Comparison of linear modularization criteria using the relational formalism

Patricia Conde Céspedes

Université Paris Nord
L2TI

December 16th, 2014

Table of contents

1. Introduction and objective
   - Definitions
2. Relational approach
3. Comparison of linear criteria
4. Applications
5. Conclusions

1. Introduction and objective
   - Definitions
2. Relational approach
   - Mathematical Relational modelling
   - Properties verified by linear modularization criteria
3. Comparison of linear criteria
4. Applications
   - The generalized Louvain algorithm
   - Examples: real large graphs
5. Conclusions

Description of the problem

Nowadays, we can find networks everywhere: biology, sociology, computer programming, marketing, etc). cyber-marketing, cyber-Security.
It is difficult to analyse a network directly because of its big size.
Therefore, we need to decompose it in clusters or modules $\Longleftrightarrow$ **modularize** it.
Different modularization criteria have been formulated in different contexts in the last few years and we need to compare them.

Objective: Compare the partitions found by different linear criteria

We will provide a **unified** notation of different linear modularization criteria to understand the properties of the clusters found by their optimization. Moreover, this notation allows to easily identify the criteria having a **resolution limit**.

Definitions and notations

A **graph** $G(V, E)$ is a set of objects V, called nodes, linked by edges E.

Definitions and notations

A **graph** $G(V, E)$ is a set of objects V, called nodes, linked by edges E.

$N = |V|$ is the number of nodes and $M = |E|$ is the number of edges.

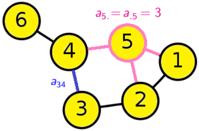
Definitions and notations

A **graph** $G(V, E)$ is a set of objects V, called nodes, linked by edges E.

$N = |V|$ is the number of nodes and $M = |E|$ is the number of edges.
A graph is completely described by a $N \times N$ matrix called the **Adjacency Matrix A** defined as follows

$$a_{ii'} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } i', \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

Example: given a graph with $N = 6$ and $M = 7$.



its adjacency matrix is:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

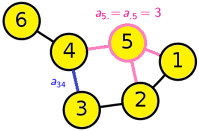
Definitions and notations

A **graph** $G(V, E)$ is a set of objects V, called nodes, linked by edges E.

$N = |V|$ is the number of nodes and $M = |E|$ is the number of edges.
A graph is completely described by a $N \times N$ matrix called the **Adjacency Matrix A** defined as follows

$$a_{ii'} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } i', \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Example: given a graph with $N = 6$ and $M = 7$.



its adjacency matrix is:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

The **degree** d_i of node i is the number of edges incident to i.
$d_i = \sum_{i'} a_{ii'} = a_{i.} = a_{.i}$
The **average degree** of the graph is $d_{av} = \frac{2M}{N}$.
The **Density of the graph** is $\delta = \frac{2M}{N^2}$.

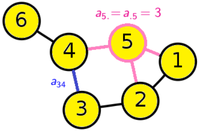
Definitions and notations

A **graph** $G(V,E)$ is a set of objects V, called nodes, linked by edges E.

$N = |V|$ is the number of nodes and $M = |E|$ is the number of edges.
A graph is completely described by a $N \times N$ matrix called the **Adjacency Matrix A** defined as follows

$$a_{ii'} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } i', \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

Example: given a graph with $N = 6$ and $M = 7$.



its adjacency matrix is:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

The **degree** d_i of node i is the number of edges incident to i.
$d_i = \sum_{i'} a_{ii'} = a_{i.} = a_{.i}$
The **average degree** of the graph is $d_{av} = \frac{2M}{N}$.
The **Density of the graph** is $\delta = \frac{2M}{N^2}$.

Just in case the graph is weighted we will denote the adjacency matrix **W**.

