Communities Detection in Large Networks

Andrea Capocci[1], Vito D.P. Servedio[1,2],
Guido Caldarelli[1,2], and Francesca Colaiori[2]

[1] Centro Studi e Ricerche e Museo della Fisica "E. Fermi",
Compendio Viminale, Roma, Italy
[2] INFM UdR Roma1-Dipartimento di Fisica Università "La Sapienza",
P.le A. Moro 5, 00185 Roma, Italy

Abstract. We develop an algorithm to detect community structure in complex networks. The algorithm is based on spectral methods and takes into account weights and links orientations. Since the method detects efficiently clustered nodes in large networks even when these are not sharply partitioned, it turns to be specially suitable to the analysis of social and information networks. We test the algorithm on a large-scale data-set from a psychological experiment of word association. In this case, it proves to be successful both in clustering words, and in uncovering mental association patterns.

1 Communities in Networks

Complex networks are composed by a large number of elements (nodes), organized into sub-communities summing up to form the whole system. The partition of a complex network has no unique solution, since nodes can be aggregated according to different criteria. However, when dealing with networks an additional constraint bounds the possible choices: the entire information is coded into the adjacency matrix, whose elements describe the connection between each pair of nodes. Detecting communities, therefore, means uncovering the reduced number of interactions involving many unit elements, given a large set of pair interactions.

The partition of a graph into communities has a broad range of technological applications (from the detection of polygenic interaction in genetic networks to the development of effective tools for information mining in communication networks [1–3]). A perhaps more intriguing scientific interest in network partitioning comes from social sciences, where methods to detect groups within social networks are employed on a daily basis [4]. Moreover, when used in the analysis of large collaboration networks, such as company or universities, communities reveal the informal organization and the nature of information flows through the whole system [5, 6]. Social networks, however, are usually described by undirected graphs, where links are reciprocal. This somewhat simplifies the task, and subtler methods are required when considering directed graphs, as shown below.

Despite the potential application of the results, measurements about structures involving more than two nodes in networks mainly concern regular pat-

S. Leonardi (Ed.): WAW 2004, LNCS 3243, pp. 181–187, 2004.

terns [7–13], and particularly the clustering coefficient, which counts the number of triangles in a graph.

2 Network Partitioning Algorithm

Indeed, few scientist have developed methods and algorithms able to identify irregularly shaped communities. Traditional methods are divisive, i.e. they are based on the fragmentation of the original network into disconnected components by the removal of links.

2.1 Edge-Betweenness Methods

Recent algorithms [3, 14] are mainly based on the edge betweenness or local analogues of it. Edge betweenness measures how central an edge is: to assess this, one finds the shortest paths between each pair of nodes, and counts the fraction of them which runs over the considered edge. Removing a sufficient number of such edges from the network fragments it into separate parts. The process is iterated until the network is split into cluster including only individual nodes. The NG–algorithm builds a tree or, more exactly, a dendrogram: at each splitting event, nodes are subdivided in families, sub-families and so on.

This methods is widely assumed to give reasonable partition of networks, since the first splittings produce the most basic subgroups, whereas successive splittings gives a classification of the nodes into families at a higher resolution. Based on a similar principle, the method introduced in ref. [14] has the advantage of being faster. Despite its the outcome is independent on how sharp the partitioning of the graph is.

2.2 Spectral Methods

Spectral methods study the adjacency matrix A [15–17], whose generic element a_{ij} is equal to 1 if i points to j and 0 otherwise.

In particular, such methods analyze matrix related to A, such as the Laplacian matrix $K - A$ and the Normal matrix $K^{-1}A$, where K is the diagonal matrix with elements $k_{ii} = \sum_{j=1}^{S} a_{ij}$ and S is the number of nodes in the network. In most approaches, referring to undirected networks, A is assumed to be symmetric.

