# Comparison of linear modularization criteria using the relational formalism, an approach to easily identify resolution limit

P. Conde-Céspedes, J. F. Marcotorchino and E. Viennet

**Abstract** The modularization of large graphs or community detection in networks is usually approached as an optimization problem of a quality function or criterion, for instance, the modularity of Newman-Girvan. There exist other clustering criteria, with their own properties leading to different solutions. In this paper we present six linear modularization criteria in relational notation such as the Newman-Girvan modularity, Zahn-Condorcet, Owsiński- Zadrożny, the Deviation to Uniformity index, the Deviation to Indetermination index and the Balanced-Modularity. We use a generic version of Louvain algorithm to approach the optimal partition of the criteria with real networks of different sizes. We have found that those partitions present important differences concerning the number of clusters. The relational formalism allows us to justify these differences from a theoretical point of view. Moreover, this notation enables to easily identify the criteria having a resolution limit (a phenomenon which causes the criterion to fail to identify modules smaller than a given scale). This finding is confirmed in artificial benchmark LFR graphs.

Patricia Conde-Céspedes
L2TI - Institut Galilée - Université Paris 13
99, av. Jean-Baptiste Clément; 93430 Villetaneuse - France
e-mail: patricia.conde_cespedes@univ-paris13.fr

Jean-François Marcotorchino
Thales Communications et Sécurité
4 av. des Louvresses; 92230 Gennevilliers - France
e-mail: jeanfrancois.marcotorchino@thalesgroup.com

Emmanuel Viennet
L2TI - Institut Galilée - Université Paris 13
99, av. Jean-Baptiste Clément; 93430 Villetaneuse - France
e-mail: emmanuel.viennet@univ-paris13.fr

# 1 Introduction

Networks are studied in numerous contexts such as biology, sociology, online social networks, marketing, etc. Graphs are mathematical representations of networks, where the entities are called nodes and the connections are called edges. Very large graphs are difficult to analyse and it is often profitable to divide them in smaller homogeneous components easier to handle. The process of decomposing a network has received different names: graph clustering (in data analysis), modularization, community structure identification. The clusters can be called communities or modules; in this paper we use those words as synonyms.

Assessing the quality of a graph partition requires a modularization criterion. This function will be optimized to find the best partition. Various modularization criteria were formulated in the past to address different practical applications. Those criteria differ in the definition given to the notion of community or cluster.

To understand the differences between the optimal partitions obtained by each criterion we show how to represent them using the same basic formalism. In this paper we use the Mathematical Relational Analysis (MRA) to express six linear modularization criteria. Linear criteria are easy to handle, for instance, the Louvain method can be adapted to linear quality functions (see [Campigotto et al., 2014]). The six criteria studied are: the Newman-Girvan modularity, the Zahn-Condorcet criterion, the Owsiński-Zadrozny criterion, the Deviation to Uniformity, the Deviation to Indetermination index and the Balanced Modularity (details in section 3). The relational representation makes clear the properties of those modularization criteria. It allows to easily identify the criteria suffering from a resolution limit, first discussed by [Fortunato and Barthelemy, 2006]. We will complete this theoretical study by some experiments on real and synthetic networks, demonstrating the effectiveness of our classification.

In this paper, we deal only with linear criteria. Nevertheless, it is important to mention that thanks to the formalism of the MRA it is also possible to express non-linear criteria in relational notations. For instance, we can mention some very well-known criteria such as the Mancoridis-Gansner criterion (see [Mancoridis et al., 1998]) in cluster-programming, the Ratio-Cuts by [Wei and Cheng, 1989], the Michalski criterion (see [Michalski and Stepp, 1983] and its relational notation given in [Decaestecker, 1992]), etc. The interested reader can see [Conde-Céspedes and Marcotorchino, 2012] and [Conde-Céspedes, 2013].

This paper is organized as follows : Section 2 presents the Mathematical Relational Analysis approach and introduces the property of *balance* for linear criteria and its relation to the property of resolution limit. In Section 3, six linear modularization criteria in the relational formalism are formulated. Next, Section 4 discusses some experiments on real and artificial graphs to confirm the theoretical properties found previously.

## 2 Relational Analysis approach

There is a strong link between the Mathematical Relational Analysis[1] 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:

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

We denote the *degree* $d_i$ of node $i$ the number of edges incident to $i$. It can be calculated by summing up the terms of the row (or column) $i$ of the adjacency matrix: $d_i = \sum_{i'} a_{ii'} = \sum_{i'} a_{i'i} = a_{i.} = a_{.i}$. We denote $\delta = \frac{2M}{N^2}$ the density of edges of the whole graph.

Partitioning a graph implies 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:

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

Modularizing a graph implies to find $\mathbf{X}$ as close as possible to $\mathbf{A}$. A modularization criterion $F(X)$ is a function which measures either a *similarity* or a distance between $\mathbf{A}$ and $\mathbf{X}$. Therefore, the problem of modularization can be written as a function to optimize $F(X)$ where the unknown $X$ is subject to the constraints of an equivalence relation. In fact, the problem of modularization can be written in the general form:

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

subject to the constraints of an equivalence relation:

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

The exact solving of this $0-1$ linear program due to the size of the constraints is impractical for big networks. So, heuristic approaches are the only reasonable way

[1] For more details about Relational Analysis theory see [Marcotorchino and Michaud, 1979] and [Marcotorchino, 1984].

to proceed.

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. In the following we denote $\kappa$ the optimal number of clusters, that means the number of clusters of the partition $\mathbf{X}$ which maximizes the criterion $F(X)$.

## 2.1 Linear balanced criteria

Every linear criterion is an affine function of $\mathbf{X}$, therefore in relational notation it can be written as:

$$F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \Phi(a_{ii'})x_{ii'} + K, \tag{4}$$

where $\Phi(a_{ii'})$ denotes any function depending only on the original data (for instance the adjacency matrix) and $K$ denotes any *constant* depending only on the original data. Therefore, $K$ does not intervene in the optimization problem.

**Definition 1 (Property of linear balance).** A linear criterion is *balanced* if it can be written in the following general form:

$$F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \phi(a_{ii'})x_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{\phi}(a_{ii'})\bar{x}_{ii'} + K. \tag{5}$$

where $\phi(.)$ and $\bar{\phi}(.)$ are non negative functions depending only on the original data and verifying $\sum_{i=1}^{N}\sum_{i'=1}^{N}\phi_{ii'} > 0$ and $\sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{\phi}_{ii'} > 0$.
So, they can not be all null simultaneously.

