Analysis of community structure in networks of correlated data

Sergio Gómez,[1] Pablo Jensen,[2] and Alex Arenas[1]
[1]*Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain*
[2]*IXXI–Institut des Systèmes Complexes, 5 rue du Vercors, 69007 Lyon, France*
(Received 5 December 2008; revised manuscript received 16 March 2009; published 22 July 2009)

We present a reformulation of modularity that allows the analysis of the community structure in networks of correlated data. The modularity preserves the probabilistic semantics of the original definition even when the network is directed, weighted, signed, and has self-loops. This is the most general condition one can find in the study of any network, in particular those defined from correlated data. We apply our results to a real network of correlated data between stores in the city of Lyon (France).

DOI: 10.1103/PhysRevE.80.016114 PACS number(s): 89.75.Hc, 02.10.Ox, 02.50.−r

I. INTRODUCTION

Complex networks are graphs representative of the intricate connections between elements in many natural and artificial systems [1–4], whose description in terms of statistical properties have been largely developed looking for a universal classification of them. However, when the networks are locally analyzed, some characteristics that become partially hidden in the global statistical description emerge. The most relevant is perhaps the discovery in many of them of *community structure*, meaning the existence of densely (or strongly) connected groups of nodes, with sparse (or weak) connections between these groups [5].

The study of the community structure helps to elucidate the organization of the networks and, eventually, could be related to the functionality of groups of nodes [6]. The most successful solutions to the community detection problem, in terms of accuracy and computational cost required, are those based in the optimization of a quality function called *modularity* proposed by Newman and Girvan [7] that allows the comparison of different partitioning of the network. The extension of modularity to weighted [8] and directed networks [9,10] has been the first step towards the analysis of the community structure in general networks. Nevertheless, we point out that the use of modularity in community detection suffers under a resolution limit [11]. This problem can make modularity useless in situations where real communities present sizes below a certain resolution scale; however the problem is alleviated by applying multiresolution methods, based on the same structural properties of modularity, to resolve these scales [9,12].

Very often networks are defined from correlation data between elements. The common analysis of correlation matrices uses classical or advanced statistical techniques [13]. Nevertheless an alternative analysis in terms of networks is possible. The network approach usually consists in to filter the correlation data matrix, by eliminating poorly correlated pairs according to a threshold, and by keeping unsigned the value of the correlation, producing a network of positive links and no self-loops (self-correlations). Recently, some authors pointed out the possibility to analyze these networks via spectral decomposition [14,15]. We devise also the possibility to analyze them in terms of Newman's modularity to reveal the community structure (clusters) of the correlated data. However, any of these approaches can be misleading because of two facts: first, the sign of the correlation is important to avoid the mixing of correlated and anticorrelated data, and second, the existence of self-loops is critical for the determination of the community structure [9]. Here we propose a method to extract the community structure in networks of correlated data, which accounts for the existence of signed correlations and self-correlations, preserving the original information. To this end, we extend the modularity to the most general case of directed, weighted, and signed links. We will show the performance of our method in a real network of correlations between commercial activities, previously analyzed in [16] using a Potts model.

II. GENERALIZATION OF MODULARITY

Given an undirected network partitioned into communities, the modularity of a given partition is, up to a multiplicative constant, the probability of having edges falling within groups in the network minus the expected probability in an equivalent (null case) network with the same number of nodes, and edges placed at random preserving the nodes' strength, where the strength of a node stands for the sum of the weights of its connections [8]. In mathematical form, being C_i the community to which node i is assigned, modularity is expressed in terms of the weighted adjacency matrix w_{ij}, which represents the value of the weight in the link between i and j (0 if no link exists), as [8]

$$Q = \frac{1}{2w}\sum_i \sum_j \left(w_{ij} - \frac{w_i w_j}{2w} \right) \delta(C_i, C_j), \quad (1)$$

where the Kronecker delta function $\delta(C_i, C_j)$ takes the values, 1 if nodes i and j are into the same community, 0 otherwise, the strengths $w_i = \Sigma_j w_{ij}$, and the total strength $2w = \Sigma_i w_i = \Sigma_i \Sigma_j w_{ij}$.