The Laplacian and Normal spectrum have peculiar shapes. The largest eigenvalue of the Normal spectrum is equal to 1 by definition, since it corresponds to the normalization to one imposed to elements on a same matrix row. But non-principal eigenvalues close to 1 gives an insight about the network partitioning, since they correspond roughly to the number of clear components of the graph. A perturbative approach helps to understand this: a block matrix would have an eigenvalue equal to 1 for each matrix block, since the normalization applies to each block as well. If some link across the block is present, this slightly perturbs the spectrum, but eigenvalues close to one are still meaningful if a clear partition

is present. A similar argument applies for the Laplacian matrix, but in this case eigenvalues close to 0 are to be looked at.

The eigenvectors associated to the largest eigenvalues of the Laplacian and Normal spetrum have a characteristic structure too: the components corresponding to nodes within the same cluster have very similar values x_i, so that, as long as the partition is sufficiently sharp, the profile of each eigenvector, sorted by components, is step–like. The number of steps in the profile corresponds again to the number of communities.

This is explained by mapping the matrix eigenproblem into a constrained minimization process. With the most general applications in mind, we replace the adjacency matrix A by the weight matrix W, whose elements w_{ij} are assigned the intensity of the link (i, j). We consider undirected graphs first, and then we pass to the most general directed case.

The spectrum of such matrices are related to the minimization of

$$z(\mathbf{x}) = \frac{1}{2} \sum_{i,j=1}^{S} (x_i - x_j)^2 w_{ij} \,, \tag{1}$$

where x_i are values assigned to the nodes, with some constraint on the vector $\mathbf{x}$, expressed by

$$\sum_{i,j=1}^{S} x_i x_j m_{ij} = 1 \,, \tag{2}$$

where m_{ij} are elements of a given symmetric matrix M.

Writing $\frac{dz}{dx_i} = 0$ for all x_i in vector formalism, given the constraint (2) leads to equation

$$(D - W)\mathbf{x} = \mu M \mathbf{x} \,, \tag{3}$$

where D is the diagonal matrix $d_{ij} = \delta_{ij} \sum_{k=1}^{S} w_{ik}$, and μ is a Lagrange multiplier.

The constraint Matrix M determine the kind of eigenproblem to solve. Choosing $M = D$ to $D^{-1}W\mathbf{x} = \mu\mathbf{x}$ (related to the generalized Normal spectrum), while $M = 1$ leads to $(D - W)\mathbf{x} = \mu\mathbf{x}$ (corresponding to the Laplacian case).

The absolute minimum corresponds to the trivial eigenvector, which is constant. The other stationary points correspond to eigenvectors where components associated to well connected nodes assume similar values.

As an example, we show in Fig. 2 the step-like profile of the second eigenvectors of $D^{-1}W$ for the simple graph shown in Fig. 1 with $S = 19$ nodes, where random weights between 1 and 10 were assigned to the links.

Yet, studying the eigenvector profile is meaningful only if such a sharp partition exists. In most cases, especially in large networks systems, eigenvector profiles are far too complicate to detect steps, and components do not cluster around a few values. Nevertheless, the minimization problem still applies, so that clustered node correspond to components with similar values in many eigenvectors.

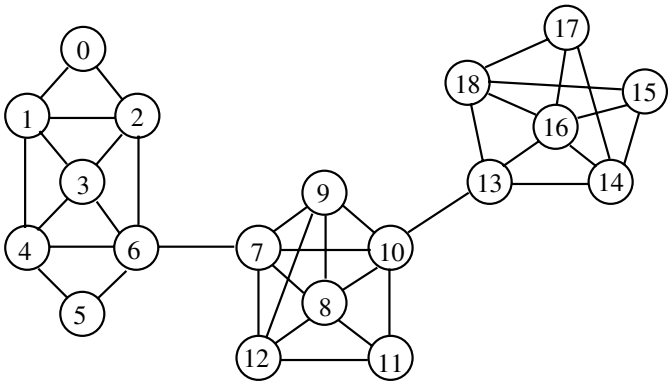

Fig. 1. Network employed as an example, with $S = 19$ and random weights between 1 and 10 assigned to the links. Three clear clusters appear, composed by nodes $0 - 6$, $7 - 12$ and $13 - 19$