By replacing $\bar{x}$ by its definition $1 - x_{ii'}$, equation (5) can be rewritten as follows:

$$F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} (\phi_{ii'} - \bar{\phi}_{ii'})x_{ii'} + K. \tag{6}$$

### 2.1.1 Interpretation of functions $\phi(.)$ and $\bar{\phi}(.)$

At this point, we can give the intuition behind functions $\phi(.)$ and $\bar{\phi}(.)$. From expression (6) we deduce the importance of the property of *balance* for linear criteria. If the criterion is a function to maximize, the presence and/or absence of the terms $\phi_{ii'}$

and $\bar{\phi}_{ii'}$ has the following impact on the optimal solution:

- If $\bar{\phi}_{ii'} = 0 \forall i, i'$ the solution that maximizes $F(X)$ is the partition where all nodes are clustered together in a single cluster, so $\kappa = 1$ and $x_{ii'} = 1 \quad \forall (i,i')$ and $F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi_{ii'}$.
- If $\phi_{ii'} = 0 \forall i, i'$ then the optimal solution that maximizes $F(X)$ is the partition where all nodes are separated, so $\kappa = N$ and $x_{ii'} = 0 \forall i \neq i'$ and $x_{ii} = 1 \forall i$ therefore $F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}_{ii}$.

In other words, the optimization of a linear criterion who does not verify the property of balance will either *cluster all the nodes in a single cluster* or *isolate each node in its own cluster*, therefore forcing the user to fix the number of clusters in advance.

We can deduce from the previous paragraphs that the values taken by the functions $\phi$ and $\bar{\phi}$ create a sort of *balance* between the fact of generating as many clusters as possible, $\kappa = N$, and the fact generating only one cluster, $\kappa = 1$.

In the following we will call the quantity $\sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) x_{ii'}$ the term of *positive agreements* and the quantity $\sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}(a_{ii'}) \bar{x}_{ii'}$ the term of *negative agreements*.

## 2.2 Different levels of balance

We define two levels of balance for all linear balanced criterion:

**Definition 2 (Property of local balance).** A balanced linear criterion whose functions $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ depend only upon the pair $(i,i')$ (therefore not depending on global properties of the graph) has the property of local balance.

Some remarks about definition 2:

- When we talk about global properties we refer to the total number of nodes, the total number of edges or other properties describing the global structure of the graph.
- For the particular case of local balance where $\phi_{ii'} + \bar{\phi}_{ii'} = K$ (that is $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ sum up to a constant), we can conclude that whereas $\phi_{ii'}$ increases $\bar{\phi}_{ii'}$ decreases and vice versa.

Let us consider the special case where $\phi(a_{ii'}) = a_{ii'}$, the general term of the adjacency matrix. A *null model* is a graph with the same total number of edges and nodes and where the edges are randomly distributed. Let us denote the general term of the adjacency matrix of this random graph $\bar{\phi}(a_{ii'})$. A criterion based on a null model considers that a random graph does not have community structure. The goal of such a criterion is to maximize the deviation between the real graph, represented by $\phi(a_{ii'})$ and the null model version of this graph, represented by $\bar{\phi}(a_{ii'})$ as shown in equation (6). Since the original graph and the null model have the same number of edges $M$, we have $\sum_{i=1}^{N}\sum_{i'=1}^{N}\phi_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{\phi}_{ii'} = 2M$. If this constraint causes $\bar{\phi}_{ii'}$ to depend upon the total number of edges $M$, then a criterion based on a null model does not verify the property of local balance. Consequently, it is not scale invariant because it depends on a global characteristic of the graph.

The definition of null model for linear criteria can be generalized as follows:

**Definition 3 (Criterion based on a null model).** A balanced linear criterion that seeks to maximize the deviation between the real graph and a null model is a criterion based on a null model. In its formulation, the real graph is represented by $\phi(a_{ii'})$ whereas the null model is represented by $\bar{\phi}(a_{ii'})$. The functions $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ satisfy the following condition:

$$\sum_{i=1}^{N}\sum_{i'=1}^{N}\phi_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{\phi}_{ii'}$$

such that the functions $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ depend on global properties of the graph.

The global properties of the graph can be, for example, the total number of edges or the total number of nodes.

We can deduce from definitions 2 and 3 that a linear criterion cannot be locally balanced and based on a null model at the same time.

In the particular case where $\bar{\phi}$ decreases with the size of the network, it becomes negligible for large graphs. As explained previously, if this term tends towards zero, the optimization of the criterion will tend to group the nodes more easily. For instance, a single edge between two sub-graphs would be interpreted by the criterion as a sign of a strong correlation between the two clusters, and optimizing the criterion would lead to the merge of the two clusters. Such a criterion is said to have a *resolution limit*.

The resolution limit was introduced by [Fortunato and Barthelemy, 2006], where the authors studied the resolution limit of the modularity of Newman-Girvan. They demonstrated that modularity optimization may fail to identify modules smaller than a given size which depends on global characteristics of the graph. Even weakly interconnected complete sub-graphs – the best identifiable communities – would be merged by this kind of optimization criteria if the network is sufficiently large. According to [Kumpula et al., 2007] the resolution limit is present in any modularization criterion based on global optimization of intra-cluster edges and extra-community links and on a comparison to any null model.

In section 4, we will show how criteria having a resolution limit fail to detect certain groups of densely connected nodes.

## 3 Modularization criteria in relational notation

Graph clustering criteria depend strongly on the meaning given to the notion of *community*. In this section, we describe six linear modularization criteria and their relational coding in table 1. We assume that the graphs we want to modularize are scale-free, that means that their degree distribution follows a power law.

1. **The Zahn-Condorcet criterion (1785, 1964)**: C.T. Zahn 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 in [Zahn, 1964]. However the criterion defined by Zahn corresponds to the dual Condorcet's criterion (see [Condorcet, 1785]) introduced in Relational Consensus and whose relational coding is given in [Marcotorchino and Michaud, 1979]. This criterion requires that every node in each cluster be connected with at least as half as the total nodes inside the cluster. Consequently, for each cluster the fraction of within cluster edges is at least 50% (see appendix and [Conde-Céspedes, 2013] for proof).

2. **The Owsiński-Zadrożny criterion (1986)** (see [Owsiński and Zadrożny, 1986]) it is a generalization of Condorcet's function. It has a parameter $\alpha$, which allows, according to the context, to define the minimal percentage of required within-cluster edges: $\alpha$. For $\alpha = 0.5$ this criterion is equivalent to Condorcet's criterion. The parameter $\alpha$ defines the balance between the positive agreements term and the negative agreements term. For each cluster the density of edges is at least $\alpha\%$ (see [Conde-Céspedes, 2013]).

