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

Patricia Conde-Céspedes*, Jean-François Marcotorchino**, Emmanuel Viennet*, [1]

*L2TI - Institut Galilée - Université Paris 13
99, av. Jean-Baptiste Clément; 93430 Villetaneuse - France
prenom.nom@univ-paris13.fr

**Thales Communications et Sécurité
4 av. des Louvresses; 92230 Gennevilliers - France
jeanfrancois.marcotorchino@thalesgroup.com

Résumé. La modularisation de grands graphes ou recherche de communautés est abordée comme l'optimisation d'un critère de qualité, l'un des plus utilisés étant la modularité de Newman-Girvan. D'autres critères, ayant d'autres propriétés, aboutissent à des solutions différentes. Dans cet article, nous présentons une réécriture relationnelle de six critères linéaires: Zahn-Condorcet, Owsiński-Zadrożny, l'Ecart à l'Uniformité, l'Ecart à l'Indétermination et la Modularité Equilibrée. Nous utilisons une version générique de l'algorithme d'optimisation de Louvain pour approcher la partition optimale pour chaque critère sur des réseaux réels de différentes tailles. Les partitions obtenues présentent des caractéristiques différentes, concernant notamment le nombre de classes. Le formalisme relationnel nous permet de justifier ces différences d'un point de vue théorique. En outre, cette notation permet d'identifier facilement les critères ayant une limite de résolution (phénomène qui empêche en pratique la détection de petites communautés sur de grands graphes). Une étude de la qualité des partitions trouvées dans les graphes synthétiques LFR permet de confirmer ces résultats.

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 beneficial 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.

1. This work is supported by REQUEST project between Thales and Paris 13 University.

Assessing the quality of a graph partition requires a modularization criterion. This function will be optimized to find the best partition. Various modularization criteria have been 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 allows to understand the properties of those modularization criteria. It allows to easily identify the criteria suffering from a resolution limit, first discussed by Fortunato et Barthelemy (2006). We will complete this theoretical study by some experiments on real and synthetic networks, demonstrating the effectiveness of our classification.

This paper is organized as follows : Section 2 presents the Mathematical Relational Analysis approach, we introduce the property of *balance* for linear criteria and its relation to the property of resolution limit. In Section 3, we present the six linear modularization criteria in the relational formalism. Next, Section 4 presents 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 [2] 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 :

2. For more details about Relational Analysis theory see Marcotorchino et Michaud (1979) and Marcotorchino (1984).

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

Modularizing a graph implies to find **X** as close as possible to **A**. A modularization criterion $F(X)$ is a function which measures either a *similarity* or a distance between **A** and **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 [3].

We define as well $\bar{\mathbf{X}}$ and $\bar{\mathbf{A}}$ as the inverse relation of **X** and **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 κ the optimal number of clusters, that means the number of clusters of the partition **X** which maximizes the criterion $F(X)$.

2.1 Linear balanced criteria

Every linear criterion is an affine function of **X**, therefore in relational notation it can be written as :

$$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) x_{ii'} + K, \tag{3}$$

where the function $\phi(a_{ii'})$ depends only on the original data (for instance the adjacency matrix). In the following we will use K to denote any *constant* depending only on the original data.

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

$$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) x_{ii'} + \sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}(a_{ii'}) \bar{x}_{ii'} + K. \tag{4}$$

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 $\sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}_{ii} > 0$.

3. In fact, the problem of modularization can be written in the general form :

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

subject to the constraints of an equivalence relation :

$x_{ii'} \in \{0,1\}$		Binary
$x_{ii} = 1$	$\forall i$	Reflexivity
$x_{ii'} - x_{i'i} = 0$	$\forall (i,i')$	Symmetry
$x_{ii'} + x_{i'i''} - x_{ii''} \leq 1$	$\forall (i,i',i'')$	Transitivity

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 to proceed.

