Comparing different modularization criteria using relational metric

P. Conde Céspedes[1] and J.F. Marcotorchino[2]

[1] Laboratoire de Statistique théorique et Appliquée (LSTA), Université Pierre et Marie Curie, Paris, France
patricia.conde_cespedes@upmc.fr
[2] Thales Communications et Sécurité, TCS, Gennevilliers, France and (LSTA) Université Pierre et Marie Curie, Paris, France
jeanfrancois.marcotorchino@thalesgroup.com

Abstract. In this paper we use the relational metric to represent some linear modularization criteria such as Newman-Girvan, Zahn-Condorcet and Owsiński- Zadrożny. The relational coding allows us to compare and deduce the properties of those criteria. Furthermore, we introduce two modularization criteria: the balanced-modularity and the Deviation to indetermination Index. The first one based on the Newman-Girvan modularity and the second one based on the "deviation from indetermination" structure. The partitions obtained with all the criteria are tested using the generalized Louvain algorithm.

Keywords. Graph, modularization, modularization function, modularity, clustering, communities, Mathematical Relational Analysis.

1 Introduction

Nowadays, the increasing use of social networks has considerably reinforced their complexity. Furthermore, networks can be found in various contexts such as biology, computer programming, marketing, etc. Graphs are mathematical representations of networks, where the entities are called nodes and the connections are called edges.

It is difficult to analyze directly complex networks because of their big size. Therefore, we need to divide it in smaller components easy to handle. The process of splitting a network has received different names: graph clustering (in data analysis) or modularization; depending on the context, the clusters can be called communities, modules or clusters.

To evaluate the decomposition in clusters of a network, it is necessary to dispose of a modularization criterion to optimize. All criteria differ in the definition given to the notion of *community*.

We use the Mathematical Relational Analysis (ARM)[3] to represent linear modularization criteria as a linear optimization problem subject to linear constraints forcing the output to represent an equivalence relation (a clustering in ARM language). This model allows to compare different criteria in order to understand their properties. We study mainly three criteria: Newman-Girvan, Zahn-Condorcet and Owsiński- Zadrożny. We introduce as well two modularization criteria: the balanced modularity and the Deviation to indetermination index. The first one is derived from the Newman-Girvan modularity and the second one is based on the "deviation from indetermination" structure.

In section 2 we present a brief summary of the Mathematical Relational Analysis (MRA) approach; section 3 exposes diverse linear modularization criteria in relational notations; section 4 introduces two new modularity criteria: the balanced modularity and the deviation from indetermination structure index.

2 Relational Analysis approach

There is a strong link between the Mathematical Relational Analysis[4] and graph theory: *A graph is a mathematical structure that represents binary relations between objects belonging to the same set.* Therefore, a non-oriented and non-weighted graph $G = (V, E)$, with $N = |V|$ nodes and $M = |E|$ edges, is a binary symmetric relation on its set of nodes V represented by its adjacency matrix $\mathbf{A}$ as follows:

$$a_{ii'} = \begin{cases} 1 & \text{if there exists an edge between } i \text{ and } i' \ \forall (i, i') \in V \times V \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Partitioning a graph is nothing else than defining an equivalence relation on the set of nodes V, that means a symmetric, reflexive and transitive relation. Mathematically, an equivalence relation is represented by a square matrix $\mathbf{X}$ of order $N = |V|$, whose entries are defined as follows:

$$x_{ii'} = \begin{cases} 1 & \text{if } i \text{ and } i' \text{ are in the same cluster } \forall (i, i') \in V \times V \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Modularizing a graph implies to define $\mathbf{X}$ as close as possible to $\mathbf{A}$. A modularity criterion is a function which measures either a *similarity* or a distance between $\mathbf{A}$ and $\mathbf{X}$. Therefore, the problem of modularization will be written as a function to optimize in the general form:

[3] Analyse Relationnelle Mathématique in French (ARM).
[4] For more details about Relational Analysis theory see [MAM79], [MAR84], [MIC87], [MAY91], [MAR91].

$$\underset{X}{Max}(F(A,X)) \tag{3}$$

subject to the constraints of an equivalence relation:

$$\begin{aligned} x_{ii'} &\in \{0,1\} && \text{Binary} \\ x_{ii} &= 1 \quad \forall i && \text{Reflexivity} \\ x_{ii'} - x_{i'i} &= 0 \quad \forall (i,i') && \text{Symmetry} \\ x_{ii'} + x_{i'i''} - x_{ii''} &\leq 1 \ \forall (i,i',i'') && \text{Transitivity} \end{aligned} \tag{4}$$

The exact solving of this $0-1$ linear program due to de size of the constraints for big networks, (for example facebook has more than one billion users in march 2013), is impossible. So, heuristic approaches are the only reasonable way to proceed. In particular, to simplify the complexity, linear criteria are suitable.