3. **The Newman-Girvan criterion (2004)** (see [Newman and Girvan, 2004]): It is the best known modularization criterion, called sometimes simply *modularity*. It relies upon a null model. Its definition involves a comparison of the number of within-cluster edges in the real network and the expected number of such edges

in a random graph where edges are distributed following the *independence structure* (a network without regard to community structure). In fact, the *modularity* measures the *deviation to independence*.
As mention in the previous section, this criterion, based on a null model and it has a resolution limit (see [Fortunato and Barthelemy, 2006]). In fact, as the network becomes larger $M \longrightarrow \infty$, the term $\bar{\phi}_{ii'} = \frac{a_{i.}a_{.i'}}{2M}$ tends to zero since the degree distribution follows a power law.

4. **The Deviation to Uniformity (2013)** This criterion maximizes the deviation to the *uniformity structure*, it was proposed in [Conde-Céspedes, 2013]. It compares the number of within-cluster edges in the real graph and the expected number of such edges in a random graph (the null model) where edges are uniformly distributed, thus all the nodes have the same degree equal to the average degree of the graph. This criterion is based on a null model and it has a resolution limit. indeed $\delta \longrightarrow 0$ as $N \longrightarrow \infty$.

5. **The Deviation to Indetermination (2013)** Analogously to Newman-Girvan function, this criterion compares the number of within-cluster edges in the real network and the expected number of such edges in a random graph where edges are distributed following the *indetermination structure*[2] (a graph without regard to community structure), introduced in [Marcotorchino, 2013] and [Marcotorchino and Conde-Céspedes, 2013]. The Deviation to Indetermination is based on a null model, therefore it has a resolution limit.

6. **The Balanced modularity**[3] **(2013)** This criterion, introduced in [Conde-Céspedes and Marcotorchino, 2013], was constructed by adding to the Newman-Girvan modularity a term taking into account the absence of edges $\bar{\mathbf{A}}$. 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 $\frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}$. It is based on a null model and it has a resolution limit.

The six linear criteria of table 1 verify the property of *balance*, so it is not necessary to set in advance the number of clusters. Table 2 specifically focuses on the fonctions $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ for each criterion.

From tables 1 and 2 one can easily deduce that two criteria: Zahn-Condorcet and Owsiński-Zadrożny verify the property of local balance. Furthermore, table 2 clearly shows that the functions $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ add up to a constant $K_{ii'}$ for these two

[2] There exists a duality between the independence structure and the indetermination structure (see [Marcotorchino, 1984], [Marcotorchino, 1985] and [Ah-Pine and Marcotorchino, 2007]).

[3] Although the name of this criterion contains the word *balanced*, its definition is not related to the property of balance given in definition 1.

**Table 1** Relational notation of linear modularity functions.

| 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) | $F_{Z_{OZ}}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} ((1-\alpha) a_{ii'} x_{ii'} + \alpha \bar{a}_{ii'} \bar{x}_{ii'})$<br>with $0 < \alpha < 1$ |
| Newman-Girvan (2004) | $F_{NG}(X) = \frac{1}{2M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( a_{ii'} - \frac{a_{i.} a_{.i'}}{2M} \right) x_{ii'}$ |
| Deviation to Uniformity (2013) | $F_{\text{UNIF}}(X) = \frac{1}{2M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( a_{ii'} - \frac{2M}{N^2} \right) x_{ii'}$ |
| Deviation to Indetermination (2013) | $F_{DI}(X) = \frac{1}{2M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( a_{ii'} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{2M}{N^2} \right) x_{ii'}$ |
| The Balanced Modularity (2013) | $F_{BM}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} ((a_{ii'} - P_{ii'}) x_{ii'} + (\bar{a}_{ii'} - \bar{P}_{ii'}) \bar{x}_{ii'})$<br>where $P_{ii'} = \frac{a_{i.} a_{.i'}}{2M}$ and $\bar{P}_{ii'} = \left( \bar{a}_{ii'} - \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M} \right)$ |

**Table 2** Balance property for linear criteria.

| | General balance | | |
|---|---|---|---|
| **Criterion** | **Local Balance** | **Null model** | **Comment** |
| Zahn-Condorcet | X | | $\phi_{ii'} + \bar{\phi}_{ii'} = a_{ii'} + \bar{a}_{ii'} = 1.$ |
| Owsiński-Zadrożny | X | | $\phi_{ii'} + \bar{\phi}_{ii'} = (1-\alpha) a_{ii'} + \alpha \bar{a}_{ii'}.$ |
| Newman-Girvan | | X | $\sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}_{ii'} = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{a_{i.} a_{.i'}}{2M} = 2M.$ |
| Deviation to Uniformity | | X | $\sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}_{ii'} = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{2M}{N^2} = 2M$ |
| Deviation to Indetermination | | X | $\sum_{i=1}^{N} \sum_{i'=1}^{N} \left( \frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2} \right) = 2M$ |
| Balanced modularity | | X | $\sum_{i,i'=1}^{N} \sum_{i'=1}^{N} \bar{p}_{ii'} = \sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{a}_{ii'} = N^2 - 2M$ |

criteria. The quantity $\bar{\phi}_{ii'}$ decreases with the size of the graph for all criteria that have a resolution limit.

## 4 The impact of merging two clusters

We modularized five real networks of different sizes: Jazz [Gleiser and Danon, 2003], Internet [Hoerdt and Magoni, 2003], Web nd.edu [Albert et al., 1999], Amazon [Yang and Leskovec, 2012][4] and Youtube [Mislove et al., 2007]. We ran a generic version of Louvain Algorithm (see [Campigotto et al., 2014] and [Blondel et al., 2008]) until achievement of a stable value of each criterion. The number of clusters obtained for each network is shown in table 3.