By replacing $\bar{x}$ by its definition $1 - x_{ii'}$, equation (4) can be rewritten as follows :

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

From this expression we can 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 ϕ 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 $\bar{\phi}_{ii'}$ satisfy*

$$\phi_{ii'} + \bar{\phi}_{ii'} = K_L \quad \forall \, (i, i')$$

where K_L is a constant depending 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 :
— Since K_L depends only on properties of the pair (i, i') , that is local properties, we call this property **local balance**.
— 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.
— In the particular case of local balance where K_L is constant $\forall (i, i')$, that is $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ sum up to a constant, we have the following situation : 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 (5).

That implies $\sum_{i=1}^{N}\sum_{i'=1}^{N}\phi_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{\phi}_{ii'} = 2M$. This constraint implies that $\bar{\phi}_{ii'}$ depends upon the total number of edges M. Consequently, the decision of clustering together two sub-graphs depends on a characteristic of the whole network and the criterion is not scale invariant because it depends on a global property 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 whose functions $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ satisfy the following conditions :*

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

$$\phi_{ii'} + \bar{\phi}_{ii'} = g(K_G) \quad \forall\,(i,i')$$

where $g(K_G)$ is a function depending on global properties of the graph K_G is a criterion based on a null model.

K_G can be for example the total number of edges or nodes. We can deduce from definitions 2 and 3 that a linear criterion can not be local balanced and based on a null model at the same time.

In the particular case where $\bar{\phi}$ decreases if the size of the network increases, it becomes negligible for large graphs. As explained previously, if this term tends to zero, the optimization of the criterion will tend to put together 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 et 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 scale which depends on global characteristics of the graph even weakly interconnected complete graphs, which represent 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 identify 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 (see Zahn (1964)) 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 Condorcet (1785)) introduced in Relational Consensus and whose relational coding is given in Marcotorchino et Michaud (1979). This criterion requires that every node in each cluster be connected to 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 Conde-Céspedes (2013) for the demonstration).

2. **The Owsiński-Zadrożny criterion (1986)** (see Owsiński et Zadrożny (1986)) it is a generalization of Condorcet's function. It has a parameter α, which allows, according to the context, to define the minimal percentage of required within-cluster edges : α. For $\alpha = 0.5$ this criterion is equivalent to Condorcet's criterion. The parameter α 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 et 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 et 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 for 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*[4] (a graph without regard to community structure), introduced in Marcotorchino (2013) and Marcotorchino et Conde-Céspedes (2013). The Deviation to Indetermination is based on a null model, therefore it has a **resolution limit**.

6. **The Balanced modularity (2013)** This criterion, introduced in Conde-Céspedes et 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**.

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 $0 < \alpha < 1$
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)	$F_{\mathrm{UNIF}}(X) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{2M}{N^2}\right)x_{ii'}$
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)	$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left((a_{ii'} - P_{ii'})\,x_{ii'} + (\bar{a}_{ii'} - \bar{P}_{ii'})\bar{x}_{ii'}\right)$ 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 1 – *Relational notation of linear modularity functions.*

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

The six linear criteria of Table 1 verify the property of *balance*, so it is not necessary to fix in advance the number of clusters, more specifically :

	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	$\sum_{i,i'=1}^{N}\sum_{i'=1}^{N} \bar{p}_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{a}_{ii'} = N^2 - 2M$

TABLE 2 – *Balance Property for Linear criteria.*

From Tables 1 and 2 one can easily deduce that for the criteria having a resolution limit the quantity $\bar{\phi}_{ii'}$ decreases when the size of the graph becomes larger.

4 Tests with real and artificial networks

We modularized six real networks of different sizes : Jazz (Gleiser et Danon (2003)), Internet (Hoerdt et Magoni (2003)), Web nd.edu (Albert et al. (1999)), Amazon (Yang et Leskovec (2012) [5]) 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 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 α, that is the minimal required fraction of within-cluster edges, so the criterion becomes more flexible.

Only ground-truth overlapping communities are defined on these previuos real networks. This fact makes difficult to judge the quality of the obtained partitions. That si why we generated five benchmark LFR graphs (see Lancichinetti et al. (2008)) of different sizes 1000, 5000, 10000, 100000 and 500000. The input parameters are the same as those considered in Lancichinetti et Fortunato (2009). The average degree is 20, the maximum degree 50, the exponent

5. the data was taken from `http://snap.stanford.edu/data/com-Amazon.html`.