Table of contents

1 Introduction and objective

2 Relational approach

- Mathematical Relational modelling
- Properties verified by linear modularization criteria

3 Comparison of linear criteria

4 Applications

5 Conclusions

Mathematical Relational modelling

Let **X** be a square matrix of order N defining an equivalence relation on V as follows:

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

Mathematical Relational modelling

Let **X** be a square matrix of order N defining an equivalence relation on V as follows:

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

We present a modularization criterion as a function F to optimize:

$$\max_X \text{ or } \min_X F(A, X). \tag{3}$$

subject to the constraints of an equivalence relation:

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

Finding the exact solution of this problem turns impractical for large graphs, therefore we will use heuristics ad-hoc.

Properties verified by linear modularization criteria

A criterion is **linear** if it can be written in the general form:

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

Properties verified by linear modularization criteria

A criterion is **linear** if it can be written in the general form:

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

Besides that, the criterion has the property of **General balance** if it can be written in the form:

$$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) x_{ii'} + \sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}(a_{ii'}) \bar{x}_{ii'} \tag{6}$$

Properties verified by linear modularization criteria

A criterion is **linear** if it can be written in the general form:

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

Besides that, the criterion has the property of **General balance** if it can be written in the form:

$$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) x_{ii'} + \sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}(a_{ii'}) \bar{x}_{ii'} \tag{6}$$

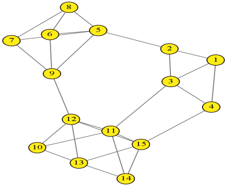
where K is any constant depending only on the original data and $\bar{x}_{ii'} = (1 - x_{ii'})$ (the opposite relation of $\mathbf{X}$);
$\phi(a_{ii'}) \geq 0 \, \forall i, i'$ and $\bar{\phi}(a_{ii'}) \geq 0 \, \forall i, i'$ are non negative functions verifying:
$\sum_{i=1}^{N} \sum_{i'=1}^{N} \phi_{ii} > 0$ and $\sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}_{ii} > 0;$.

The quantities $\sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) x_{ii'}$ and $\sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}(a_{ii'}) \bar{x}_{ii'}$ are called positive **(+)** and negative **(-)** agreements respectively.

The impact of the property of General balance

Let κ denote the number of clusters obtained after optimization of the criterion.

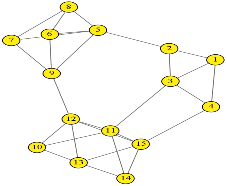
Given a graph



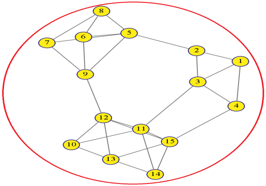
The impact of the property of General balance

Let κ denote the number of clusters obtained after optimization of the criterion.

Given a graph



If **(-)** agreements missing ($\bar{\phi}_{ii'} = 0 \, \forall i, i'$)

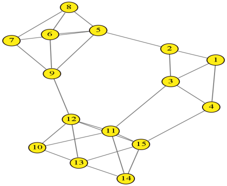


then all nodes are clustered together, $\kappa = 1$

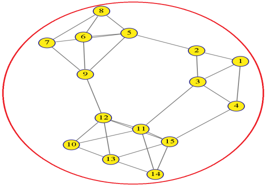
The impact of the property of General balance

Let κ denote the number of clusters obtained after optimization of the criterion.

Given a graph

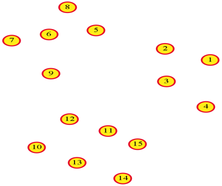


If **(-)** agreements missing ($\bar{\phi}_{ii'} = 0 \, \forall i, i'$)



then all nodes are clustered together, $\kappa = 1$

If **(+)** agreements missing ($\phi_{ii'} = 0 \, \forall i, i'$)



then all nodes are separated, $\kappa = N$

Contribution: Different levels of general balance for linear criteria

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') has the property of local balance.

Therefore K_L must not depend on global properties of the graph.

Contribution: Different levels of general balance for linear criteria

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.

A linear criterion can not be local balanced and based on a null model at the same time.

Resolution limit

If $\bar{\phi}$ tends to zero with the graph size the criterion has a **resolution limit**.

Table of contents

Some linear criteria in relational notation

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

Table : Relational notation of linear modularity functions.