**Table 3** Ref: Zahn-Condorcet (ZC), Owsiński- Zadrożny (OZ), Deviation to Uniformity (UNIF), Newman-Girvan (NG), Deviation to Indetermination(DI) and Balanced Modularity (BM).

| **Network** | **Jazz** | **Internet** | **Web nd.edu** | **Amazon** | **Youtube** |
|---|---|---|---|---|---|
| $N \sim$ | 198 | 70k | 325k | 334k | 1M |
| $M \sim$ | 3k | 351k | 1M | 925k | 3M |
| $\delta$ | 0,14 | $1.44 \times 10^{-04}$ | $2.77 \times 10^{-05}$ | $1.65 \times 10^{-05}$ | $4.64 \times 10^{-06}$ |
| **Criterion** | $\kappa$ | $\kappa$ | $\kappa$ | $\kappa$ | $\kappa$ |
| ZC | 38 | 40,123 | 201,647 | 161,439 | 878,849 |
| OZ $\alpha = 0.4$ | 34 | 30,897 | 220,967 | 121,370 | 744,680 |
| OZ $\alpha = 0.2$ | 23 | 24,470 | 184,087 | 77,700 | 601,800 |
| UNIF | 20 | 173 | 711 | 265 | 51,584 |
| NG | 4 | 46 | 511 | 250 | 5,567 |
| DI | 6 | 39 | 324 | 246 | 13,985 |
| BM | 5 | 41 | 333 | 230 | 6,410 |

Table 3 shows that the Zahn-Condorcet and Owsiński- Zadrożny criteria generate many more clusters than the other criteria having a resolution limit, for which the number of clusters is rather comparable. Moreover, this difference increases with the network size. Notice that the number of clusters for the Owsiński- Zadrożny criterion decreases with $\alpha$, that is the minimal required fraction of within-cluster edges, so the criterion becomes more flexible.

In order to explain these differences we measure the impact of merging two clusters on the value of each criterion. Let us suppose we want to merge two clusters $\mathscr{C}_1$ and $\mathscr{C}_2$ in the network of sizes $n_1$ and $n_2$ respectively. Let us suppose as well they are connected by $l$ edges as shown in figure 1.

Let us denote $C_F$ the contribution of merging two clusters to the value of a criterion $F$. The contribution $C_F$ can be easily calculated from (6) (for the proof see [Conde-Céspedes, 2013]):

$$C_F = \sum_{i \in \mathscr{C}_1}^{n_1} \sum_{i' \in \mathscr{C}_2}^{n_2} (\phi_{ii'} - \bar{\phi}_{ii'}) \qquad (7)$$

[4] the data was taken from http://snap.stanford.edu/data/com-Amazon.html.

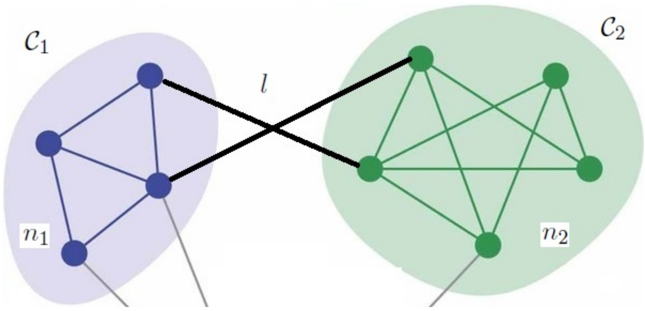

**Fig. 1** Two sub graphs of the entire network we want to merge.

- If $C > 0$ the merger of the two clusters increases the value of the criterion.
- If $C < 0$ the merger of the two clusters decreases the value of the criterion.

Equation (7) shows that the decision of merging or not the two clusters depends on a comparison between the quantity $\sum_{i\in\mathscr{C}_1}^{n_1}\sum_{i'\in\mathscr{C}_2}^{n_2} \phi_{ii'}$ and the quantity $\sum_{i\in\mathscr{C}_1}^{n_1}\sum_{i'\in\mathscr{C}_2}^{n_2} \bar{\phi}_{ii'}$. Giving the fact that both are positive, it is the one with the highest value that decides to merge or not to merge. Thus, whereas the first one is *for* fusion the second one is *against* the fusion.

Table 4 shows the explicit expression of the contribution for the linear criteria described below[5] .

**Table 4** Contribution of merging two clusters for linear criteria.

| **Criterion:** $F$ | $C_F = \sum_{i\in\mathscr{C}_1}^{n_1}\sum_{i'\in\mathscr{C}_2}^{n_2} (\phi_{ii'} - \bar{\phi}_{ii'})$ |
|---|---|
| Zahn-Condorcet | $C_{ZC} = \left(l - \frac{n_1 n_2}{2}\right)$ |
| Owsiński-Zadrożny | $C_{OZ} = (l - n_1 n_2 \alpha) \quad 0 < \alpha < 1$ |
| Deviation to Uniformity | $C_{\text{UNIF}} = (l - n_1 n_2 \delta)$ |
| Newman-Girvan | $C_{NG} = \left(l - n_1 n_2 \frac{d_{av}^1 d_{av}^2}{2M}\right)$ |
| Deviation to Indetermination | $C_{DI} = \left(l - n_1 n_2 \left(\frac{d_{av}^1}{N} + \frac{d_{av}^2}{N} - \frac{2M}{N^2}\right)\right)$ |

[5] The contribution for the Balanced Modularity will be given later.

where $d_{av} = \frac{\sum_{i \in V}^{N} a_{i.}}{N}$ is the average degree of the whole graph, $d_{av}^{1} = \frac{\sum_{i \in \mathscr{C}_1}^{n_1} a_{i.}}{n_1}$ and $d_{av}^{2} = \frac{\sum_{i' \in \mathscr{C}_2}^{n_2} a_{.i'}}{n_2}$ are the average degrees of $\mathscr{C}_1$ and $\mathscr{C}_2$ respectively.

We can remark from table 4 that for the five criteria the contribution compares ”the number of edges between $\mathscr{C}_1$ and $\mathscr{C}_2$: $l$” to the quantity in bold. We can see as well that the *contribution* for locally balanced criteria depends only upon local properties: $l, \bar{l}, n_1, n_2$. In fact, locally balanced criteria are scale invariant. In contrast, for the other criteria having a resolution limit the contribution depends and is *decreasing* on the global size of the network. We remark as well that for three criteria: Newman-Girvan, Deviation to Indetermination and Balanced Modularity the contribution depends on the degree distribution of the two clusters. According to [Barabasi and Albert, 1999] many real networks fall into the class of scale-free networks, meaning that their degree distribution follows a power-law. In a scale-free network a few nodes called hubs have many connexions whereas most nodes have few connexions.

## 4.1 Impact on the optimal number of clusters

From the previous results we can deduce the main characteristics of the optimal partition found by the optimization of each criterion (see table 5). In addition, we remark the following facts:

- **The Zahn-Condorcet criterion:** According to table 4 for merging the two clusters $\mathscr{C}_1$ and $\mathscr{C}_2$, these ones must be connected by at least as many edges as the half of the maximum possible number of edges[6], that is $l > \frac{n_1 n_2}{2}$.

