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*

(Dated: January 16, 2009)

We present a reformulation of modularity that allows the analysis of the community structure in networks of correlated data. The new 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).

PACS numbers: 89.75.Hc, 02.10.Ox, 02.50.-r

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 in the curse 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 perhaps is 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 steps towards the analysis of the community structure in general networks.

Very often networks are defined from correlation data between elements. The common analysis of correlation matrices uses classical or advanced statistical techniques [11]. Nevertheless an alternative analysis in terms of networks is possible. The network approach usually consist 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 [12, 13] . We devise also the possibility to analyze them in terms of 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 anti-correlated 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, that 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 obtained from a simple physical model [14].

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 [15]. 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}, that represents the value of the weight in the link between i and j (0 if no link exists), as [15]

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

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 $w_i = \sum_j w_{ij}$, and the total strength $2w = \sum_i w_i = \sum_i \sum_j w_{ij}$.

The larger the modularity the best the partitioning is, cause 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 are equal to the Bell [16] or exponential numbers, which grow at least exponentially in the number of nodes N. Indeed, optimization of modularity is a NP-hard (Non-deterministic Polynomial-time hard) problem [17]. Sev-

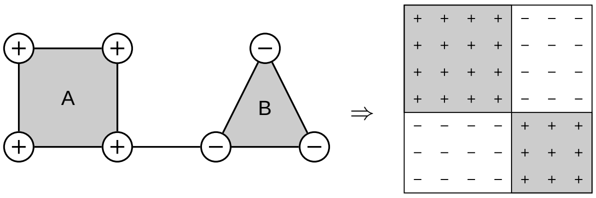

FIG. 1: Network with well-defined community structure and its correlation matrix.

eral authors have attacked the problem proposing different optimization heuristics [18–23].

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, $R_{ij} = S_i S_j$ 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 $R_{ij} = \frac{w_i w_j}{2w}$ 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

$$p_i p_j = \frac{w_i w_j}{(2w)^2}\,. \tag{3}$$

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

$$\begin{aligned} w_{ij}^{+} &= \max\{0, w_{ij}\}\,, &(5)\\ w_{ij}^{-} &= \max\{0, -w_{ij}\}\,. &(6) \end{aligned}$$

The positive and negative strengths are given by

$$\begin{aligned} w_i^{+} &= \sum_j w_{ij}^{+}\,, &(7)\\ w_i^{-} &= \sum_j w_{ij}^{-}\,, &(8) \end{aligned}$$

and the positive and negative total strengths by

$$\begin{aligned} 2w^{+} &= \sum_i w_i^{+} = \sum_i \sum_j w_{ij}^{+}\,, &(9)\\ 2w^{-} &= \sum_i w_i^{-} = \sum_i \sum_j w_{ij}^{-}\,. &(10) \end{aligned}$$

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

$$\begin{aligned} p_i^{+} &= \frac{w_i^{+}}{2w^{+}}\,, &(13)\\ p_i^{-} &= \frac{w_i^{-}}{2w^{-}}\,. &(14) \end{aligned}$$

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

$$\begin{aligned} Q^{+} &= \frac{1}{2w^{+}} \sum_i \sum_j \left(w_{ij}^{+} - \frac{w_i^{+} w_j^{+}}{2w^{+}} \right) \delta(C_i, C_j)\,, &(15)\\ Q^{-} &= \frac{1}{2w^{-}} \sum_i \sum_j \left(w_{ij}^{-} - \frac{w_i^{-} w_j^{-}}{2w^{-}} \right) \delta(C_i, C_j)\,. &(16) \end{aligned}$$

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

$$Q = \frac{2w^+}{2w^+ + 2w^-} Q^+ - \frac{2w^-}{2w^+ + 2w^-} Q^- \,. \qquad (17)$$

An alternative equivalent form for modularity Q is

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

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. directed $Q(C, \{w_{ij}\}) = -Q(C, \{-w_{ij}\})$ .The extension to directed networks is simply obtained by the substitutions

$$w_i^{\pm} \;\rightarrow\; w_i^{\pm,\mathrm{in}} = \sum_k w_{ki} \,, \qquad (19)$$

$$w_j^{\pm} \;\rightarrow\; w_j^{\pm,\mathrm{out}} = \sum_k w_{jk} \,. \qquad (20)$$

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 [14] 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 [24] 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

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

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

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

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 neighbourhood 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 Montecarlo sampling, are taken into account in the adjacency matrix.

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

	original modularity	new modularity
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 [25]. We perform a comparison between the different partitions obtained optimizing Eq. (1) (4 communities) and Eq. (18) (6 communities), against the Lyon's Commerce Chamber retail activities classification (9 communities), in terms of the Rand Index [26], Jaccard Index [27], and normalized mutual information (NMI) [28] (see Table I). All indices show a better performance of Eq. (18) discriminating the actual communities provided by the Lyon's Commerce Chamber.