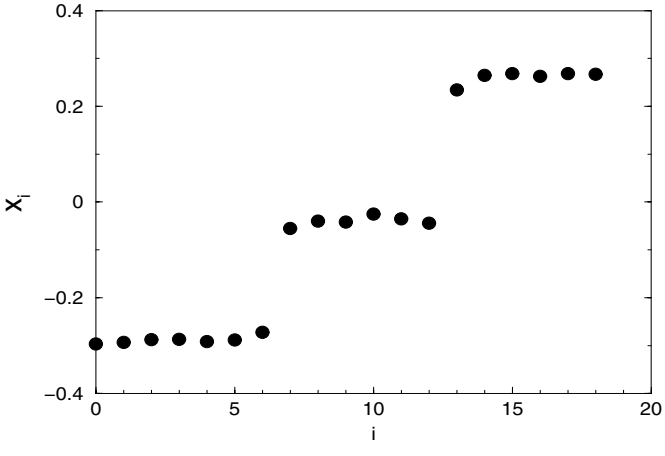

Fig. 2. Values of the 2nd eigenvector components for matrix $D^{-1}W$ relative to the graph depicted in figure 1

By measuring the correlation

$$r_{ij} = \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{[(\langle x_i^2 \rangle - \langle x_i \rangle^2)(\langle x_j^2 \rangle - \langle x_j \rangle^2)]^{\frac{1}{2}}}\,, \tag{4}$$

where the average $\langle \cdot \rangle$ is over the first few nontrivial eigenvectors, one is then able to detect such a rigid displacement of the components across different eigenvectors.

The quantity r_{ij} measures the community closeness between node i and j. Though the performance may be improved by averaging over more and more

eigenvectors, with increased computational effort, we find that indeed a small number of eigenvectors suffices to identify the community to which nodes belong, even in large networks.

3 Spectral Methods for Directed Networks

When dealing with a directed network, links do not correspond to any equivalence relation. Rather, pointing to common neighbors is a significant relation, as suggested in the sociologists' literature where this quantity measures the so-called *structural equivalence* of nodes [18]. Accordingly, in a directed network, clusters should be composed by nodes pointing to a high number of common neighbors, no matter their direct linkage.

To detect the community structure in a directed network, we therefore replace, in the previous analysis, the matrix W by matrix $Y = WW^T$. This corresponds to replacing the directed network with an undirected weighted network, where nodes pointing to common neighbors are connected by a link whose intensity is proportional to the total sum of the weights of the links pointing from the two original nodes to the common neighbors. The previous analysis wholly applies: the function to minimize in this case is

$$y(\mathbf{x}) = \sum_{ijl}^{1,S} (x_i - x_j)^2 w_{il} w_{jl} \,. \tag{5}$$

Defining Q as the diagonal matrix $q_{ij} = \delta_{ij} \sum_{lj=1}^{S} w_{il} w_{jl}$, the eigenvalue problem for the analogous of the generalized normal matrix,

$$Q^{-1} Y \mathbf{x} = \lambda \mathbf{x} \tag{6}$$

is equivalent to minimizing the function (5) under the constraint $\sum_{ijl=1}^{S} x_i x_j q_{ij} = 1$.

4 A Test: The Word Association Network

To test this spectral correlation-based community detection method on a real complex network, we apply the algorithm to data from a psychological experiment reported in reference [19]. Volunteering participants to the research had to respond quickly by freely associating a word (response) to another word given as input (stimulus), extracted by a fixed subset. Scientists conducting the research have recorded all the stimuli and the associated responses, along with the occurrence of each association. As in ref. [20], we build a network were words are nodes, and directed links are drawn from each stimulus to the corresponding responses, assuming that a link is oriented from the stimulus to the response. The resulting network includes $S = 10616$ nodes, with an average in-degree equal to about 7. Taking into account the frequency of responses to a given stimulus, we construct the weighted adjacency matrix W. In this case, passing to the

matrix Y means that we expect stimuli giving rise to the same response to be correlated.