Some linear criteria in relational notation (continuation)

Criterion	**Relational notation**
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 : Relational notation of linear modularity functions.

Properties of these linear criteria

The 6 criteria have the property of **General balance**.

	Global balance	
Criterion	**Local Balance**	**Null model**
Zahn-Condorcet	X	
Owsiński-Zadrożny	X	
Newman-Girvan		X
Deviation to Uniformity		X
Deviation to Indetermination		X
Balanced modularity		X

Table : Balance Property for Linear criteria

First approach: the deviation form notation

Criterion	**Notation** $F(X)=\sum_{i=1}^{N}\sum_{i'=1}^{N}(\phi_{ii'}-\bar{\phi}_{ii'})x_{ii'}$
Zahn-Condorcet	$F_{ZC}(X)=\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}-\frac{1}{2}\right)x_{ii'}$
Owsiński-Zadrożny	$F_{OZ}(X)=\sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'}-\alpha)x_{ii'}$
Deviation to uniformity	$F_{\mathrm{UNIF}}(X)=\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}-\frac{2M}{N^2}\right)x_{ii'}$
Newman-Girvan	$F_{NG}(X)=\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}-\frac{a_{i.}a_{.i'}}{2M}\right)x_{ii'}$
Deviation to indetermination	$F_{DI}(X)=\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}-\left(\frac{a_{i.}}{N}+\frac{a_{.i'}}{N}-\frac{2M}{N^2}\right)\right)x_{ii'}$

Comparison between Newman-Girvan, Deviation to Indetermination and the Balanced Modularity

Maximizing the Balanced Modularity turns out to maximize the following expressions depending upon the Newman-Girvan criterion and the Deviation to Indetermination respectively.

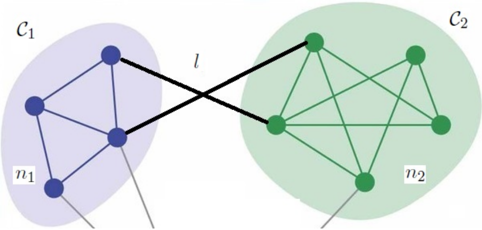
$$F_{BM} = 2F_{NG} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{2M(1-\delta)}\right) x_{ii'}.$$

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

The **Balanced Modularity** behaves as a **regulator** between the Newman-Girvan criterion and the Deviation to Indetermination.

Second approach: Impact of merging two clusters

Now let us suppose we want to merge two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ in the network of sizes n_1 and n_2 respectively. Let us suppose as well they are connected by l edges and they have average degree d_{av}^1 et d_{av}^2 respectively.



Impact of merging two clusters

What is the contribution of merging two clusters to the value of each criterion?

Impact of merging two clusters

What is the contribution of merging two clusters to the value of each criterion?
The **contribution** C of merging two clusters will be:

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

Impact of merging two clusters

What is the contribution of merging two clusters to the value of each criterion?
The **contribution** C of merging two clusters will be:

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

The objective is to compare function $\phi(.)$ to function $\bar{\phi}(.)$

- If $C > 0$ the criterion merges the two clusters, it is a **gain**.
- If $C < 0$ the criterion separates the two clusters, it is a **cost**.

The Contribution of merging two clusters

Contribution of merging two clusters for linear criteria.

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

Summary by criterion

Criterion	Characteristics of the clustering
Zahn-Condorcet	- The **density** of edges of each cluster is at least equal to 50%. - No resolution limit. - 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**, α. - For $\alpha = 0.5$ the Owsiński-Zadrożny criterion $\equiv$ the Zahn-Condorcet criterion. - No resolution limit.
Deviation to Uniformity	- A particular case of Owsiński-Zadrożny criterion with $\alpha = \delta$. - The **density** of within cluster edges of each cluster is at least δ. - It has a resolution limit.