We define as well $\bar{\mathbf{X}}$ and $\bar{\mathbf{A}}$ as the inverse relation of $\mathbf{X}$ and $\mathbf{A}$ respectively. Their entries are defined as $\bar{x}_{ii'} = 1 - x_{ii'}$ and $\bar{a}_{ii'} = 1 - a_{ii'}$ respectively.

The partitions obtained by each criterion differ according to the properties the criterion verifies. In this paper we consider two properties: *linearity* and *separability*. The relational codings of these properties are shown in table 1.

Table 1. Properties verified by modularity criteria

The criterion has the property	**If it can be written as**
Linearity	$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} a_{ii'} x_{ii'} + K$
Separability	$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) \psi(x_{ii'}) + K$

In this article we will use K to denote any *constant* depending only on the original data.

According to Table 1, the property of linearity entails that the criterion is an affine function of $\mathbf{X}$.
The function $\phi(a_{ii'})$ depends only on the original data (i.e. the adjacency matrix). $\psi(x_{ii'})$ is a function of the unknown variable $x_{ii'}$. The property of separability implies that the criterion can be written as a scalar product of two vectors, the first one depending only upon the original data and the second one depending upon the unknown variable. Consequently, every linear criterion is separable. The property of separability entails that the criterion separates the variable part from the data part.

3 Modularization criteria in relational notation

Graph clustering criteria depend strongly on the meaning given to the notion of *community*. Various modularization criteria have been defined in different fields, each one having its own definition of *community*. However, all definitions have something in common: *dense connections within the community and only sparse connections between communities*. That is, community detection is possible only if the graph is dense. In this section, we present some linear modularization criteria in relational coding, this notation will help us compare those criteria and identify their main properties. The relational coding of all linear criteria is given in table 2.

1. **The Zahn-Condorcet criterion (1785, 1964)**: C.T. Zahn (see [ZAH64]) was the first author who studied the problem of finding an equivalence relation **X**, which best approximates" a given symmetric relation **A** in the sense of minimizing the distance of the symmetric difference. However the criterion defined by Zahn corresponds to the dual Condorcet's criterion (see [CON1785]) introduced in his on Relational Consensus and whose relational coding is given in [MAM79].
   This criterion requires that every node in each cluster be connected to at least as half as the total nodes inside the cluster. Consequently, the clustering coefficient of each cluster is greater than 50%.

2. **The Owsiński-Zadrożny criterion (1986)** (see [OWZ86]) it is a generalization of Condorcet's function. It is more flexible because it has a parameter α, which allows the user, according to the context, to define the minimal percentage of required within-cluster edges: α. Everything depends on the definition given to the notion of *community*. For $\alpha = 0.5$ this Owsiński - Zadrożny function is equal to Condorcet's criterion multiplied by 0.5.

3. **The Demaine-Immorlica or the Correlation clustering criterion (2002):** The problem of correlation clustering was introduced by [BAB02]. Later, in [DMI03] the authors formulated the quality function to optimize this problem. This criterion considers a graph with real non-negative edge weights labeled + and −, the purpose is to partition the nodes into clusters to minimize the total cut of + edges and uncut − edges. Considering the relational notation of this criterion in table 2, it is easy to remark that it is a variant of Condorcet's criterion for a weighted graph with positive and negative weights.

4. **The Newman-Girvan criterion (2004)** (see [NEW04]): It is the best known modularization criterion, called sometimes simply *modularity*. Its definition involves a comparison of the number of within-community edges in a real network and the expected number of such edges in a random graph (without regard to community structure). In fact, the *modularity* maximizes the deviation to independence, that is, the numerator of the well known χ^2

index or the Belson's index (see [BEL59]) in contingency theory.
The relational notation made the number of clusters of the optimal partition disappear from the original formulation. The *modularity* is a null model, that means that it is null if all the nodes are in the same cluster. It has been shown as well (see [BRA08]) main disadvantages of Newman-Girvan modularity, such as non-locality and resolution limit (see [FOB07]), besides that, in (see [DEM12]), the authors showed that by clustering regular graphs such as lattice or grid graph, who do not have community structure at all it is possible to get a value of *modularity* asymptotically equal to 1.