The larger the modularity the better the partitioning since more deviates from the null case. Note that the optimization of the modularity cannot be performed by exhaustive search since the number of different partitions is equal to the Bell [17] or exponential numbers, which grow at least exponentially in the number of nodes N. Indeed, optimization of modularity is a nondeterministic polynomial-time (NP)-hard problem [18]. Several authors have attacked the problem

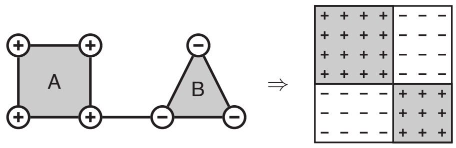

FIG. 1. Network with well-defined community structure and its correlation matrix.

proposing different optimization heuristics [19–24].

To demonstrate the flaws of modularity when trying to extract the community structure of correlated data we show the following example. Suppose we have a network with a well-defined community structure as the one presented in Fig. 1. Let us pretend that each community is indeed a functional community, in such a way that nodes in every group have different states. To simplify the mathematics we will consider that the nodes in community A are in a state +1 and nodes in community B are in a state −1. After, we define the correlation between these data as, for example, $R_{ij}=S_iS_j$ being S_i and S_j the corresponding states of nodes i and j. The question is: can we infer communities A and B from the correlated data represented in matrix R? The answer is that applying modularity, no. Let us sketch the proof, the matrix R is blockwise composed of submatrices R_{AA}, R_{AB}, R_{BA}, and R_{BB}. The blocks R_{AA} and R_{BB} are all valued +1, and R_{AB} and R_{BA} are valued −1. Any matrix of this form results in zero modularity Eq. (1) for all partitions since $R_{ij}=\frac{w_iw_j}{2w}$ for all pairs.

To reveal the community structure in the network presented in Fig. 1 from its correlation matrix, it is necessary to revise the formulation of modularity. Let us suppose we have a weighted undirected complex network with weights w_{ij} as above. The relative strength p_i of a node

$$p_i = \frac{w_i}{2w} \tag{2}$$

may be interpreted as the probability that this node makes links to other ones, if the network were random. This is precisely the approach taken by Newman and Girvan to define the modularity null case term, which reads

$$p_ip_j = \frac{w_iw_j}{(2w)^2}. \tag{3}$$

The introduction of negative weights destroys this probabilistic interpretation of p_i since in this case the values of p_i are not guaranteed to be between zero and one. The problem is the implicit hypothesis that there is only one unique probability to link nodes, which involves both positive and negative weights. However, if we suppose there are two different probabilities to form links, one for positive and the other for negative weights, the problem disappears.

Let us formalize this approach. First, we separate the positive and negative weights,

$$w_{ij} = w_{ij}^+ - w_{ij}^-, \tag{4}$$

where

$$w_{ij}^+ = \max\{0, w_{ij}\}, \tag{5}$$

$$w_{ij}^- = \max\{0, -w_{ij}\}. \tag{6}$$

The positive and negative strengths are given by

$$w_i^+ = \sum_j w_{ij}^+, \tag{7}$$

$$w_i^- = \sum_j w_{ij}^-, \tag{8}$$

and the positive and negative total strengths by

$$2w^+ = \sum_i w_i^+ = \sum_i \sum_j w_{ij}^+, \tag{9}$$

$$2w^- = \sum_i w_i^- = \sum_i \sum_j w_{ij}^-. \tag{10}$$

Obviously it is satisfied that

$$w_i = w_i^+ - w_i^- \tag{11}$$

and

$$2w = 2w^+ - 2w^-. \tag{12}$$

With these definitions at hand, the connection probabilities with positive and negative weights are, respectively,

$$p_i^+ = \frac{w_i^+}{2w^+}, \tag{13}$$

$$p_i^- = \frac{w_i^-}{2w^-}. \tag{14}$$

Now there are two terms which contribute to modularity: the first one takes into account the deviation of actual positive weights against a null case random network given by probabilities p_i^+, and the other is its counterpart for negative weights. Thus, it is useful to define

$$Q^+ = \frac{1}{2w^+}\sum_i \sum_j \left(w_{ij}^+ - \frac{w_i^+w_j^+}{2w^+}\right)\delta(C_i, C_j), \tag{15}$$

$$Q^- = \frac{1}{2w^-}\sum_i \sum_j \left(w_{ij}^- - \frac{w_i^-w_j^-}{2w^-}\right)\delta(C_i, C_j). \tag{16}$$