- **The Owsiński-Zadrożny criterion:** For merging the two clusters $\mathscr{C}_1$ and $\mathscr{C}_2$, these ones must be connected by at least as $\alpha\%$ as the maximum possible number of edges.

- **The Deviation to Uniformity:** According to table 4 for the merge to take place the fraction of edges between $\mathscr{C}_1$ and $\mathscr{C}_2$ must be at least equal to the global density of the whole graph.

- **Newman-Girvan criterion:** From table 4 we can deduce that the optimal partition does not have clusters with a single node (this result was already demonstrated in [Brandes et al., 2008]). In fact, if $\mathscr{C}_1$ has only one node with only one connection to $\mathscr{C}_2$, thus $n_1 = 1$, $d_{av}^{1} = 1$, $l = 1$ and consequently the contribution

[6] This result is a consequence of the rule this criterion relies on: ”*The rule of absolute majority of Condorcet*” in voting theory.

is always positive: $C_{NG} = \left(1 - \frac{\sum_{i=1}^{n_2} a_{i.}}{2M}\right) > 0$.

- **Balanced Modularity:** It is easy to understand the behavour of the contribution of Balanced Modularity when we compare it to those of Newman-Girvan and Deviation to Indetermination (see [Conde-Céspedes, 2013] for proof)[7]. Indeed, we demostrated in [Conde-Céspedes, 2013] that:

$$C_{BM} = 2C_{NG} + n_1 n_2 \frac{(d_{av}^1 - d_{av})(d_{av}^2 - d_{av})}{2M(1-\delta)} \tag{8}$$

and

$$C_{BM} = 2C_{DI} + n_1 n_2 \left(2 - \frac{1}{\delta}\right) \frac{(d_{av}^1 - d_{av})(d_{av}^2 - d_{av})}{N^2(1-\delta)}. \tag{9}$$

Although the contribution for the Balanced Modularity is increasing in both the contribution of Newman Girvan $C_{NG}$ and in the contribution of Deviation to Indetermination $C_{DI}$, in both cases $C_{BM}$ has an additional term that we can treat as *regulator*: $\left(n_1 n_2 \frac{(d_{av}^1 - d_{av})(d_{av}^2 - d_{av})}{2M(1-\delta)}\right)$ and $\left(n_1 n_2 \left(2 - \frac{1}{\delta}\right) \frac{(d_{av}^1 - d_{av})(d_{av}^2 - d_{av})}{N^2(1-\delta)}\right)$ respectively. These two regulators have opposite sign for real networks. In fact, the coefficient $\left(2 - \frac{1}{\delta}\right)$ of the second regulator is almost surely negative for real graphs because the density $\delta << 0.5$ for scale-free networks. That is why the Balanced Modularity behaves as a regulator between both criteria: Newman-Girvan and Balanced Modularity. However, when the network size increases $N \longrightarrow \infty$ and $M \longrightarrow \infty$ the regulator terms tend to zero.

Only ground-truth overlapping communities are defined on real networks in table 3. This fact makes difficult to judge the quality of the obtained partitions because we can not directly compare a partition to overlapping communities. That is why in the next section we will consider artificial networks with a predefined community structure.

[7] These expressions are deduced from the two following expressions of Balanced Modularity in terms of Newman-Girvan and Deviation to Indetermination criteria:

$$F_{BM} = 2F_{NG} + \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{2M(1-\delta)}\right) x_{ii'}$$

and

$$F_{BM} = 2F_{DI} + \left(2 - \frac{1}{\delta}\right) \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{N^2(1-\delta)}\right) x_{ii'}.$$

## 5 Experiments with artificial networks

In order to judge the quality of the partitions obtained by each criterion we generated benchmark LFR graphs[8] (see [Lancichinetti et al., 2008]) of different sizes 1000, 5000, 10000, 50000, 100000 and 500000. The input parameters are the same as those considered in [Lancichinetti and Fortunato, 2009]. The average degree is 20, the maximum degree 50, the exponent of the degree distribution is -2 and that of the community size distribution is -1. In order to test the existence of resolution limit we chose small communities sizes, ranging from 10 to 50 nodes, and low values of mixing parameter, 0.10, 0.20 et 0.30. Figure 2 shows the average number of clusters for 100 runs of the generic Louvain algorithm.

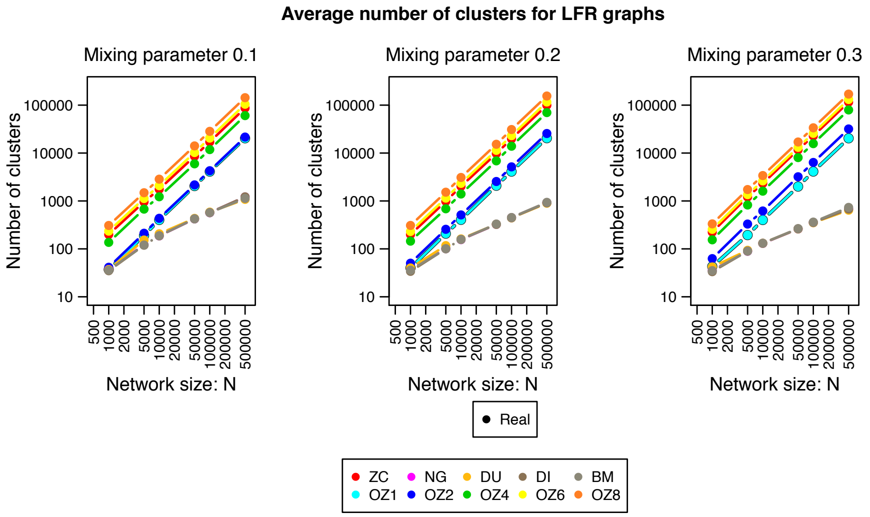

**Fig. 2** Average number of cluster for artificial LFR graphs (logarithmic scale). The curve of the real number of clusters (in black) it is almost overlapped with that of OZ1 and OZ2

In figure 2 it is hard to see the curve of the real number of clusters (in black) beacuse it is almost overlapped with those of OZ1 and OZ2.
Figure 2 shows clearly the difference between the behavior of those criteria having a resolution limit (NG, DU, DI and BM) and the behavior of criteria locally defined (ZC and OZ). As the size of the network increases the four criteria suffering from resolution-limit detect fewer clusters than those predefined. The number of clusters

[8] LFR graphs are benchmark graphs introduced in [Lancichinetti et al., 2008] that aim to reproduce as much as possible the structure that reflects the real properties of nodes and communities found in real networks. These artificial graphs have predefined community structure based on the mixing parameter of each node. As stated in [Lancichinetti et al., 2008], for each node the mixing parameter is the fraction of its links it shares with the nodes of the network outside its community.

is rather comparable for these four functions, one reason can be the fact that the term of negative agreements tends to zero when the network gets bigger. Conversely, the number of clusters of criteria locally defined increases nearly at the same rate as the real number of clusters. Whereas OZ with high $\alpha$ identifies more clusters than those predefined, the criterion which best approaches the real number of clusters is OZ with low values of $\alpha = 0.2$ and $\alpha = 0.1$.

