reason, the best partition has a high modularity. Many different kinds of algorithms based on modulatory measures have been theorized.

The dynamic algorithms compose another class of methods. It is possible to include in this group different kinds of methods based on different concepts, as for instance *random walk* (Yen et al., 2009), *synchronization* (Boccaletti et al., 2007), and *label propagation* (Raghavan et al., 2007).

Finally, it is possible to find in literature some models that cannot be included in the previous categories as for instance the *L-shell* method proposed by Bagrow and Bollt (2008), and the models related to the benchmark proposed for instance by Girvan at al. (2002) and Lancichinetti et al. (2009). These methods are included in the category *other methods*.

3. Other community detection algorithms

In the following, the attention is focused on four algorithms. These are: *Louvain* (Blondel et al., 2008), *Label Propagation* (Raghavan et al., 2007), *Walktrap* (Pons and Latapy, 2005) and *Edge Betweenness* (Newman and Girvan, 2004).

Specifically, *Louvain* has been developed by Blondel et al. (2008). It is based on *modularity* and on the method proposed by Clauset et al. (2004). The algorithm works in two phases. Firstly, it assigns each node of the network to different communities. In this initial partition, each node corresponds to one community. Then, the algorithm considers a node *i* and a node *j* and evaluates if it is possible to record an increase in modularity merging the two points in a same community. This step is repeated for all nodes and each node is located in the community that generates the highest increase in modularity. This process is repeated until all nodes are located and until no further improvement can be achieved.

Label propagation community detection algorithm is based on the concept of label propagation and on the model theorized by Raghavan et al. (2007). The algorithm focuses on the hypothesis that a node *x* has as neighbours x_1, x_2, …, x_k, and that each neighbour carries a label denoting the community to which it belongs to. Then, the node *x* defines its community based on the labels of its neighbours. Raghavan et al. (2007) have hypothesized that *each node chooses to join the community to which the maximum number of its neighbours belong to, with ties broken uniformly randomly* (Raghavan et al., 2007, p.4). At the starting point, every node presents a unique label and the labels propagate through the network. After this propagation, the densely connected group of nodes share the same label. Nodes that have the

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

same labels are grouped together as one community. Some labels dominate, other disappear within the network. Using this process the communities are defined as groups of vertices having identical labels at convergence. The algorithm converges when a global consensus among groups is reached and the communities are identified.

Walktrap is based on the distance between vertices and on a hierarchical clustering algorithm. Pons and Latapy (2005) have introduced a new measure of distance *r*, between the vertices, which is able to capture the community structure on the graph. The distance between two communities is found taking into account the random walk. Specifically, the starting vertex is chosen randomly and uniformly among the vertices of the community. Moreover, the distance is used in the search of the communities in combination with a hierarchical cluster algorithm. The algorithm is based on an agglomerative approach and uses Ward's method.

Edge Betweenness community detection algorithm is based on the concept of edge betweenness, i.e.: the ratio of the number of the shortest paths going through the edge to the shortest paths of all node pairs (Newman & Girvan, 2004). The algorithm involves computation of the edge betweenness of the graph, obtained by removing the edge with the highest edge betweenness score and recalculating edge betweenness of the edges and repeating this refinement until furhter improvements are not possible.

To sum up, the above-mentioned algorithms are based on different measures used to identify the communities inside the network. Generally, they are able to capture the strength of the relationship created inside the network and to join the more tied nodes. It has been hypothesized that the algorithms allows us to identify a specific community structure inside a network and to identify similarities and differences within the communities.

4. An application: Zackary network and Friendship network of a UK university faculty

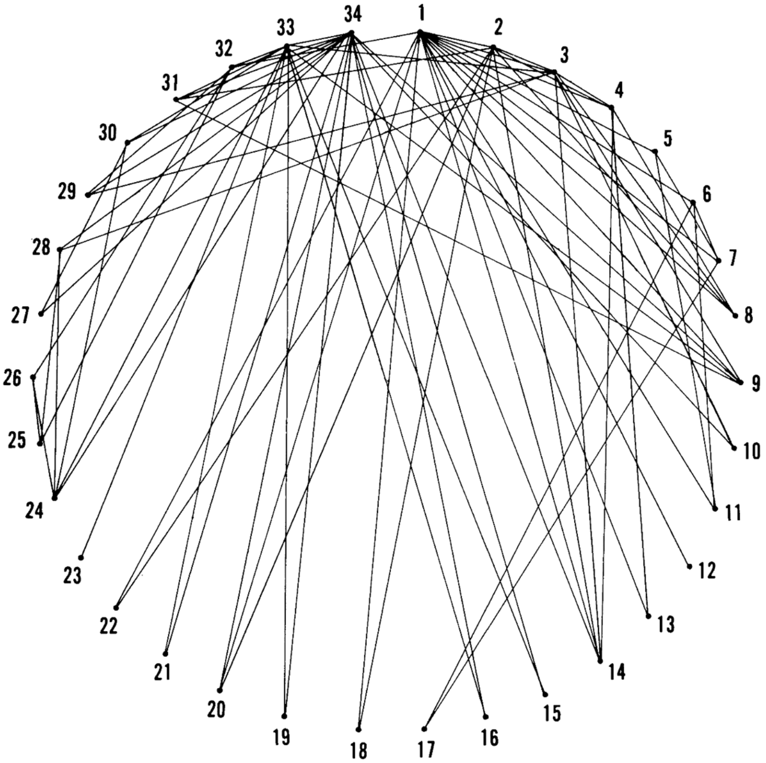
The above-mentioned algorithms have been performed on two networks. We choose to use one indirect and one direct graph[1] in order to identify similarities and differences in the application of the algorithms.