Criterion	**Relational notation**
Zahn-Condorcet (1785, 1964)	$F_{ZC}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'}x_{ii'} + \bar{a}_{ii'}\bar{x}_{ii'})$
Owsiński - Zadrożny(1986)	$F_{Z_OZ}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}((1-\alpha)a_{ii'}x_{ii'} + \alpha\bar{a}_{ii'}\bar{x}_{ii'})$ with $0 < \alpha < 1$
Demaine-Immorlica (2002)	$F_D(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(w^+_{ii'}x_{ii'} + w^-_{ii'}\bar{x}_{ii'})$
Newman-Girvan (2004)	$F_{NG}(X) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}\right)x_{ii'}$

Table 2. Relational notation of linear modularity functions.

Table 2 shows that Zahn, Owsiński-Zadrożny and Demaine-Immorlica criteria are variants of Condorcet's criterion (1785). The Owsiński-Zadrożny criterion distinguishes from Condorcet's criterion by the parameter α used to define the importance of the positive agreements part and of the negative agreements part. The Demaine's criterion (correlation clustering) differs from Condorcet's criterion in the type of input data, whereas Condorcet's criterion treats binary data, the correlation clustering criterion treats real data.

4 Two new modularization criteria

1. **The balanced modularity** This criterion is a balanced version of the Newman-Girvan modularity. In fact, it was constructed by adding to the Newman-Girvan modularity a term taking into account the absence of edges $\bar{\mathbf{A}}$. The *balanced modularity* is given by the following formula in relational notation:

$$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}\right)x_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\bar{a}_{ii'} - \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}\right)\bar{x}_{ii'} \tag{5}$$

Whereas Newman-Girvan modularity compares the actual value of $a_{ii'}$ to its equivalent in the case of a random graph $\frac{a_{i.}a_{.i'}}{2M}$, the new term compares the value of $\bar{a}_{ii'}$ to its version in case of a random graph $\frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}$.

2. **The Deviation to indetermination index**
Analogously to Newman-Girvan function, which maximizes the deviation to the independence structure; this new criterion maximizes the deviation to the indetermination structure (see [JAV82], [MAR84], [MAR85] and [AHM07]). The expression of the Deviation to indetermination index is written as follows:

$$F_{DI}(X) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{2M}{N^2}\right)x_{ii'} \tag{6}$$

Analogously to Newman-Girvan modularity, criterion (6) is also a null model, because $\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{2M}{N^2}\right) = 0$.

5 Applications

The partitions obtained with all criteria are tested using the generalized Louvain algorithm (see [BLO08]) with real networks. In this text, we present in detail only the results obtained with "the College football network" (see [GIN02]). During the talk more examples with real networks will be presented.

The *college football* network is a representation of the schedule of games for the 2000 season. Nodes in the graph represent teams (N=115) and edges represent regular season games between two teams ($2M = 1226$ in total). This network has the advantage of incorporating a known community structure because the teams are divided into 12 *conferences*. Games are more frequent between members of the same conference.

To compare the partitions obtained with different criteria we calculate the *percentage of agreements with the conference partition* ρ_{agree} as follows:

$$\rho_{agree} = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N}(y_{ii'}x_{ii'} + \bar{y}_{ii'}\bar{x}_{ii'})}{N^2} \tag{7}$$

where $\mathbf{Y}$ and $\mathbf{X}$ represent the relational matrix of the partition in *conferences* and the relational matrix of the partition found by optimizing the criterion respectively. Table 3 shows the results obtained with all criteria.

The four criteria identify the conference structure with a high percentage of agreements. However, the number of clusters vary from the Zahn-Condorcet criterion to the three others. In fact, Zahn-Condorcet function splits the *conferences* (groups) in subgroups. It is interesting to remark that the percentage of

Criterion	**Number of clusters** κ	**Clusters correctly identified**	ρ_{agree} (%)
Newman-Girvan	10	6	96,9%
Zahn-Condorcet	16	7	97,7%
Balanced modularity	10	6	96,9%
Deviation to indetermination Index	10	6	96,9%

Table 3. Partitions obtained by clustering the "College football" network.

agreements is the highest for this quality function. The partitions found with the two new criteria are exactly the same as that obtained with the Newman-Girvan criterion (although their definitions are quite different).

By analyzing the partitions found with other real networks such as a "the collaboration network of jazz musician" with $N = 198$ nodes and $M = 2742$ edges (see [GLD03]) or a big sub-network of the internet with $N = 69949$ nodes and $M = 351380$ edges (see [HOM03]) we found that the number of clusters found by the 5 criteria differs as N tends to the infinity. The Zahn-Condorcet criterion generates many clusters with a single node. The partitions found with the two new criteria are nearly the same as that found by Newman-Girvan modularity. However there are small differences. Concerning the Balanced modularity we remarked that nodes with small degree can easily join the cluster of their neighbors if they are clustered with other nodes of small degree, however if their neighbors are clustered with nodes of high degree the criterion separates nodes with small degree. In contrast, the partitions obtained with Newman-Girvan function do not have clusters containing only a single node. Concerning the deviation to the indetermination structure this criterion favors big clusters with high average degree and small clusters with low average degree.

