COMMUNITIES IN COMPLEX NETWORKS: IDENTIFICATION AT DIFFERENT LEVELS

Alex Arenas, Jordi Duch and **Sergi Gómez**
Departament Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Spain

Leon Danon
Mathematics Institute, University of Warwick, Great Britain

Albert Díaz-Guilera
Departament Física Fonamental, Universitat de Barcelona, Spain

Keywords: Communities, hierarchies, overlap, dynamics

Contents

Acknowledgements
Glossary
Bibliography
Biographical Sketches

Summary

We present here and compare the most common approaches to community structure identification in terms of sensitivity and computational cost. The work is intended as an introduction as well as a proposal for a standard benchmark test of community detection methods.

1. Introduction

The analysis of complex networks has received a vast amount of attention from the scientific community during the last decade. Statistical physicists in particular have become interested in the study of networks describing the topologies of a wide variety of systems, from biological technological or social networks. Although several questions have been addressed (see the review paper by Costa et al. for a complete set of measurements), many important ones still resist complete resolution. One such problem is the analysis of modular structure found in many networks. Distinct modules or communities within networks can loosely be defined as subsets of nodes which are more densely linked, when compared to the rest of the network. Such communities, as usually called in social sciences, have been observed, using some of the methods we shall go on to describe, in many different contexts, including biological networks, economic networks and most notably social networks. As a result, the problem of identification of communities has been the focus of many recent efforts. As a concrete example we show in Figure 1 the network representing the Spanish research community of Statistical and Nonlinear Physicists (FISES, http://www.fises.es).

We consider two scientists linked if they have co-authored a panel contribution to any of the conferences. To be able to consider the historical structure of this network we ``accumulate" the network over all the conferences, that is, once a link is created, it remains, even if the authors never collaborated again. The final network (accumulated over all the years) is comprised of 784 nodes with 655 (84%) of those belonging to the giant component. Green nodes denote the member of the scientific committees.

Nodes belonging to the same community are more than likely to have other properties in common and hence community detection in large networks is potentially very useful for instance when trying to understand dynamical properties. In the world wide web, community analysis has uncovered thematic clusters. In biochemical or neural networks, communities may be functional groups, and separating the network into such groups could simplify the functional analysis considerably.

The problem of community detection has been the subject of study in various disciplines. A simpler version of this problem, the graph bi-partitioning problem (GBP) has been the topic of study in the realm of computer science for decades. Here one looks to separate the graph into two equal-size communities, which are connected with the

minimum number of links. This is indeed an NP complete problem; however several methods have been proposed to reduce the complexity of the task. In real networks one cannot assume how many communities there are, but in general it is more than two. This makes the process much more costly.