The total modularity must be a trade off between the tendency of positive weights to form communities and that of negative weights to destroy them. If we want that Q^+ and Q^- contribute to modularity proportionally to their respective positive and negative strengths, the final expression for modularity Q is

$$Q = \frac{2w^+}{2w^+ + 2w^-}Q^+ - \frac{2w^-}{2w^+ + 2w^-}Q^-. \tag{17}$$

An alternative equivalent form for modularity Q is

$$Q = \frac{1}{2w^+ + 2w^-}\sum_i \sum_j \left[w_{ij} - \left(\frac{w_i^+w_j^+}{2w^+} - \frac{w_i^-w_j^-}{2w^-}\right)\right]\delta(C_i, C_j). \tag{18}$$

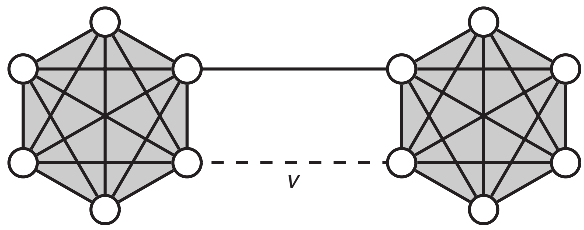

FIG. 2. Network with two well-defined communities. Solid lines correspond to positive links, and the dashed line to the only negative link, with weight $v<0$.

The main properties of Eq. (18) are: without negative weights, the standard modularity is recovered; modularity is zero when all nodes are together in one community; and it is antisymmetric in the weights, i.e., $Q(C,\{w_{ij}\}) = -Q(C,\{-w_{ij}\})$.

The extension to directed networks [25] is simply obtained by the substitutions

$$w_i^{\pm} \rightarrow w_i^{\pm,\text{out}} = \sum_k w_{ik}^{\pm}, \tag{19}$$

$$w_j^{\pm} \rightarrow w_j^{\pm,\text{in}} = \sum_k w_{kj}^{\pm}. \tag{20}$$

III. COMPARISON WITH OTHER METHODS

In Fig. 2 we show a simple example of a network for which the original Newman modularity Eq. (1) and the Potts model in [16] do not yield the expected partition in two communities, whereas our modularity Eq. (18) succeeds. It consists in two cliques, formed by positive links, and connected by two edges, one positive and the other negative. All positive links have a weight +1, and the negative a weight $v<0$. Any size of the cliques greater than or equal to three does the job.

First, the Potts model in [16] is based on a Hamiltonian which only takes into account the difference between positive and negative weights within the modules, and is equivalent to modularity but without the null case term. In the network Fig. 2, if $|v|<1$, the strength between the two cliques is $1+v>0$; thus the Potts model is rewarded to join both cliques in the same module. Clearly, the absence of the null case is responsible of this incorrect result. On the other hand, the original definition of modularity Eq. (1), although it has a null case, it was not designed to cope with negative weights. In this example, its optimal partition is again a single module containing all the nodes if the value of $|v|$ is greater than the number of positive links.

In a recent preprint [26], the authors proposed a method that also copes with positive and negative strengths. Although in the same spirit of the current work, the authors approach the problem differently, they propose to decompose the positive and negative contribution at the level of modules, aggregating the link weights by addition. Given that the structure of modularity is nonlinear (the null term is quadratic), this prescription presents important differences with our approach, concerning normalization and the relative strength of the positive versus the negative modularity parts.

After this work was finished, the authors became aware of another recent preprint [27] proposing an alternative definition of modularity for positive and negative links. Their work, based on a Potts model representation of the network communities' assignment [28], is totally compatible with the definition presented in the current work, and equivalent for the values of their parameters $\lambda=\gamma=1$.

IV. APPLICATION TO A REAL NETWORK

We now turn to an example of community structure detection using our method in a specific social network. We deal with the spatial distribution of retail activities in the city of Lyon, thanks to data obtained at the Lyon's Commerce Chamber [29]. We have shown in [16] how to transform data on locations into a matrix of correlated data, in this case of attractions/repulsions (i.e., positive and negative links) between retail activities. To compute the interaction between activities A and B, the idea is to compare the concentrations of B stores in the neighborhood of A stores to a reference concentration obtained by locating the B stores randomly. To compute the random reference, the idea [30] is to locate the B stores on the array of *all existing* store sites. This is the best way to take into account automatically the geographical peculiarities of each town. The logarithm of the ratio of the actual concentration to the reference concentration gives the interaction coefficient, which is positive for attractions and negative for repulsions, as anticipated.

