arXiv:physics/0504059 v1 8 Apr 2005

Improved spectral algorithm for the detection of network communities

Luca Donetti* and Miguel A. Muñoz†

*Departamento de Electrónica y Tecnología de Computadores,
Universidad de Granada, 18071 Granada, Spain
†Departamento de Electromagnetismo y Física de la Materia e
Instituto de Física Teórica y Computacional Carlos I,
Universidad de Granada, 18071 Granada, Spain

Abstract. We review and improve a recently introduced method for the detection of communities in complex networks. This method combines spectral properties of some matrices encoding the network topology, with well known hierarchical clustering techniques, and the use of the *modularity* parameter to quantify the goodness of any possible community subdivision. This provides one of the best available methods for the detection of community structures in complex systems.

Complex networks have recently been an active topic of investigation in physics because of their relevance in the modeling of many real complex systems ranging from social and communication networks to biology and neural sciences [1]. A common feature of many of these real networks is the presence of communities, that is subsets of nodes with high mutual interconnectivity and only few links to the rest of the network.

The importance of their proper detection stems from many different causes: first of all they provide a coarse-grained structure that can notoriously simplify the analysis of a large network. Moreover, communities can be identified as functional units in several cases of biochemical or neural networks. Therefore, even if there is no commonly accepted quantitative definition of community, many algorithms have been proposed to split a network into densely interconnected subsets [2, 3, 4]. A recent comparative review of most of the available community finding methods can be found in [5].

For other problems, similar in spirit to this one, as for example graph partitioning (in a given number of subsets), image partitioning, or graph visualization, spectral techniques have proven to be very useful [6]. Such methods are based on the spectral analysis of a suitable matrix encoding the corresponding network topology. Similar techniques can also be exploited for the detection of communities [7, 8].

Here we give a brief outline of a method we recently introduced [8] which combines spectral properties, hierarchical clustering techniques, and the optimization of the modularity (a quantity introduced to quantify the validity of any given community subdivision) [9].

The nodes of a given network are represented as points in a D-dimensional space whose coordinates are the components of the first D non-trivial eigenvectors of the corresponding Laplacian matrix [11]. Once the nodes have been embedded in a space, a distance (Euclidean, angular, etc [8]) between them can be defined. Afterwards, stan-

dard methods such as hierarchical clustering techniques [10, 8] are employed to group the nodes according to their mutual distances: nearby sites are progressively grouped together. Proceeding like this, a dendrogram, that is a tree representing the hierarchy, is obtained. In order to determine at which level the "tree" should be looked at to obtain the best community-splitting, we have to quantify the quality of the partitions. For this purpose, the *modularity* Q, defined as the fraction of internal edges minus its expected value for a random graph with the same number of links for each community, has been introduced. The output of the algorithm is therefore the partition of the dendrogram giving the highest value of Q.

The justification for using the eigenvectors of the Laplacian matrix representing the network, can be understood by exploiting the connection between the eigenvalue problem and the minimization of the quadratic form

$$\sum_{\text{links}} (x_i - x_j)^2 = x^{\mathrm{T}} \mathbf{L} x, \tag{1}$$

where the $x = \{x_i\}$ is a vector of real values assigned to the nodes and $\mathbf{L}$ is the Laplacian matrix [11]. Minimizing this expression is a way to impose the condition that connected nodes should be given a similar value of x. Indeed, it is easy to see that minimizing equation (1) with a normalization condition on vector x ($\sum x_i^2 = 1$) yields the eigenvalue equation for matrix $\mathbf{L}$. The first eigenvector is trivial (constant) and the corresponding eigenvalue is zero: actually if all x_i are equal the sum (1) is zero and it is its minimum possible value. The following eigenvector (with an eigenvalue larger than 0 for any connected network) corresponds to the non-trivial minimum and therefore its components can be used to partition the nodes. Indeed, as shown in [8], also the following eigenvectors contain useful information and can be profitably used to find communities in the network. The number of eigenvectors D that have to be taken into account in order to obtain a good detection of communities is *a priori* not known. Therefore, the whole procedure is repeated in the algorithm for different D's and the subdivision corresponding to the highest value of the modularity is selected.