Figure 3 shows the *Normalized Mutual Information*[9] (NMI) for the partitions in figure 2.

Figure 3 shows that the average NMI decreases with the network size for criteria having a resolution limit. Moreover, they almost overlap. Conversely, the NMI of the criteria locally defined seem to increase with the network size. The criterion with the highest NMI is OZ with low values of $\alpha$, 0.1 and 0.2.

Figure 4 shows the average *Normalized Mutual Information* for the mixing parameter ranging from 0.1 to 0.8 for different network sizes.

Figure 4 shows that for all the criteria previously presented the NMI decreases as the mixing parameter increases. This figure demonstrates once more the differences between the behavior of criteria with resolution limit and that of the criteria locally defined. For the first ones the quality decreases abruptly beyond mixing parameter equal to 0.6. For the second ones, the quality seems to decrease at a lower rate. However, it is important to remark that the quality of criteria with a resolution limit decreases not only with the mixing parameter but also with the network size. Converserly, the behavior of the NMI of locally defined criteria seem to have the same behavour independtly of the size of the whole network.

[9] The the normalized mutual information (NMI) is a measure of similarity of two partitions. It was originated in information theory to measure the departure from independence between two random variables. Given a set of objects $V$ and two partitions $P_1$ and $P_2$ defined on $V$, intuitively, the mutual information measures the information that $P_1$ and $P_2$ share. It is normalized between 0 and 1. It is worth 0 if the two partitions are independent and 1 if they are identical. Let $p$ and $q$ be the total number of clusters of partitions $P_1$ and $P_2$ respectively. The NMI is calculated as follows:

$$NMI(P_1,P_2) = \frac{2I(P_1,P_2)}{H(P_1)+H(P_2)}$$

where:

- $I(P_1,P_2) = \sum_{u=1}^{p}\sum_{v=1}^{q} p_{uv} \ln\left(\frac{p_{uv}}{p_{u.}p_{.v}}\right)$ is the mutual information of partitions $P_1$ and $P_2$. $I$ tells how much we learn about $P_1$ if we know $P_2$ and vice versa. The quantity $p_{uv} = \frac{n_{uv}}{N}$ is the fraction of objects who belong simultaneously to cluster $u$ of partition $P_1$ and to cluster $v$ of partition $P_2$. Analogously $p_{uv} = \frac{n_{u.}}{N}$ is the fraction of objects who belong to cluster $u$ of partition $P_1$ and $p_{uv} = \frac{n_{.v}}{N}$ is the fraction of objects who belong to cluster $v$ of partition $P_2$ and $|V| = N$. In the case $n_{uv} = 0$ we assume $\ln\left(\frac{p_{uv}}{p_{u.}p_{.v}}\right) = 0$.
- $H(P_1) = -\sum_{u=1}^{p} p_{u.} \ln p_{u.}$ represents the Shanon entropy of $P_1$ and $H(P_2) = -\sum_{v=1}^{q} p_{.v} \ln p_{.v}$ represents the Shanon entropy of $P_2$ (see [Shannon, 1948]).

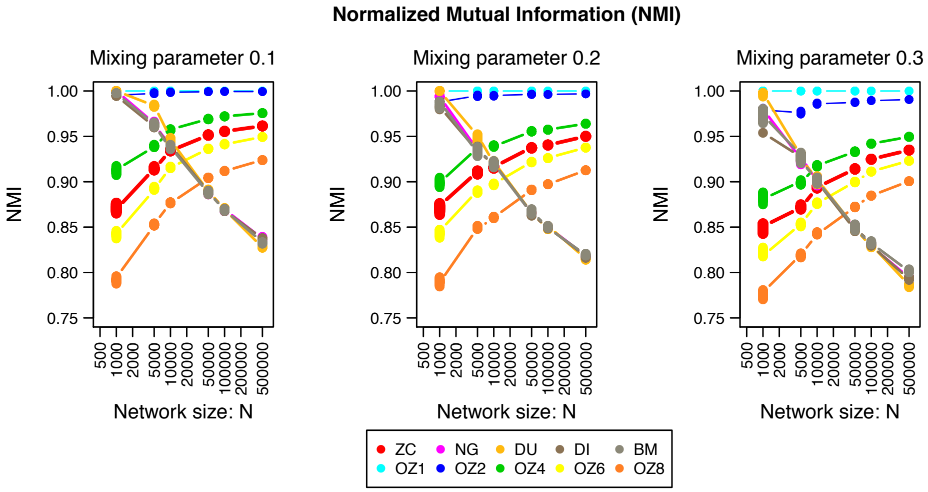

**Fig. 3** The Average Normalized Mutual Information (NMI) on the graphs in figure 2 (logarithmic scale).

Another point to remark is that even when the mixing parameter is high all the criteria find a community structure. In fact, the pre-defined communities in the LFR graphs are based on mixing parameter, whereas all the criteria analysed in this article have their own definition of *graph with no community structure* which is not based on the mixing parameter.

Table 5 presents a summary of the results found by the previous analysis.

## 6 Conclusions

We presented six linear modularization criteria in relational notation, Zahn-Condorcet, Owsiński- Zadrożny, the Newman-Girvan modularity, the Deviation to Uniformity index, the Deviation to Indetermination index and the Balanced-Modularity. This notation allowed us to easily identify the criteria suffering from a resolution limit. We found that the first two criteria had a local definition, whereas the others, based on a null model, had a resolution limit. These findings were confirmed by modularizing real and artificial graphs using a generic version of the Louvain algorithm. We compared the number of clusters found by the six criteria and the Normalized Mutual information for artificial graphs. The results showed that criteria based on a local definition had a better performance than those based on a null model when the size of the graph increases, experimentally the crition having the best behavior was Owsiński- Zadrożny with low values of parameter $\alpha$. However, it is important