More precisely, the (self-)interaction of N_A A stores embedded in a larger set of N_t locations is

$$a_{AA}(r) = \log_{10}\frac{N_t - 1}{N_A(N_A-1)}\sum_{i=1}^{N_A}\frac{N_A(A_i)}{N_t(A_i)}, \tag{21}$$

where $N_A(A_i)$ and $N_t(A_i)$ represent the number of A stores and the total number of stores in the neighborhood of store A_i, i.e., locations at a distance smaller than r. Similarly, the coefficient characterizing the spatial distribution of the B_i around the A_i is

$$a_{AB}(r) = \log_{10}\frac{N_t - N_A}{N_A N_B}\sum_{i=1}^{N_A}\frac{N_B(A_i)}{N_t(A_i) - N_A(A_i)}, \tag{22}$$

where $N_A(A_i)$, $N_B(A_i)$, and $N_t(A_i)$ are, respectively, the A, B, and total number of locations in the neighborhood of point A_i (not counting A_i). Both a_{AA} and a_{AB} are defined so that they take value 0 when there are no spatial correlations. In the case of the a_{AB} coefficient, this means that the local B spatial concentration is not perturbed, on average, by the presence of A stores, and is equal to the average concentration over the whole town, $\frac{N_B}{N_t-N_A}$. Only coefficients which deviate significantly from 0, using a Monte Carlo sampling, are taken into account in the adjacency matrix. The final result of the analysis of the 11 629 stores in Lyon is a directed network with 97 nodes (retail activities) and 1131 links, 715 positive and 416 negative.

TABLE I. Comparison between the different partitions and the Lyon Chamber of Commerce classification.

	Optimal partition of Eq. (1)	Optimal partition of Eq. (18)
Rand Index	0.6168	0.6952
Jaccard Index	0.1336	0.1426
NMI	0.1458	0.2310

We analyze the community structure of the resulting network using the modularity defined in Eq. (18). The optimization method used is Tabu search [9] that for this case gave the highest modularity when compared to others [31]. We perform a comparison between the different partitions obtained optimizing independently Eq. (1) (resulting in four communities) and Eq. (18) (resulting in six communities), against the Lyon's Commerce Chamber retail activities classification (nine communities predefined). The similarity of the first two partitions in front of the third one is measured using three different indices, namely, the Rand index [32], the Jaccard index [33], and the normalized mutual information (NMI) [34] (see Table I). The larger their values, the more similar the partitions are. All indices show a better performance of Eq. (18) discriminating the actual communities provided by the Lyon's Commerce Chamber. Note that in both modularities we have used all the positive and negative links; thus the increase in performance can only be attributed to a proper use of the information embedded in the links.

Our method is also helpful to understand the spatial organization of retail stores. To interpret the information conveyed by the network links, we make use of the z score (Z) [21]. The basic idea consists in to compute the z score of the internal strength of each node with respect to the internal strength of the community to which is assigned. To be consistent with our approach along the paper both quantities should be evaluated consistently with the sign of the interactions and with the directionality of links, then

$$Z_i^{\pm,\text{in/out}} = \frac{w_{i,\text{int}}^{\pm,\text{in/out}} - \langle w_{\text{int}}^{\pm,\text{in/out}} \rangle}{\sigma(w_{\text{int}}^{\pm,\text{in/out}})}, \qquad (23)$$

where subindices "int" express that links are restricted within the community to which node i belongs to, "in/out" refers to the direction of links, and $\langle \cdots \rangle$ and σ are the average and standard deviation of the corresponding variables, respectively.

Using the z score we can answer some questions about the role of nodes in their communities, as for example, for each community, which are: the most attractive retailers (max Z_+^{in}), the most repulsive retailers (max Z_-^{out}), the most attracted retailers (max Z_+^{out}), and the most repelled retailers (max Z_-^{in}). In Table II we show the two highest results of these z scores obtained for the largest community found (33 stores). This group gathers the proximity stores, which means mainly food stores. Here are some examples of the understanding of the spatial organization of retail stores allowed by our method. Sports facilities and funeral services (third in list of most attractive) are peculiar because they strongly attract (and are attracted) by some specific activities that go along with them almost systematically, e.g., car repairs and small hardware stores. Gas stations enjoy a paradoxical situation in this group, since they represent the most attractive and the most repelled activity. There is an interesting commercial interpretation of this paradox: gas stations tend to have the most specific commercial environment, strongly attracting some of the group's activities (such as supermarkets) and being strongly repelled by others which however are in the proximity store group (for example, butchers or fine pastries stores almost never have gas stations close to them). Dairy products and fine pastries strongly repel some specific of the activities that belong to their same group, such as car repairs or firm's restaurants.