If we assign to the nodes a weight proportional to their degree, the normalization condition becomes $\sum k_i x_i^2 = x^{\mathrm{T}} \mathbf{D} x = 1$; in this case the minimization of equation (1) is transformed into the eigenvalue equation for the matrix $\mathbf{L}' = \mathbf{D}^{-1}\mathbf{L}$. As before, the first non trivial eigenvector corresponds to the non-trivial minimum of the sum (1). Therefore, we can wonder how the original method performance (as presented in [8]) is affected by replacing the eigenvectors of $\mathbf{L}$ by those of $\mathbf{L}'$.

First of all, we applied both algorithms (with $\mathbf{L}$ and with $\mathbf{L}'$ respectively) [1] to computer generated networks with a given community structure [2]. These networks contain 128 nodes, split into 4 equal-size communities; edges are randomly extracted in such a way that each node has, on the average, k_{in} links to other nodes in the same community and k_{out} to to the rest of the network, with $k_{\text{in}} + k_{\text{out}} = 16$. For small k_{out} the communities are almost disconnected, while increasing this value they become less and less separated, so that detecting them becomes a very difficult and not clearcut task. Since the communities are known, we can measure the quality of the algorithm by counting the number of nodes

[1] An implementation of the algorithms can be found at `http://www.ugr.es/~donetti/`.

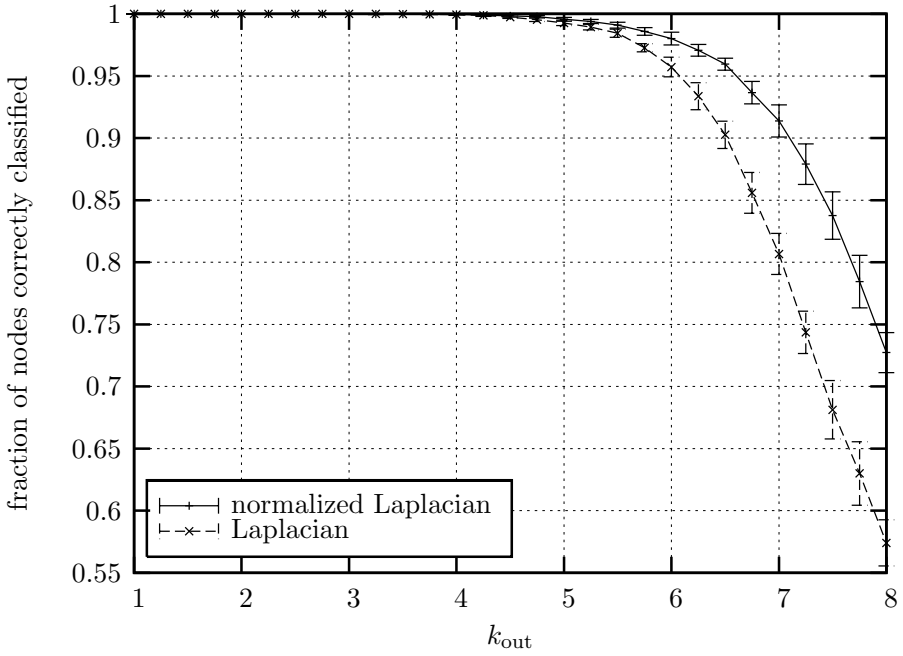

FIGURE 1. Fraction of nodes correctly classified by the algorithm (averaged over 200 networks) as a function of k_{out}, using the eigenvectors of $\mathbf{L}$ and $\mathbf{L}'$. In both cases angular distance and complete linkage clustering are used (see [8]).

that are correctly classified. In figure 1 we plot the corresponding fraction of nodes, and we can see that when the eigenvectors of the normalized Laplacian matrix $\mathbf{L}'$ are used, the method produces much better results. Moreover, in a very recent independent and systematic comparison of different community-finding methods performed by Danon et al. [5], it has been found that our method, equipped with the normalized Laplacian matrix, exhibits an extremely good performance and is among the most convenient choices.