[1] The graph can be classified as direct and indirect. The direct graph is composed by directed edges and in this case is called digraph. In direct graph, it is important to take into account the direction of the link. Instead, if the edges are not directed in the graph, it is identified an undirected graph.

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

The first network is the undirected Zachary's karate club network. It is a well-known social network composed of 34 members of a karate club. Wayne W. Zachary has studied this network for a period of three years from 1970 to 1972. It describes the relationships developed between pairs of members outside the club, as shown Figure 1. Each member of the network is identified through a number. Two members, the number 1 and the number 34, have a specific role: they are respectively the instructor and the president of the Karate club.

Figure 1 – The social relationships among the 34 individuals in the karate club (Zachary, 1977)



The different community detection algorithms applied on Zachary's network have allowed identifying four different community structures within the network. Each structure presents a specific number of communities and the Zachary's members are joined together in different way. Specifically, the number of communities changes from three to five, as show in Figure 2. The

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

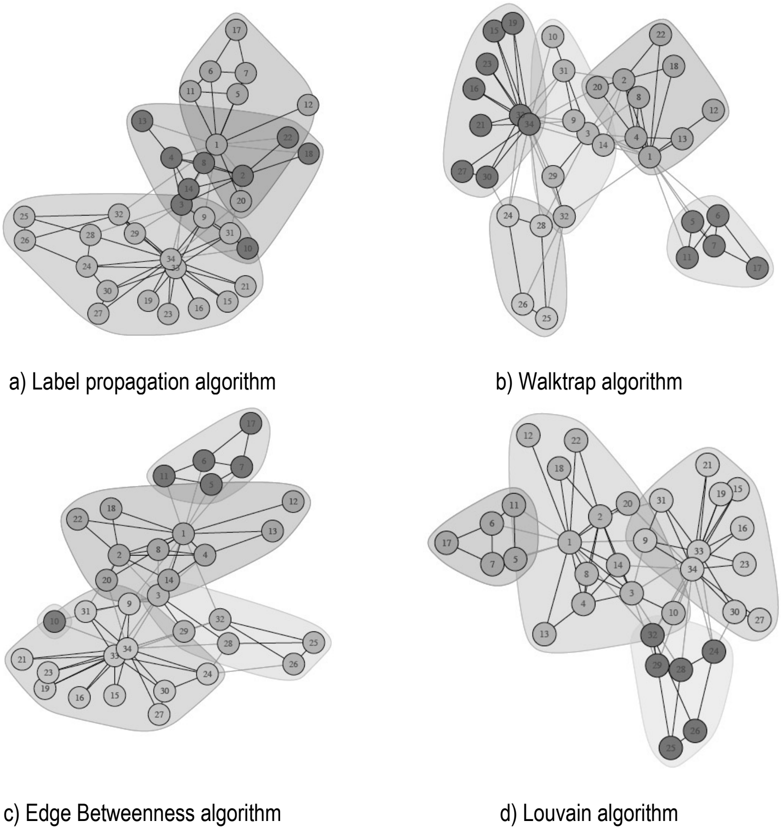
lowest number of communities is identified through the Label propagation algorithm, the highest with the Walktrap and the Edge Betweenness algorithms. Moreover, analysing the four communities structures, it is possible to evidence how members have both a specific and a central role inside the communities. This role is determined by elevate number of connection both within the community and between communities, as show in Figure 2. The same members play the central role in the most of identified communities through the algorithms. They are the members identified with the number 1, 34 and 5, as shown in Table 1. Additionally, it is interesting to highlight how the community composed by the member 5, 6, 7, 11 e 17, is included in all network structures generated by the different algorithms. This means that the five Zachary' members are strongly linked between them and this solid relationship is recognised by all algorithms. Moreover, some members are always located in the same community because they are linked together through a strong relationship that has been recognised by all algorithms.