TABLE II. Roles of retailers within communities.

+ attractive	+ repulsive	+ attracted	+ repelled
Gas Station	Dairy products	Funeral Services	Gas Station
Sports facility	Cake shop	Sports facility	Flea market

V. CONCLUSIONS

Summarizing, we have proposed a formulation of modularity that allows for the analysis of any complex network, in general with links directed, weighted, signed, and with self-loops, preserving the original probabilistic semantics of modularity. With this definition one can afford the analysis of networks coming from correlated data without the necessity to symmetrize the network, or skipping autocorrelation, or considering the unsigned value of the correlations. We devise that other methods are also likely to be appropriate for this task, after its pertinent adaptation, for example the analysis via clique percolation [35], or specifically methods based on the minimization of the energy function of an equivalent spin-glass system, where weighted signed links can be interpreted in terms of ferromagnetic and antiferromagnetic interactions between spins [28]. It is worth mentioning that our generalization, as well as all those measures based on modularity, suffers under a resolution limit [11]. In cases where an extra resolution is needed, one can take advantage of multiresolution methods [9] simply substituting the original modularity by the general modularity presented here.

We have analyzed within the scope of the modularity an interesting model of attraction-repulsion of retail stores in a large city, previously reported in [16]. The results overcome those obtained using the original definition of modularity when compared to the Lyon Chamber of Commerce classification, and also point out the necessity of defining new roles of nodes based on directionality and sign of the weights of links, as we have proposed for the z score.

ACKNOWLEDGMENTS

We thank J. Borge for help with the simulations. This work has been partially supported by the Spanish DGICYT under Project No. FIS2006-13321-C02-1. A.A. acknowledges support of the Director, Office of Science, Computational and Technology Research, U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We gratefully acknowledge Christophe Baume and Bernard Gagnaire from Lyon's Commerce Chamber who have kindly provided the location data.

[1] S. H. Strogatz, Nature (London) **410**, 268 (2001).
[2] C. M. Song, S. Havlin, and H. A. Makse, Nature (London) **433**, 392 (2005).
[3] A.-L. Barabási, Science **308**, 639 (2005).
[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[5] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[6] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[7] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[8] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).
[9] A. Arenas, A. Fernández, and S. Gómez, New J. Phys. **10**, 053039 (2008).
[10] E. A. Leicht and M. E. J. Newman, Phys. Rev. Lett. **100**, 118703 (2008).
[11] S. Fortunato and M. Barthelemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[12] J. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, Eur. Phys. J. B **56**, 41 (2007).
[13] P. X.-K. Song, *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer Series in Statistics (Springer, New York, 2007).
[14] T. Heimo, J. Saramäki, J.-P. Onnela, and K. Kaski, Physica A **383**, 147 (2007).
[15] T. Heimo, G. Tibély, J. Saramäki, K. Kaski, and J. Kertész, Physica A **387**, 5930 (2008).
[16] P. Jensen, Phys. Rev. E **74**, 035101(R) (2006).
[17] E. T. Bell, Am. Math. Monthly **41**, 411 (1934).
[18] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE Trans. Knowl. Data Eng. **20**, 172 (2008).
[19] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[20] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[21] R. Guimerà and L. A. N. Amaral, J. Stat. Mech. (2005) P02001.
[22] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[23] J. M. Pujol, J. Béjar, and J. Delgado, Phys. Rev. E **74**, 016107 (2006).
[24] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[25] A. Arenas, J. Duch, A. Fernández, and S. Gómez, New J. Phys. **9**, 176 (2007).
[26] T. Kaplan and S. Forrest, e-print arXiv:0801.3290.
[27] V. A. Traag and J. Bruggeman, e-print arXiv:0811.2329.
[28] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).
[29] The Commerce Chamber classifies retail activities according to commercial criteria derived from an experienced knowledge of the field. This classification is adopted here as reference.
[30] G. Duranton and H. G. Overman, Rev. Econ. Stud. **72**, 1077 (2005).
[31] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech. (2005) P09008.
[32] W. M. Rand, J. Am. Stat. Assoc. **66**, 846 (1971).
[33] P. Jaccard, New Phytol. **11**, 37 (1912).
[34] A. Strehl and J. Ghosh, J. Mach. Learn. Res. **3**, 583 (2003).
[35] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).