Network	**Jazz**	**Internet**	**Web nd.edu**	**Amazon**	**Youtube**
$N \sim$	198	70k	325k	334k	1M
$M \sim$	3k	351k	1M	925k	3M
δ	0,14	$1{,}44 \cdot 10^{-04}$	$2{,}77 \cdot 10^{-05}$	$1{,}65 \cdot 10^{-05}$	$4{,}64 \cdot 10^{-06}$
Criterion	κ	κ	κ	κ	κ
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 – *Ref : Zahn-Condorcet (ZC), Deviation to Uniformity (UNIF), Newman-Girvan (NG), Deviation to Indetermination(DI) and Balanced Modularity (BM).*

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 a low mixing parameter, 0.10. So, the communities are clearly defined. Figure 1 shows the average number of clusters for 100 runs of the generic Louvain algorithm.

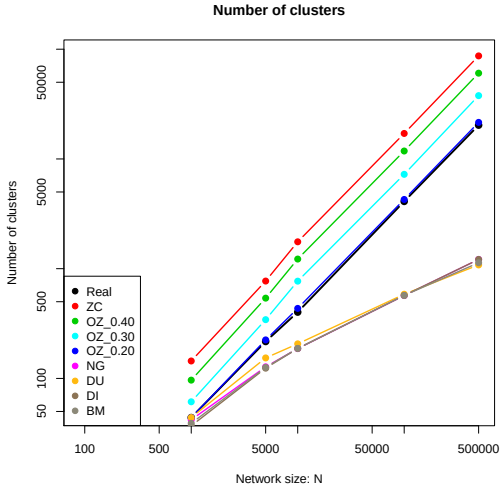

FIGURE 1 – *Average number of cluster for artificial LFR graphs (logarithmic scale).*

Figure 1 shows clearly the difference between the behaviour of those criteria having a resolution limit (NG, DU, DI and BM) and the behaviour 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 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 criteria locally defined identified more clusters than those predefined, specially ZC. The criterion which best approaches the real number of clusters is OZ with $\alpha = 0.2$. Figure 2 shows the average *Normalized Mutual Information* for the partitions in Figure 1.

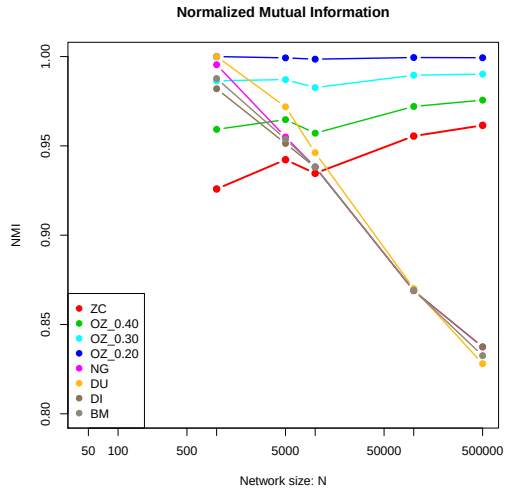

FIGURE 2 – *The Average Normalized Mutual Information (NMI) on the graphs in 1.*

Figure 2 shows that the average NMI decreases with the network size for criteria having a resolution limit. The criterion with the highest NMI is OZ with $\alpha = 0.2$ which guarantees an within-cluster density of 20%.

5 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 those criteria ba-

sed on a local definition had a better performance than those based on a null model when the size of the graph increases.

Références

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

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

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

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

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, P. et F. Marcotorchino (2013). Comparison different modularization criteria using relational metric. In F. Nielsen et F. Barbaresco (Eds.), *Proc. First International Conference, Geometric Science of Information*, Number 1, Paris, France, pp. 180–187. Springer-Verlag.

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.

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

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

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

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

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

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

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*, 1–57.

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

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

Marcotorchino, F. et P. Michaud (1979). *Optimisation en Analyse ordinale des données*. Paris : Masson.

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

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

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

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

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

Summary

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 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 allows 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.