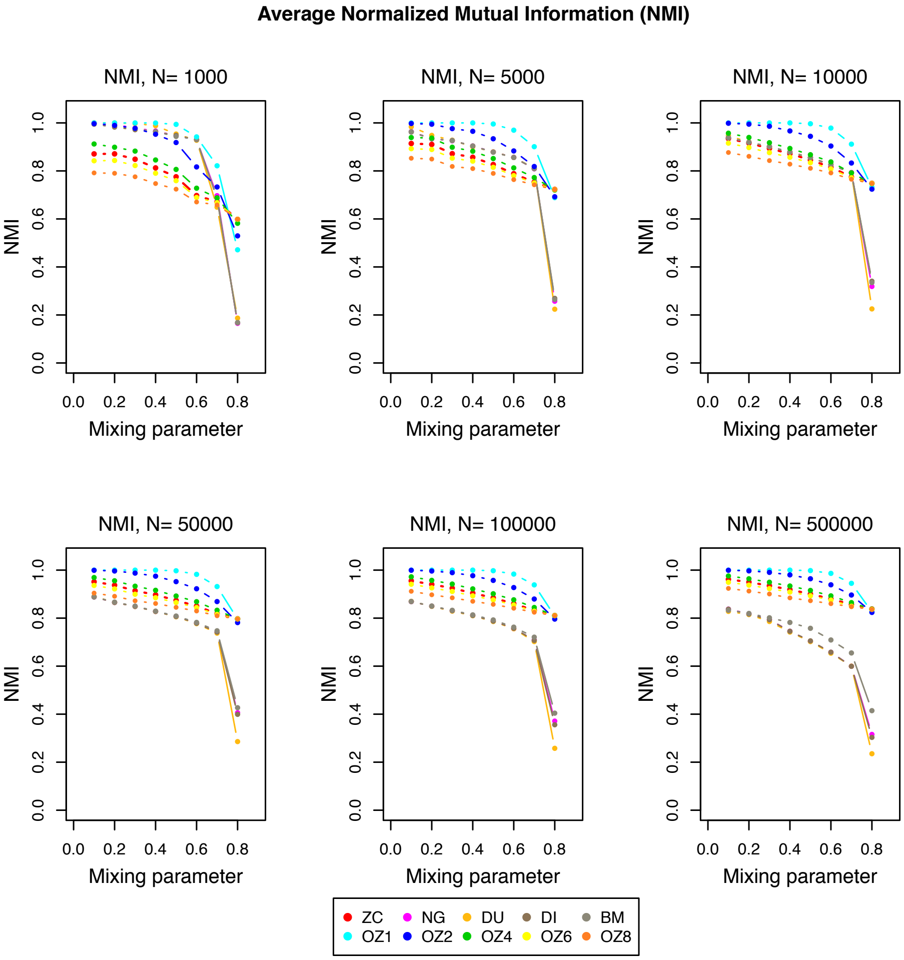

**Fig. 4** The Average Normalized Mutual Information (NMI) according to mixing parameter for networks of 6 different sizes: 1000, 5000, 10000, 50000, 100000 and 500000.

to remark that these results are based on a particular kind of graphs, more precisely, graphs with a low mixing parameter, small communities[10], node degrees and community sizes distributed according to a power law.

**Acknowledgements** This work is supported by REQUEST and Open Food System projects.

[10] What we call *small* are communities ranging from 10 to 50 nodes, that is the same sizes considered by the authors of LFR graphs (see [Lancichinetti and Fortunato, 2009]).

**Table 5** Summary by criterion

| **Criterion** | **Characteristics of the optimal partition** |
|---|---|
| Zahn-Condorcet | • The density of edges of each cluster is at least equal to 50%.<br>• No resolution limit.<br>• For real networks the optimal partition contains many small clusters or single nodes. |
| Owsiński-Zadrożny | • It gives the choice to define the minimum required within-cluster density, $\alpha$.<br>• For $\alpha = 0.5$ the Owsiński-Zadrożny criterion $\equiv$ the Zahn-Condorcet criterion.<br>• No resolution limit.<br>• The optimal partition depends on the parameter $\alpha$ |
| Deviation to Uniformity | • A particular case of Owsiński-Zadrożny criterion with $\alpha = \delta$.<br>• The density of within cluster edges of each cluster is at least the global density $\delta$.<br>• It has a resolution limit. |
| Newman-Girvan | • It depends on the degree distribution.<br>• It has a resolution limit.<br>• The optimal partition has no single nodes. |
| Deviation to Indetermination | • It depends on the degree distribution.<br>• It has a resolution limit. |
| Balanced modularity | • It depends on the degree distribution.<br>• It has a resolution limit. |

## Appendix

**Theorem 1 (The density of clusters obtained by maximization of Zahn-Condorcet criterion is least 50% ).** *Given a connected, non-oriented and unweighted graph $G = (V, E)$, the optimal partition obtained by optimizing the Zahn-Condorcet criterion has the following property: the number of within-cluster edges of each cluster is at least as half as the possible maximum existing within-cluster edges, that is to say the number of existing edges in the case the cluster is a clique. Furthermore, every node in each cluster is connected with at least as half as the total nodes inside the cluster.*

*Proof.* Considering the constraints of reflexivity and symmetry of the relational variable $x_{ii'}$ (i.e. $x_{ii} = 1 \forall i$ and $x_{ii'} = x_{i'i}$), the expression of Zahn-Condorcet criterion in table 2 can be written as follows:

$F_{ZC}(X) = \sum_{i>i'}(a_{ii'} - \bar{a}_{ii'})x_{ii'} + N^2 - 2M - N.$

where:

- $\sum_{i>i'} a_{ii'}x_{ii'}$ is the number of within-cluster edges for all clusters.
- $\sum_{i>i'} \bar{a}_{ii'}x_{ii'}$ is the number of missing within-cluster edges for all clusters.

If we denote $E_j$ the number of within edges of cluster $j$, the total number of missing edges for the cluster $j$ will be $\left(\frac{n_j(n_j-1)}{2} - E_j\right)$. So, the criterion Zahn-Condorcet will become:

$$F_{ZC}(\mathscr{C}) = \sum_{j=1}^{\kappa} \left(E_j - \left(\frac{n_j(n_j-1)}{2} - Ej\right)\right) + N^2 - 2M - N,$$

or

$$F_{ZC}(\mathscr{C}) = \sum_{j=1}^{\kappa} (2E_j - \frac{n_j(n_j-1)}{2}) + N^2 - 2M - N.$$

the term $(2E_j - \frac{n_j(n_j-1)}{2})$ represents the contribution of cluster $j$ to the value of the criterion. For each cluster of the optimal partition this term must be positive or null. Otherwise it would be possible to obtain a better partition by isolating each node in cluster $j$ (the contribution to the value of the criterion by a cluster of an isolated node is null). This implies:

$$(2E_j - \frac{n_j(n_j-1)}{2}) \geq 0, \text{ or } E_j \geq \frac{n_j(n_j-1)}{4}.$$

So, each cluster $j$ has a density of at least 50%.