Table 1 – The community structure of Zachary's karate club network generated by the community detection algorithms

Algorithms	**Community 1**	**Community 2**	**Community 3**	**Community 4**	**Community 5**
Louvain	5, 6, 7, 11, 17	1, 2, 3, 4, 8, 10, 12, 13, 14, 18, 20, 22	24, 25, 26, 28, 29, 32	9, 15, 16, 19, 21, 23, 27, 30, 31, 33, 34	
Central role Louvain	5	1	32	34	
Walktrap	1, 2, 4, 8, 12, 13, 18, 20, 22	3, 9, 10, 14, 29, 31, 32	15, 16, 19, 21, 23, 27, 30, 33, 34	24, 25, 26, 28	5, 6, 7, 11, 17
Central role Walktrap	1	3	34	24	5
Label propagation	1, 2, 3, 4, 8, 12, 13, 14,18, 20, 22	5, 6, 7, 11, 17	9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34		
Central role Label propagation	1	5	34		
Edge Betweenness	1, 2, 4, 8, 12, 13, 14, 18, 20, 22	3, 25, 26, 28, 29, 32	5, 6, 7, 11, 17	9, 15, 16, 19, 21, 23, 24, 27, 30, 31, 33, 34	10
Central role Edge Betweenness	1	3	5	34	

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

Figure 2 – Community detection algorithms applied on Zachary's karate club network



a) Label propagation algorithm

b) Walktrap algorithm

c) Edge Betweenness algorithm

d) Louvain algorithm

The second network is a direct graph called *UKfaculty*. It is a social network composed of *81 members. It has been proposed by* Nepusz et al. (2008). It describes the personal friendship network of academic staff of a UK university.

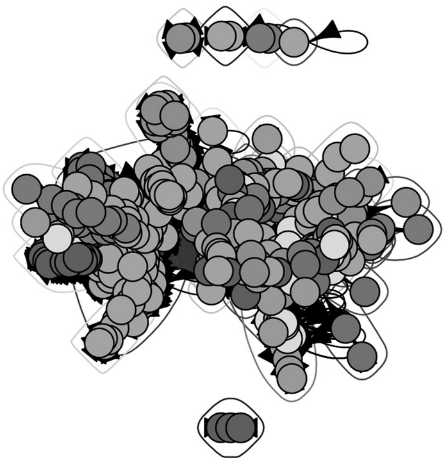
Some problems have been identified in the application of community detection algorithms. Not all algorithms are able to analyse the direct graphs. Specifically, Label propagation algorithm has been developed just to analyse the indirect graph. It involves joining together the nodes that share the same label. The propagation is possible only if there is no predefined direction in the connection. In the same way, the Louvain algorithm cannot be applied on direct graphs. In fact, it allows computing the modularity for indirect graph only.

Only the Edge Betweenness and the Walktrap algorithms are able to identify communities inside the network, as shown in Figure 3. However, it is

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

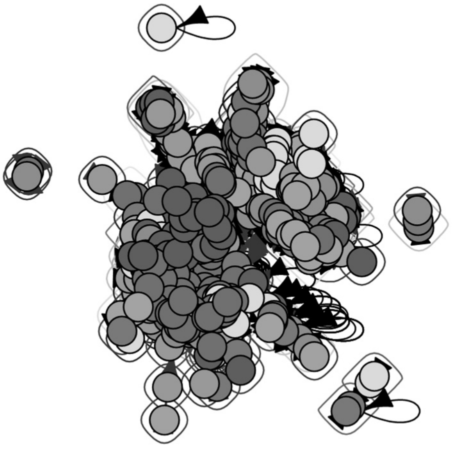
necessary to evidence how the Walktrap algorithm ignores the direction of the edges and transforms the directed graphs in indirect graphs. In this way, the specific characteristics of the network are ignored and the network is analysed as an indirect graph. In the same way, the transformation from direct to indirect graph can be considered as a possible solution for the application of the community detection algorithms. However, this solution is not considered correct because it ignores the real nature and structure of the network.

Figure 3 – Community detection algorithms applied on Friendship network of a UK university faculty



a) Edge Betweenness algorithm

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701



b) Walktrap algorithm

5. Concluding remarks

To sum up, it is possible to state that the four chosen algorithms are useful to detect communities only for indirect graphs. In fact, the algorithms have identified specific community structures inside Zachary's karate club network. Moreover, each structure is characterised by a different number of groups and by the union of specific nodes. However, it is possible to identify some similarities. For instance, some members are always located in the same community because they are linked together through a strong relationship that has recognised by all algorithms. Finally, it is possible to recognise a specific and central role for some members of the communities, as for instance the instructor and the president that present an elevate number of connections inside the network.