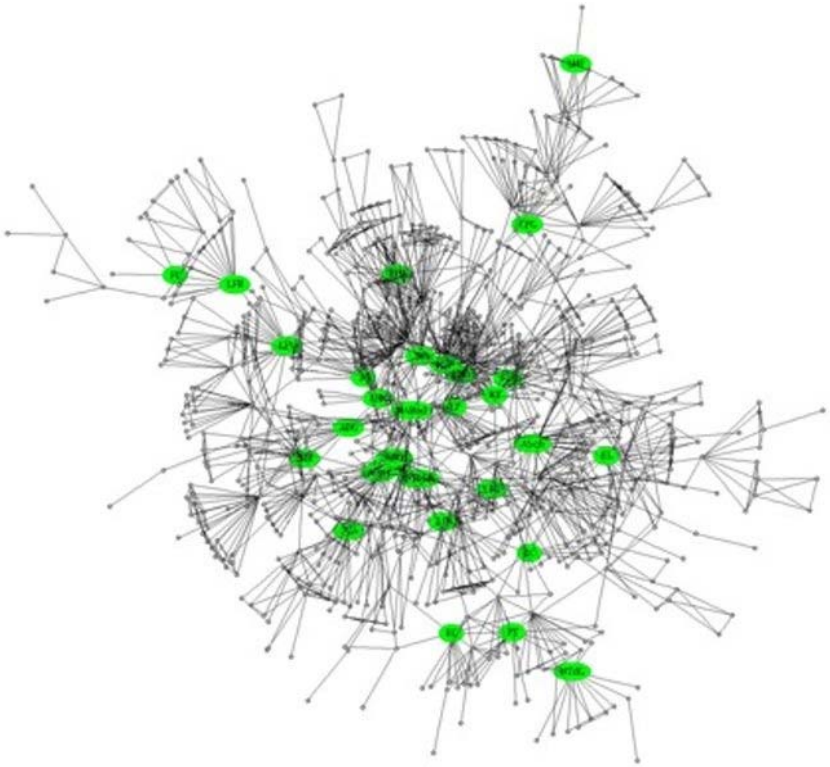

Figure 1. FisEs network. Network of scientists that contributed to the Statistical Physics (Física Estadística) conferences in Spain.

Furthermore communities can be organized in hierarchies, meaning that different organizational levels can be simultaneously important and the question to the best partition has not a single answer. This hierarchical organization strongly affects the dynamical properties of networks. Another additional issue is that sometimes there is not a clear separation among communities and they present a certain degree of overlapping.

In this chapter we would like to present the recent advances made in the field of community identification in networks in a clear and simple fashion. To this end, the sections are organized as follows. In the next section we describe some ways to define communities in a network context. Following this, we present a method to evaluate a particular partition of a network. Then, we go on to describe the various recent methods starting with link removal methods, going on to agglomerative methods, followed by methods optimizing modularity and finally "other" methods. Some of the methods presented do not necessarily fit into just one of these classifications, and there may be some overlap. We finally introduce different structural organizations in networks and dynamical applications of modular networks.

2. Definitions of Communities

There is not a unique definition of what a community is, instead the idea of communities is different and has been evolving depending on the field that defines it. The first definitions of community come from the field of social networks, where the communities are studied and understood according to the effect that an individual player has on the network and vice versa. Some of these ideas have been used and developed by some of the methods we present below, while new approaches have also been adopted from other fields such as physics or mathematics.

The different definitions of what is a community are all based in the concept of a subgraph, that is, groups of nodes and all the connections between them. The definitions can be classified into two main conceptual categories, those who use self-referral information and those based on comparative definitions.

Self referring definitions only use information of the structure of the network to decide what groups of nodes can be considered as a community. The most restricting and simple community structure is a clique, defined as a subgraph that is fully connected (i.e. it has all the possible edges between its nodes). Since this constraint is rarely fulfilled in real sparse networks, there are other approaches that relax it, such as n-cliques, n-clans and n-clubs. Self-referring definitions, while useful in characterizing communities, which are already known, are not the best choice while trying to find them since the methods to find the cliques in a network is very costly.

A second type of definitions use topological information to compare if a group of nodes is a community or not, for instance, counting how many links have the nodes of the subgraph inside of it and how many links have them with nodes outside the subgraph. The strong definition of community requires that all the nodes of a community must have a larger number of links to members of the same community than to members of other communities. A lighter version of this definition is the weak definition of community proposed by Radicchi et al., where it is required that the sum of links inside the community is larger than the total number of links to the outside. This definition and some small variations of it is the most used in the majority of the methods that we will present later, since comparing the internal structure of a community to the external structure gives rise to a measure of how good a particular partition is.

3. Evaluating Community Identification

Once a partition of the network into communities has been identified, the problem turns on to evaluate how good is the partition. Girvan and Newman proposed a simple approach, based on the intuitive idea of lack of community structure in random networks. Consider an arbitrary partition of a given network into N_c communities. We can define a $N_c \times N_c$ size matrix $\mathbf{e}$ where the elements e_{ij} represent the fraction of total links starting at a node in partition i and ending at a node in partition j. Then, the sum of any row of $\mathbf{e}$, $a_i = \sum_j e_{ij}$ corresponds to the fraction of links connected to i. If there is no community structure in the network the expected value of the fraction of

links within partitions can be estimated. It is simply the probability that a link begins at a node in i, a_i, multiplied by the fraction of links that end at a node in i, a_i. Then the expected number of intra-community links is just $a_i a_i$. We also know that the *real* fraction of links exclusively within a partition is e_{ii}. Comparing the two and summing over all the partitions in the graph we get

$$Q = \sum_{i=1}^{c} (e_{ii} - a_i^2) . \qquad (1)$$

This is a measure known as *modularity*. As an example, we can consider a network comprised of two disconnected components. If we then have two partitions, corresponding exactly to the two components, modularity will have a value of 1. For particularly "bad" partitions, for example, when all the nodes are in a community of their own, the value of modularity can take negative values.