This result can be extended to every node of each cluster of the optimal partition. In fact, let us suppose that there is a cluster $j$ containing a node $n_0$ which is connected with less than half of the total nodes in the cluster. Let us denote $E_{j_0}$ the connexions of $n_0$ to nodes in $C_j$. So, $E_{j_0} <= \frac{(n_j-1)}{2}$.

It is always possible to obtain a better partition by isolating $n_0$. In fact, the contribution of the two resulting clusters after isolation of node $n_0$ is:

$$2(E_j - E_{j_0}) - \frac{(n_j-1)(n_j-2)}{2}$$

this last expression is greater than the contribution of cluster $j$, given by $(2E_j - \frac{n_j(n_j-1)}{2})$, if $n_0$ is connected with less than half of nodes in $C_j$.

This also proves why the partitions obtaining by optimizing Zahn-Condorcet criterion contain sometimes clusters of isolates nodes. □

## References

[Ah-Pine and Marcotorchino, 2007] Ah-Pine, J. and Marcotorchino, F. (2007). Statistical, geometrical and logical independences between categorical variables. *Proc. of the ASMDA2007 Symposium, Chania, Greece*.

[Albert et al., 1999] Albert, R., Jeong, H., and Barabási, A. (1999). Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131.

[Barabasi and Albert, 1999] Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

[Blondel et al., 2008] Blondel, V., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment P10008*.

[Brandes et al., 2008] Brandes, U., Delling, D., Gaertler, M., Grke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188.

[Campigotto et al., 2014] Campigotto, R., Conde-Céspedes, P., and Guillaume, J. (2014). A generalized and adaptive method for community detection. *CoRR*, abs/1406.2518.

[Conde-Céspedes, 2013] Conde-Céspedes, P. (2013). *Modélisations et extensions du formalisme de l'Analyse Relationnelle Mathématique à la modularisation des grands graphes*. Thèse de doctorat, Université Pierre et Marie Curie.

[Conde-Céspedes and Marcotorchino, 2013] Conde-Céspedes, P. and Marcotorchino, F. (2013). Comparison different modularization criteria using relational metric. In Nielsen, F. and Barbaresco, F., editors, *Proc. First International Conference, Geometric Science of Information*, number 1, pages 180–187, Paris, France. Springer-Verlag.

[Conde-Céspedes and Marcotorchino, 2012] Conde-Céspedes, P. and Marcotorchino, J. (2012). Modularisation et recherche de communautés dans les réseaux complexes par unification relationnelle. *Revue des Nouvelles Technologies de l'Information*, Apprentissage Artificiel et Fouille de Données, RNTI-A-6:71–97.

[Condorcet, 1785] Condorcet, C. A. M. d. (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. *Journal of Mathematical Sociology*, 1(1):113–120.

[Decaestecker, 1992] Decaestecker, C. (1992). *Apprentissage en classification conceptuelle incrémentale*. PhD thesis, Université Libre de Bruxelles (Faculté des Sciences).

[Fortunato and Barthelemy, 2006] Fortunato, S. and Barthelemy, M. (2006). Resolution limit in community detection. In *Proceedings of the National Academy of Sciences of the United States of America*.

[Gleiser and Danon, 2003] Gleiser, P. and Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems (ACS)*, 06(04):565–573.

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

[Kumpula et al., 2007] Kumpula, J., Saramäki, J., Kaski, K., and Kertesz, J. (2007). Limited resolution in complex network community detection with potts model approach. *The European Physical Journal B*, 56(1):41–45.

[Lancichinetti and Fortunato, 2009] Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117.

[Lancichinetti et al., 2008] Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78(4).

[Mancoridis et al., 1998] Mancoridis, S., Mitchell, B., Rorres, C., Chen, Y., and Gansner, E. (1998). Using automatic clustering to produce high-level system organizations of source code. In *In the IEEE Proceedings of the 1998 International Workshop on Program Understanding (IWPC'98)*, pages 45–52, Ischia, Italy. IEEE Computer Society.

[Marcotorchino, 1984] Marcotorchino, F. (1984). Utilisation des comparaisons par paires en statistique des contingences (partie i). *Publication du Centre Scientifique IBM de Paris, F057, et Cahiers du Séminaire Analyse des Données et Processus Stochastiques Université Libre de Bruxelles*, pages 1–57.

[Marcotorchino, 1985] Marcotorchino, F. (1985). Utilisation des comparaisons par paires en statistique des contingences (partie iii). *Etude F-081 du Centre Scientifique IBM de Paris*, pages 1–39.

[Marcotorchino, 2013] Marcotorchino, F. (2013). Optimal transport, spatial interaction models and related problems, impacts on relational metrics, adaptation to large graphs and networks modularity. *Internal Publication of Thales*.

[Marcotorchino and Conde-Céspedes, 2013] Marcotorchino, F. and Conde-Céspedes, P. (2013). Optimal transport and minimal trade problem, impacts on relational metrics and applications to large graphs and networks modularity. In Nielsen, F. and Barbaresco, F., editors, *Proc. First International Conference, Geometric Science of Information*, number 1, pages 169–179, Paris, France. Springer-Verlag.

[Marcotorchino and Michaud, 1979] Marcotorchino, F. and Michaud, P. (1979). *Optimisation en Analyse ordinale des données*. Masson, Paris.

[Michalski and Stepp, 1983] Michalski, R. and Stepp, R. (1983). Learning from observation: Conceptual clustering. In Michalski, R., Carbonell, J., Mitchell, T., and Kaufmann, M., editors, *Machine Learning: An Artificial Intelligence Approach*, volume 1, chapter 11, pages 331–364. Tioga.

[Mislove et al., 2007] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., and Bhattacharjee, B. (2007). Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA.

[Newman and Girvan, 2004] Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E.*, 69(2).

[Owsiński and Zadrożny, 1986] Owsiński, J. and Zadrożny, S. (1986). Clustering for ordinal data: a linear programming formulation. *Control and Cybernetics*, 15(2):183–193.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

[Wei and Cheng, 1989] Wei, Y. and Cheng, C. (1989). Towards efficient hierarchical designs by ratio cut partitioning. *IEEE International Conference on Computer-Aided Design*, pages 298–301.

[Yang and Leskovec, 2012] Yang, J. and Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. In *International Conference on Data Mining*, volume abs/1205.6233, pages 745–754. IEEE Computer Society.

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