On the contrary, only one of the four analysed algorithms has been able to identify a community structure inside the friendship network of a faculty

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

in a UK university. It is possible to state that only the Edge Betweenness algorithm is able to analyse a direct graph, because the Walktrap algorithm ignores the direction of the edges and transform the direct graph in an indirect graph.

References

Agarwal G., Kempe D. (2008). Modularity-maximizing graph communities via mathematical programming. *The European Phyical Journal B*, 66(3): 409-418.

Barabási A.L., Albert R., Jeong H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1): 173-187.

Bedi P., Sharma C. (2016). Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3): 115-135.

Berg J., Lässig M. (2002). Correlated random networks. *Physical review letters*, 89(22), 228701: 1-4.

Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 10, 10008: 1-12.

Boccaletti S., Bianconi G., Criado R., Del Genio C. I., Gómez-Gardenes J., Romance M., Sendina-Nadal I., Wang Z., Zanin M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, 544(1): 1-157.

Borgatti S.P., Halgin D.S. (2011). On network theory. *Organization science*, 22(5): 1168-1181.

Chen T., Liu X., Lu. W. (2007). Pinning complex networks by a single controller. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54(6): 1317-1326.

Clauset A., Newman M.E.J., Moore C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6), 066111: 1-6.

Clementi A., Di Ianni M., Gambosi G., Natale E., Silvestri R. (2015). Distributed community detection in dynamic graphs. *Theoretical Computer Science*, 584: 19-41.

De Meo P., Ferrara E., Fiumara G., Provetti A. (2013). Enhancing community detection using a network weighting strategy. *Information Sciences*, 222: 648-668.

Easley D., Kleinberg J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Euler L. (1741). *Solutio problematis ad geometriam situs pertinentis. Commentarii academiae scientiarum Petropolitanae*: 128-140.

Fortunato S. (2010). Community detection in graphs. *Physics reports*, 486(3): 75-174.

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

Hoffman M., Steinley D., Gates K.M., Prinstein M.J., Brusco M.J. (2017). Detecting clusters/communities in social networks. *Multivariate behavioural research*: 1-17.

Hric D., Darst R.K., Fortunato S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E*, 90(6), 062805: 1-25.

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701

Kolascyk E.D. (2013). *Statistical analysis of network data. SAMSI program on Complex networks*. Boston University.

Lancichinetti A., Fortunato S. (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1), 016118: 1-9.

Leicht E.A., Newman M.E.J. (2008). Community structure in directed networks. *Physical Review Letters*, 100(11), 118703.

Nepusz T., Petroczi A., Negyessy L., Bazso F. (2008). Fuzzy communities and the concept of bridgeless in complex networks. *Physical Review E*, 77, 016107: 1-13.

Newman M.E.J. (2010). *Networks: an introduction*. Oxford University Press.

Newman M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), 066133: 1-5.

Newman M.E.J., Girvan M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 026113: 1-15.

Paliouras G., Papadopoulos S., Vogiatzis D. (2015). Discovery of complex user communities. In *User Community Discovery*. Springer, pp. 1-22.

Pons P., Latapy M. (2005). Computing communities in large networks using random walks. In *International symposium on computer and information sciences*. Springer, pp. 284-293.

Porter M.A., Onnela J.P., Mucha P.J. (2009). Communities in networks. *Notices of the AMS*, 56(9): 1082-1097.

Radicchi F. (2014). A paradox in community detection. *Europhysics Letters* (EPL), 106(3), 38001: 1-5.

Radicchi F., Castellano C., Cecconi F., Loreto V., Parisi D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9): 2658-2663.

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: 1-12.

Schaeffer S.E. (2007). Graph clustering. *Computer Science Review*, 1(1): 27-64.

Sun P.G., Sun X. (2017). Complete graph model for community detection. *Physica A: Statistical Mechanics and its Applications*, 471: 88-97.

Wang M., Wang C., Yu J. X., Zhang J. (2015). Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework. *Proceedings of the VLDB Endowment*, 8(10): 998-1009.

Xu G., Tsoka S., Papageorgiou L.G. (2007). Finding community structures in complex networks using mixed integer optimisation. *The European Physical Journal B-Condensed Matter and Complex Systems*, 60(2): 231-239.

Yen L., Fouss F., Decaestecker C., Francq P., Saerens M. (2009). Graph nodes clustering with the sigmoid commute-time kernel: A comparative study. *Data &amp; Knowledge Engineering*, 68(3): 338-361.

Zachary W.W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4): 452-47.

Copyright © 2019 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788891798701