The word association network is an ideal test case for our algorithm, since words are naturally associated by their meaning, so that the performance of our method emerges immediately emerges at glance, when one looks at words falling in the same cluster.

However, in such large databases a partition is not defined, there Rather, one is interested in finding groups of highly correlated nodes, or groups of nodes highly connected to a given one. Table 1 shows the most correlated words to three test-words. The correlation are computed by averaging over just 10 eigenvectors of the matrix $Q^{-1}Y$: the results appear to be quite satisfactory, already with this small number of eigenvectors.

Table 1. The words most correlated to *science*, *literature* and *piano* in the eigenvectors of $Q^{-1}WW^T$. Values indicate the correlation

science	1	literature	1	piano	1
scientific	0.994	dictionary	0.994	cello	0.993
chemistry	0.990	editorial	0.990	fiddle	0.992
physics	0.988	synopsis	0.988	viola	0.990
concentrate	0.973	words	0.987	banjo	0.988
thinking	0.973	grammar	0.986	saxophone	0.985
test	0.973	adjective	0.983	director	0.984
lab	0.969	chapter	0.982	violin	0.983
brain	0.965	prose	0.979	clarinet	0.983
equation	0.963	topic	0.976	oboe	0.983
examine	0.962	English	0.975	theater	0.982

As sown in table 1, the results are quite satisfying: most correlated words have closely related meanings or are directly associated to the test word by a simple relation (synonymy or antinomy, syntactic role, and even by analogous sensory perception).

5 Conclusions

We have introduced a method to detect communities of highly connected nodes within a network. The method is based on spectral analysis and applies to weighted networks. When tested on a real network instance (the records of a psychological experiments) the algorithm proves to be successful: it clusters nodes or, in such case, words, according to natural criteria, and provides an automatic way to extract the most connected sets of nodes to a given one in a set of over 10^4.

The authors thank Miguel-Angel Muñoz and Ramon Ferrer Y Cancho for useful discussion. They acknowledge partial support from the FET Open Project IST-2001-33555 COSIN.

References

1. I. Simonsen, K. A. Eriksen, S. Maslov, K. Sneppen, cond-mat/0312476 (2003), to appear in *Physica A*
2. S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, *The VLDB Journal*, 639 (1999).
3. M. Girvan and M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **99**, 8271 (2002).
4. M. E. J. Newman, *SIAM Review* **45**, 167 (2003).
5. B. Huberman, J. Tyler and D. Wilkinson, in *Communities and technologies*, M. Huysman, E. Wegner and V. Wulf, eds. Kluwer Academic (2003).
6. R. Guimerà, L. Danon, A. Diaz-Guilera, F. Giralt and A. Arenas, *Phys. Rev. E* **68** 065103 (2003)
7. R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
8. S. N. Dorogovtsev and J. F. F. Mendes, *Adv. in Phys.* **51**, 1079 (2002).
9. J.P. Eckmann, E. Moses, *PNAS* **99** (9), 5825 (2002).
10. G. Bianconi and A. Capocci, *Phys. Rev. Lett.* **90** ,078701 (2003).
11. G. Caldarelli, R. Pastor-Satorras and A. Vespignani, cond-mat/0212026 (2002).
12. A. Capocci, G. Caldarelli, P. De Los Rios, *Phys. Rev.* E **68** 047101 (2003).
13. G. Caldarelli, A. Capocci, P. De Los Rios, M.A. Muñoz, *Phys. Rev. Lett.* **89** 258702 (2002).
14. F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, submitted for publication, preprint cond-mat/0309488
15. K. M. Hall, *Management Science* **17**, 219 (1970).
16. A.J. Seary and W.D. Richards, W.D. *Proceedings of the International Conference on Social Networks Volume 1: Methodology*, 47 (1995).
17. J. Kleinberg, *Journal of the ACM* **46** (5) 604 (1999).
18. M. E. J. Newman, *Eur. Phys. J. B*, in press.
19. M. Steyvers, J. B. Tenenbaum, preprint cond-mat/0110012, submitted for publication.
20. L. Da Fontoura Costa, preprint cond-mat/0309266, submitted for publication