It is tempting to think that random, Erdos-Renyi networks have little or no community structure. However, as Guimerà *et al.* showed, this in general is not the case. In fact, it is possible to find a partition which not only has a nonzero value of modularity for random networks of finite size, but that this value is quite high. For example a network of 128 nodes and 1024 links has a maximum modularity of 0.208. This suggests that community structure appears in random networks due to fluctuations.

From here on we will look at different methods of community identification presented recently. First we consider methods based on link removal.

4. Link Removal Methods

Divisive methods extract the partition into communities of a network by removing some (or all) of its links until the network is no longer connected or we have a division into communities that meets certain requirements. However, to be able to obtain useful results we need to remove the appropriate links, otherwise the communities will be meaningful. Several methods have been proposed to identify the links that we should remove, which we will revise in this section.

4.1. Shortest Path Centrality

One of the first divisive methods presented in uses the idea of centrality, a measure of how central the node or link is in the network, to decide which links need to be removed. The algorithm uses a particular type of centrality, shortest path centrality, which measures the number of shortest paths between pairs of nodes that pass through a certain node or link. The links with the highest centrality usually act as a bridge between the communities, so if we remove them we can split the network into densely connected communities.

The method works recursively eliminating all the links of the network, and stops when there are no more links and all the nodes are isolated. Every time a link is removed, all

the centralities are recalculated, otherwise we will obtain an erroneous community detection. This part of the algorithm is the one that requires most computer power and, for a network of size n with m links, using the fastest methods developed independently by Newman and Brandes the speed of calculating all link betweenness-es in one step still remains of $O(m^2 n)$ for unweighted networks. This limits the size of the graph that we can process in a reasonable time to a maximum of around 10000 nodes. Figure 2 shows the application of this algorithm to the network depicted in Figure 1.

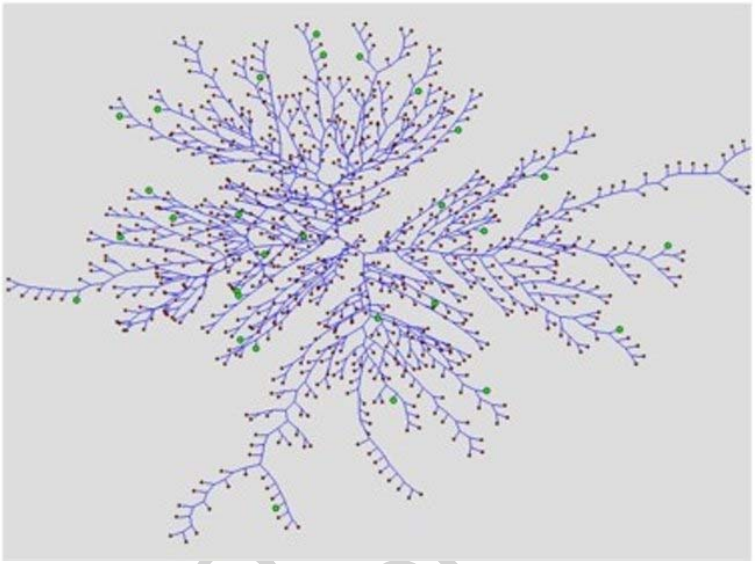

Figure 2. Binary tree showing the result of applying the Girvan-Newman algorithm and our visualization technique to the network of coauthors in FisEs.