Summary by criterion

Criterion	Characteristics of the clustering
Newman-Girvan	- It has a resolution limit. - The contribution depends on the **degree distribution** of the clusters. - The optimal partition has no single nodes.
Deviation to Indetermination	- It has a resolution limit. - The contribution depends on the **degree distribution** of the clusters.
Balanced modularity	- It has a resolution limit. - The contribution depends on the **degree distribution** of the clusters. - Depending upon δ and d_{av} this criterion behaves like a **regulator** between the NG criterion and the DI criterion.

Table of contents

Modularizing large graphs with the generalized Louvain algorithm

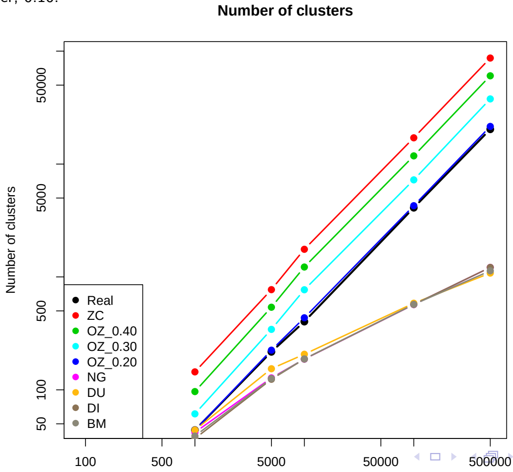
Number of clusters found by the generalized Louvain algorithm (see [Campigotto et al. (2014)])

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	1.44×10^{-04}	2.77×10^{-05}	1.65×10^{-05}	4.64×10^{-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 : Ref: Zahn-Condorcet (ZC), Deviation to Uniformity (UNIF), Newman-Girvan (NG), Deviation to Indetermination(DI) and Balanced Modularity (BM).

The number of clusters in artificial LFR graphs

Five benchmark LFR graphs of sizes 1000, 5000, 10000, 100000 and 500000. The input parameters are the same as those considered in [Lancichinetti et al (2010)]: small communities sizes, ranging from 10 to 50 nodes, and a low mixing parameter, 0.10.



The Normalized Mutual Information with LFR graphs

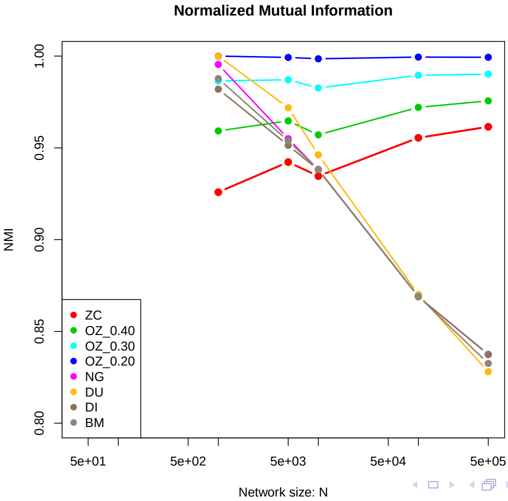


Table of contents

Conclusions

- We presented six different modularization criteria in Relational notation.

Conclusions

- We presented six different modularization criteria in Relational notation.
- We described and clearly defined the property of **balance** by making the link between this property and the **resolution limit** property.

Conclusions

- We presented six different modularization criteria in Relational notation.
- We described and clearly defined the property of **balance** by making the link between this property and the **resolution limit** property.
- The generic Louvain algorithm allowed us to modularize real **large graphs** and we could compare the number of clusters found by the different criteria.

Conclusions

- We presented six different modularization criteria in Relational notation.
- We described and clearly defined the property of **balance** by making the link between this property and the **resolution limit** property.
- The generic Louvain algorithm allowed us to modularize real **large graphs** and we could compare the number of clusters found by the different criteria.
- We characterized the partitions found by six linear modularization criteria. We saw that two criteria who have a **local definition** are based on a the **density of within-cluster edges** (Zahn-Condorcet and Owsiński-Zadrożny), whereas others are based on a **null model** (Newman-Girvan, Deviation to Uniformity, Deviation to Indetermination and the Balanced Modularity). These criteria have a **resolution limit**.

Thanks for your attention!