Another network which is used as a test for many community finding algorithm is the Zachary karate club [12]. In this case, we can compare the modularity value corresponding to the best split in the two cases: using the Laplacian eigenvectors we obtain $Q = 0.412$ while using the eigenvectors of $\mathbf{L}'$ leads to $Q = 0.419$ which is the best value obtained so far for such a workbench problem [4].

As a last example, we have studied the jazz bands network [13], which is also one of the prototypical instances studied in this field. Using the Laplacian we measure $Q = 0.437$, while with $\mathbf{L}'$ the modularity increases to $Q = 0.444$ (almost identical to the best available result [4]).

Summarizing, we outlined the connection between the detection of communities and the spectral properties of some proper matrices describing the network topology. Moreover, we improved the performance of the algorithm described in [8] by using the eigenvectors of a different matrix: the normalized Laplacian matrix. We do not have a clear understanding of why the method equipped with this matrix gives better results than with the Laplacian matrix, but as a matter of fact this is actually the case in all

the tested examples. Finally, let us mention that the method (with either matrix) can be easily generalized to the case of weighted networks.

ACKNOWLEDGMENTS

We are thankful to L. Danon and A. Arenas for a useful exchange of correspondence, and to M. Verbeni for a critical reading of the manuscript. Also, financial support from the Spanish MCyT (FEDER) under project BFM2001-2841 and the EU COSIN project IST2001-33555 is acknowledged.

REFERENCES

1. S. H. Strogatz, Nature **410**, 268 (2001). A. L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
   S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford University Press, Oxford (2003).
   R. Pastor Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics approach*, Cambridge University Press (2004).
2. M. Girvan, M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821-7826 (2002).
3. M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
   A. Clauset, M. E. J. Newman, C. Moore, Phys. Rev. E **70**, 066111 (2004).
   F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. USA **101**, 2658-2663 (2004).
   F. Wu, B. A. Huberman, Eur. Phys. J. B **38**, 331-338 (2004).
   J. Reichardt, S. Bornholdt, Phys. Rev. Lett. **93** 218701.
   A. Capocci, V. Servedio, F. Colaiori, and G. Caldarelli, in *Algorithms and models for the Web-Graph: Third International Workshop, WAW 2004, Rome, Italy, 2004. Proceedings, vol. 3243 of Lecture Notes in Computer Science, pag. 181*; and cond-mat/0402499.
   S. Fortunato, V. Latora, M. Marchiori, Phys. Rev E **70** 056104.
   Seung-Woo Son, Hawoong Jeong, and Jae Dong Noh, cond-mat/0502672.
   M. Latapy and P. Pons, cond-mat/0412368.
4. J. Duch, A. Arenas, cond-mat/0501368.
5. L. Danon, J. Dutch, A. Arenas, and A. Díaz-Guilera. Preprint (2005); and private communication.
6. M. Fiedler, Czech. Math. J. **23**, 298-305 (1973) .
   A. Pothen, H. Simon, K.-P. Liou, SIAM J. Matrix Anal. Appl. **11**, 430-452, (1990).
   U. Brandes, M. Gaertler, and D. Wagner. Proc. 11th Europ. Symp. Algorithms (ESA'03), LNCS 2832, pp. 568-579.
   A. Y. Ng, M. I. Jordan, and Y. Weiss, Advances in Neural Information Processing Systems **14**, 849 (2002).
7. C. Borgs, J. T. Chayes, M. Mahdian and A. Saberi, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge, Discovery and Data Mining (2004).
8. L. Donetti and M. A. Muñoz, J. Stat. Mech. P10012, (2004).
9. M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
10. B. S. Everitt, *Cluster Analysis*, Edward Arnold, London (1993).
11. N. L. Biggs, *Algebraic Graph Theory*, Cambridge University Press (1974).
12. W. W. Zachary, J. of Anthropological Research **33**, 452 (1977).
13. P. Gleiser and L. Danon, Advances in Complex Systems, **6**, 565 (2003).