Each branch corresponds to a real community and the tips of the branches correspond to the people that have played a major role in the different research groups. One can identify here that the members of the scientific committees over the years have indeed played an important role in the development of the community and that they are precisely quite central nodes in the respective local communities.

4.2. Extensions of the Shortest Path Centrality

The same authors of the previous method have also presented two alternative methods to detect community structure by betweenness centrality by calculating this value using two alternative approaches. However, although they are conceptually interesting, both approaches require higher computation than the previous method, and they do not improve the accuracy of it.

The first approach considers the network as a circuit, where links are assigned a unit resistance and we select two nodes that we define as unit voltage source and sink. Using Kirchoff's laws we can calculate the current flow between these two nodes. Adding the flows we will obtain a mesure similar to the centrality, where those links with the

lowest resistance (shortest path) carry the most current and, therefore, are the most central. The second approach uses random walks to determine the betweenness centrality of the links. The network is used as a substrate for signals that perform a random walk between pairs of nodes. The link betweenness in this case is simply the rate of flow of random walkers through a particular link summed over all pairs of vertices.

4.3. Information Centrality

Another divisive algorithm available uses the network efficiency measure proposed by Latora and Marchiori. This measure quantifies how efficient is a network in the context of information exchange. If we remove links of the network, its efficiency decreases a certain amount of information centrality. This method, presented by Fortunato et al., is based on the idea that if we remove the links that act as bridges between communities we should observe the largest drops in network efficiency. from this premise, the method operates similarly to the shortest path centrality method, removing recursively all the links and recalculating the efficiency of all the links at every step. The process is slower than the GN running at $O(n^4)$, but instead the accuracy obtained in the detection is better when the communities to be found are more diffuse.

4.4. Link Clustering

Another approach uses the idea that linked nodes belonging to the same community should have a high clustering coefficient, that is, they share larger number of common neighbors. Based on this idea, the algorithm of Radicchi et al. postulates that the proportion of possible number of loops that go through internal links should be much larger than the proportion of loops for links pointing to outside of the community. The algorithm also works recursively as the previous ones, but in this case by recalculating the *link-clustering coefficient*, which measures the number of loops of a certain length that pass through each link. Longer loops require more computer resources but provide more accurate results.

This algorithm provides a way to stop the detection process when a certain condition is fulfilled, instead of decomposing the whole network until all the nodes are separated. It is also faster than the previous ones, since to compute the *link-clustering coefficient* we only need local information. However, it is not very useful with networks with a very low clustering coefficient, such as trees, sparse graphs or disassortative networks, where we do not have the necessary loops to compute the *link-clustering coefficient*.

5. Agglomerative Methods

Another approach to identify the communities of a network is to start from all the nodes being in separate communities, and some strategy to join or agglomerate them in larger groups. Here we present some of these methods and their grouping algorithms.

5.1. Hierarchical Clustering

Hierarchical clustering has been used traditionally in social networks analysis to extract

the communities of the network. The idea of this method is based on the measurement of the similarity between the elements of the nodes according to some property. Starting from an empty network, the method selects those node (or groups of nodes) that have the highest similarity and joins them. This process is again repeated recursively until all the links are added or when we meet a certain condition. The method is very fast and it can work almost in linear time, being able to analyze networks that cannot be processed otherwise. However, the results are highly dependent on the similarity metric that is used to detect the communities.

5.2. L-Shell Method

A second approach focuses on identifying the community around one node of the network by agglomerating its neighbors until a condition is fulfilled. In particular, the algorithm consists on constructing a *L*-shell around one node, where a *L*-shell is a subset of the nodes with a maximum distance of the shortest path to the node origin is less or equal to L. The algorithm starts from the origin and adds more nodes by increasing the distance L until the emerging degree (number of links to nodes outside the *L*-shell) is lower than a cut-off value, and then it is stopped. Those nodes that fall inside the *L*-shell are grouped within one community.

This algorithm is particularly interesting when one is more interested in finding a single community and not in detecting the entire community structure. If we want to make the algorithm global, the authors suggest that we should repeat the process for each node, and then perform a statistical analysis of the results to detect the communities. Since the method uses local information, it is one of the fastest available.