References

AHM07. **Ah-Pine J., Marcotorchino J.F.**: "*Statistical, geometrical and logical independences between categorical Variables*", Proceedings of the Applied Stochastic Models and Data Analysis ASMDA2007 Symposium, Chania, Greece (2007).

BAB02. **Bansal N., Blum A. and Chawla S.**: "*Correlation Clustering*", MACHINE LEARNING, pp. 238–247, (2002).

BEL59. **Belson W.**: "*Matching and Prediction on the Principle of Biological Classification*", Survey Research Centre, London School of Economics and Political Science, 1959.

BLO08. **Blondel V., Guillaume J.L., Lambiotte R. and Lefebvre E.**: *"Fast unfolding of communities in large networks"*, Journal of Statistical Mechanics: Theory and Experiment, 2008.

BRA08. **Brandes U., Delling D., Gaertler M. Gőrke R., Hoefer M., Nikoloski Z. and Wagner D.**: *On Modularity Clustering*, IEEE Transactions on Knowledge and Data Engineering, vol. 20, pp. 172-188, (2008).

CON1785. **Caritat A. Marquis de Condorcet**:*Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, L'imprimerie royale, Paris, France, 1785.

DEM12. **De Montgolfier F., Soto M., Viennot L.**:*Modularité asymptotique de quelques classes de graphes*, 14èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (AlgoTel), pp. 1-4, (2012).

DMI03. **Demaine E. and Immorlica N.**: "*Correlation clustering with partial information*", In Proceedings of the 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, pp. 1–13, (2003).

FOB07. **Fortunato S. and Barthelemy M.**: *Resolution limit in community detection*, Proceedings of the National Academy of Sciences, 2007.

GLD03. **Gleiser P. and Danon L.**, *Community structure in jazz*. Preprint cond-mat/0307434 (2003).

HOM03. **Hoerdt M. and Magoni D.**, Proceedings of the 11th International Conference on Software, Telecommunications and Computer Networks 257, (2003).

JAV82. **Janson, S. and Vegelius, J.**, "*The J- Index as a Measure of Association For Nominal Scale Response Agreement*", Applied psychological measurement, 1982.

MAM79. **Marcotorchino F., Michaud P.**: "*Optimisation en Analyse Ordinale des Données*", Book by Masson pp :1- 211, (1979).

MAR84. **Marcotorchino F.** : *Utilisation des Comparaisons par Paires en Statistique des Contingences (Partie I)*, Publication du Centre Scientifique IBM de Paris, F057, pp : 1-57, Paris et Cahiers du Séminaire Analyse des Données et Processus Stochastiques Université Libre de Bruxelles , Bruxelles, (1984).

MAR85. **Marcotorchino F.**: *Utilisation des Comparaisons par Paires en Statistique des Contingences* (Partie III), Publication du Centre Scientifique IBM de Paris, F081, pp : 1-39, (1985).

MAR91. **Marcotorchino F.**: "*Seriation Problems:an overview*", Applied Stochastic Models and Data Analysis, Vol:7, n2, pp:139-151, (1991).

MAR13. **Marcotorchino F., Conde Cespedes P.**: "*Optimal Transport, Spatial Interaction Models and related Problems, impacts on Relational Metrics, adaptation to Large Graphs and Networks Modularity*" (2013).

MAY91. **Marcotorchino F., El Ayoubi N.** : "*Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association*", Revue de Statistique Appliquée, Vol :39, n2, pp:25-46 (1991).

MIC87. **Michaud P.** : "*Condorcet, a man of the avant garde*", Journal of Applied Stochastic Models and Data Analysis, Vol:3, n2, (1997).

NEW04. **Newman M. E. J. and Girvan M.**:"*Finding and evaluating community structure in networks*", Journal of Phys. Rev. E, vol. 69, (2004).

GIN02. **Girvan, M. and Newman, M. E. J.**: "*Community structure in social and biological networks*", Proceedings of the National Academy of Sciences of the United States of America, number 12, vol. 99, 7821–7826, (2002).

OWZ86. **Owsiński, Jan.W. and Zadrożny, Sławomir** (1986): *Clustering for ordinal data: a linear programming formulation.*. Control and Cybernetics, vol. 15, pp. 183-193

ZAH64. **Zahn, C.T.**(1964): *Approximating symmetric relations by equivalence relations*. SIAM Journal on Applied Mathematics, 12, 840–847.