Once the best partition has been obtained, we analyze the role of different retail stores within communities using the z-score [20]. The basic idea consists in to compute the z-score (Z) 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

$$Z_i^{\pm,\mathrm{in/out}} = \frac{w_i^{\pm,\mathrm{in/out}} - w_{\mathrm{int}}^{\pm,\mathrm{in/out}}}{\sigma_{\mathrm{int}}^{\pm,\mathrm{in/out}}} \,, \qquad (23)$$

where subindices 'int' express averages restricted to the community to which node i belongs; and 'in/out' refer to the direction of links.

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

TABLE II: Roles of retailers within communities.

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

community found (containing 33 retail stores). It is very significant the situation found for gas stations, the data tell us that gas retailers tend to have their location close to the rest of retailers in the community, while retailers do not want to have a gas station close to them. The case of sport facilities is also interesting to mention, they tend to have their location close to the rest of retailers and at the same time are very welcomed to be close. Dairy products shops and cake shops, tend to isolate from the rest of retailers, and Flea markets are repelled by the retailers within the community. Curiously, funeral services are centrally situated in the city and are welcomed by the retailers of its community.

Summarizing, we have proposed a new 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 auto-correlation, or considering the unsigned value of the correlations. We have analyzed within the scope of the new modularity an interesting model of attraction-repulsion of retail stores in a large city, previously reported in [14]. 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.

*Note added.*After this work was finished, the authors became aware of a recent preprint [**?**] of h?1T were reported. In the constituent quark model [91] and in the covariant model [93] the relation Eq. (36) was found to be satisfied, see Refs. [99,100]. In a variant of the spectator model it was found invalid [101]. It would be interesting to formulate the general conditions a quark model must satisfy such that the relation (36) holds.

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

[1] S. H. Strogatz, *Nature* **410**, 268 (2001).

[2] C. M. Song, S. Havlin, H. A. Makse, *Nature* **433**, 392 (2005).

[3] A.-L. Barabási, *Science* **308**, 639 (2005)

[4] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U., *Phys. Rep.* **424**, 175-308 (2006).

[5] M. Girvan, M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **99**, 7821 (2002).

[6] R. Guimerà, L. A. N. Amaral, *Nature* **433**, 895 (2005).

[7] M. E. J. Newman, M. Girvan, *Phys. Rev. E* **69**, 026113 (2004).

[8] M. E. J. Newman, *Phys. Rev. E* **70**, 056131 (2004).

[9] A. Arenas, A. Fernández, S. Gómez, *New J. Phys.* **10**, 053039 (2008).

[10] E. A. Leicht and M. E. J. Newman, *Phys. Rev. Lett.* **100**, 118703 (2008).

[11] P. X.-K Song, Correlated Data Analysis: Modeling, Analytics, and Applications, (Springer Series in Statistics, New York, 2007).

[12] T. Heimo, J. Saramäki, J.-P. Onnela, and K. Kaski, *Physica A* **383**, 147-151 (2007).

[13] T. Heimo, G. Tibély, J. Saramäki, K. Kaski, and J. Kertész, *Physica A* **387**, 5930-5945 (2008).

[14] P. Jensen, *Phys. Rev. E* **74**, 035101 (2006)

[15] M. E. J. Newman, *Phys. Rev. E* **70**, 056131 (2004).

[16] E. T. Bell, *Amer. Math. Monthly* **41**, 411 (1934).

[17] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, D. Wagner, *IEEE Transactions on Knowledge and Data Engineering* **20(2)**, 172 (2008).

[18] M. E. J. Newman, *Phys. Rev. E* **69**, 066133 (2004).

[19] A. Clauset, M. E. J. Newman, C. Moore, *Phys. Rev. E* **70**, 066111 (2004).

[20] R. Guimerà, L. A. N. Amaral, *J. Stat. Mech.*, P02001 (2005).

[21] J. Duch, A. Arenas, *Phys. Rev. E* **72**, 027104 (2005).

[22] J. M. Pujol, J. Béjar, J. Delgado, *Phys. Rev. E* **74**, 016107 (2006).

[23] M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **103**, 8577 (2006).

[24] G. Duranton, H. G. Overman, *The Review of Economic Studies* **72**, 1077 (2005).

[25] L. Danon, A. Díaz-Guilera, J. Duch, A. Arenas, *J. Stat. Mech.*, P09008 (2005).

[26] W. M. Rand, *J. Am. Stat. Assoc.* **66**, 846 (1971).

[27] P. Jaccard, *The New Phytologist* **11(2)**, 37 (1912).

[28] A. Strehl, J. Ghosh, *J. Machine Learning Research* **3**, 583 (2002).

[] V. A. Traag and J. Bruggeman, arXiv:0811.2329 (2008)

[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.