-
-
-

Bibliography

A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, C. Zhou (2008) Synchronization in complex networks. *Physics Reports* 469:93. [Review paper on synchronization in complex networks]

A. Arenas, A. Dáz-Guilera, and C. J. Pérez-Vicente. (2006) Synchronization Reveals Topological Scales in Complex Networks. *Phys. Rev. Lett.*, 96:114102 [First attempt in detecting hierarchical structures in complex networks based on dynamical methods]

A. Arenas, A. Fernández, and S. Gómez. (2008) Analysis of the structure of complex networks at different resolution levels. *New J. Phys.*, 10:053039+. [Multi-resolution at multiple scales]

Albert Laszlo Barabási and Reka Albert. (2002) Statistical mechanics of complex networks. *Review of Modern Physics*, 74:47–97. [The first review on complex networks].

Alex Arenas, Leon Danon, Albert Diaz-Guilera, Pablo M Gleiser, and Roger Guimerà. (2004) Community analysis in social networks. *European Physical Journal B*, 38:373–380 [Community analysis

in some examples of social networks]

C. Bron and J. Kerbosch. (1973) Finding all cliques in an undirected graph. *Communications of the ACM*, pages 575–577 [A paper on the determination of communities in a simple case].

E. N. Sawardecker, M. Sales-Pardo, and Amaral. (2009) Detection of node group membership in networks with group overlap. *The European Physical Journal B - Condensed Matter and Complex Systems*, 67:227 [Overlapping communities]

E. Oh, K. Rho, H. Hong, and B. Kahng. (2005) Modular synchronization in complex networks. *Phys. Rev. E*, 72:047101. [Dynamical properties of structured networks]

Erzsébet Ravasz and Albert-László Barabási. (2003) Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112. [Hierarchical networks]

F. Wu and B.A. Huberman. (2004) Finding communities in linear time: a physics approach. *Eurpean Physics Journal B*, 38:331–338. [Method based on dynamical properties]

Filippo Radicchi, Claudio Castellano, Frederico Cecconi, Vittorio Loreto, and Domenico Parisi. (2004) Defining and identifying communities in networks. *Publications of the National Academy of Sciences*, 101(9):2658–2663 [Weak and strong definitions of community]

John Hopcroft, Omar Khan, Brian Kulis, and Bart Selman. (2004) Tracking evolving communities in large networks. *Publications of the National Academy of Sciences USA*, 101(Suppl. 1):5249–5253. [Dynamic evolution of communities]

Jordi Duch and Alex Arenas, (2005) "Community detection in complex networks using extremal optimization", Phys. Rev. E **72**, 027104. [Method based on extremal optimization]

JReichardt and Stefan Bornholdt. (2004) Detecting fuzzy community structure in complex networks with a q-state potts model. *Phys. Rev. Lett.*, 93:218701. [Method based on spin models]

L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. (2007) Characterization of complex networks: A survey of measurements. *Adv. Phys.*, 56:167–242. [A proposal for the characterization of networks].

Leon Danon, Albert Díaz-Guilera, and Alex Arenas. (2006) The effect of size heterogeneity on community identification in complex networks. *J. Stat. Mech.*, 2006:P11010+ [Newman's fast algorithm adapted to inhomogeneous networks]

Leon Danon, Alex Arenas, and Albert D. Guilera. (2008) Impact of community structure on information transfer. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 77. [Effect of the community structure on the dynamics of information transfer]

Luca Donetti and Miguel A. Muñoz. (2004) Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, (P10012). [Method based on spectral properties]

Mark E. J. Newman. (2003) The structure and function of complex networks. *SIAM Review*, 45:167–256. [A more recent review on complex networks].

Mark E. J. Newman. (2004) Detecting community structure in networks. *European Physics Journal B*, 38:321–330. [Early review on community detection in complex networks]

Mark E. J. Newman and Michelle Girvan. (2004) Finding and evaluating community structure in networks. *Physical Review E*, 69:026113. [Computation of communities through betweenness].

Mark E. J. Newman. (2001) Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64:016132. [A paper on the properties of scientific networks].

Mark E. J. Newman. (2004) Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(066133). [Newman's fast algorithm]

Marta Sales-Pardo, Roger Guimera, Andre A. Moreira, and Luis A. Amaral. (2007) Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104:15224–15229 [Hierarchical detection of communities]

Matthieu Latapy and Pascal Pons. (2004) Computing communities in large networks using random walks. *cond-mat/0412568*. [Method based on random walks properties]

Michelle Girvan and Mark E.J. Newman. (2002) Community structure in social and biological networks. *Publications of the National Academy of Sciences USA*, 99(12):7821–7826 [Definition of modularity]

R. Pastor-Satorras, M. Rubi, and A. Díaz-Guilera, (eds.) (2003) *Proceedings of the Conference "Statistical Mechanics of Complex Networks"*. Springer. [A series of contributions on complex networks].

Roger Guimera, Leon Danon, Albert Diaz-Guilera, Francesc Giralt, and Alex Arenas. (2003) Self-similar community structure in a network of human interactions. *Physical Review E*, 68(065103) [Community analysis of the email network of Universitat Rovira I Virgili]

Roger Guimerà, Marta Sales, and Luìs N. A. Amaral. (2004) Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70:025101. [Computation of modularity in a null case].

S. Bornholdt and H. G. Schuster, (eds.) (2002) *Handbook of Graphs and Networks - From the Genome to the Internet*. Wiley-VCH, Berlin. [First edited book on complex networks]

Santo Fortunato and Marc Barthelemy. (2007) Resolution limit in community detection. *PNAS*, 104:36–41. [Discussion about the resolution limit in communities' identification]

Santo Fortunato, Vito Latora, and Massimo Marchiori. (2004) Method to find community structures based on information centrality. *Physical Review E*, 70(056104). [A betweenness related method].

Sergei N. Dorogovtsev and J. F. F. Mendes. (2003) *Evolution of Networks: From biological nets to the internet and WWW*. Oxford University Press, Oxford. [A textbook on complex networks with mathematical approach].

Stefan Boettcher and Allon G. Percus. (2001) Extremal optimization for graph partitioning. *Physical Review E*, 64 [Application of extremal optimization]

Stefan Boettcher and Allon G. Percus. (2001) Optimization with extremal dynamics. *Physical Review Letters*, 86(23):5211–5214 [Basic reference for extremal optimization methods]

Steven H. Strogatz. (2001) Exploring complex networks. *Nature*, 410:268–276. [a review on small-world effect and dynamical properties of complex networks].

Ulrik Brandes. (2001) A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, [The fastest method to compute betweenness].

Biographical Sketches

Alex Arenas: is associate professor at the University Rovira i Virgili. He doctorates in physics in 1996 at the University of Barcelona. His research covers aspects of statistical physics and computer science. He has published more than 80 papers and participated in 24 research projects.

Leon Danon: is a Research Fellow at the Harvard School of Public and the University of Warwick. He has worked on problems such as the nature of earthquake patterns, community detection in complex network and epidemic dynamics in structured populations (both human and animal). He has extensive experience in the analysis of large complex datasets and development of stochastic models of infectious diseases at multiple scales.

Albert Diaz-Guilera: Associate professor in Condensed Matter Physics at Universitat de Barcelona. PhD in Physics in 1987 from Universitat Autonoma de Barcelona. Expert in Statistical Physics, in the last years he has been specialized in the analysis of complex networks, mainly their dynamical properties.

Jordi Duch: He studied Computer Science at University Rovira Virgili. PhD in the Department of Physics, at Universitat de Barcelona under the supervision of Dr. Alex Arenas. Currently postdoctoral fellow in Luis Amaral's Group, in the Department of Chemical and Biological Engineering of Northwestern University.

Sergio Gómez: is associate professor at Universitat Rovira i Virgili, Tarragona (Spain). He has degrees in physics (1990) and mathematics (1995), and PhD in physics (1994) at Universitat de Barcelona. His

research is concentrated in two main fields, artificial neural networks and complex networks, and covers both theory and application to real world phenomena.