Comparing linear modularization criteria using the relational notation

Patricia Conde Céspedes

Université Pierre et Marie Curie
Laboratoire de Statistique Théorique et Appliquée

April 30th, 2014
## Table of contents

1. **Introduction and objective**
   - Definitions
   - The definition of community

2. **Relational approach**

3. **Contributions**

4. **Applications**

5. **Conclusions**
# Introduction and objective

- Definitions
- The definition of community

# Relational approach

- Mathematical Relational modelling
- Properties verified by linear modularization criteria

# Contributions

- Linear criteria
- Comparison of linear criteria

# Applications

- The generalized Louvain algorithm
- Examples: real large graphs

# 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 ⇐⇒ 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** $E$. 

- $N = |V|$ 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:

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

- Example: given a graph with $N = 6$ and $M = 7$, its adjacency matrix is:

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

- The **degree** $d_i$ of node $i$ is the number of edges incident to $i$: 

  $$ d_i = \sum_{i'} a_{ii'} = a_{ii} $$

- 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 $W$. 

---

Patricia Conde Céspedes

Univ. Pierre et Marie Curie Laboratoire de Statistique Théorique et Appliquée
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

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

(1)

Example: given a graph with $N = 6$ and $M = 7$. Its adjacency matrix is:

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

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

Example: given a graph with $N = 6$ and $M = 7$, its adjacency matrix is:

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

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

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

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

![Graph example](image)

Its adjacency matrix is:

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

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 $W$. 

---

Patricia Conde Céspedes
Université Pierre et Marie Curie Laboratoire de Statistique Théorique et Appliquée
The definition of community

What is a community?

Each criterion is based on its own definition of community.

Definitions of community:

- **Densely connected groups** of nodes with few connections between groups whose density is higher than the density of the whole graph.
- Set of nodes where the connections are not randomly distributed.
- many others...

The words **community**, **module** and **cluster** will be treated as synonyms.
Table of contents

1 Introduction and objective

2 Relational approach
   • Mathematical Relational modelling
   • Properties verified by linear modularization criteria

3 Contributions

4 Applications

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

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

(2)
Mathematical Relational modelling

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

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

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

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

subject to the constraints of an equivalence relation:

$$x_{ii'} \in \{0, 1\} \quad \text{Binarity}$$

$$x_{ii} = 1 \quad \forall i \quad \text{Reflexivity}$$

$$x_{ii'} - x_{i'i} = 0 \quad \forall (i, i') \quad \text{Symmetry}$$

$$x_{ii'} + x_{i'i''} - x_{i''} \leq 1 \quad \forall (i, i', i'') \quad \text{Transitivity}$$

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:

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

\[ \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) x_{ii'} + K \]
A criterion is linear if it can be written in the general form:

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

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

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

where $K$ is any constant depending only on the original data; $\phi(a_{ii'}) \geq 0 \forall i, i'$ and $\bar{\phi}(a_{ii'}) \geq 0 \forall i, i'$ are non negative functions verifying:

$$\sum_{i=1}^{N} \sum_{i'=1}^{N} \phi_{ii} > 0 \text{ and } \sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{\phi}_{ii} > 0;$$

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 $\kappa$ denote the number of clusters obtained after optimization of the criterion.

Given a graph
The impact of the property of General balance

Let $\kappa$ denote the number of clusters obtained after optimization of the criterion.

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

Given a graph

then all nodes are clustered together, $\kappa = 1$
The impact of the property of General balance

Let $\kappa$ 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.
Properties verified by linear modularization criteria

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.

### Property of Global balance

A balanced linear criterion whose functions $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ satisfy

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

where $K_G$ is a constant depending on global properties of the graph has the property of global balance.
Contribution: Different levels of general balance for linear criteria

Null Model

A global balanced linear criterion whose functions $\phi_{ii'}$ and $\bar{\phi}_{ii'}$ satisfy the following conditions:

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

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

where $K_G$ is a constant depending on global properties of the graph is a criterion based on a null model.
Different levels of general balance for linear criteria

Figure: Venn Diagram of criteria verifying the property of general balance.
Resolution limit [Fortunato and Barthelemy (2006)]

A null model has a resolution limit.

Figure: A null model may fail to identify modules smaller than a scale depending on the global network, even cliques (complete graphs with $m$ nodes $k_m$) connected by single edges.
Table of contents

1. Introduction and objective

2. Relational approach

3. Contributions
   - Linear criteria
   - Comparison of linear criteria

4. Applications

5. Conclusions
Some linear criteria in relational notation

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Relational notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zahn-Condorcet (1785, 1964)</td>
<td>( F_{ZC}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} (a_{ii'} x_{ii'} + \bar{a}<em>{ii'} \bar{x}</em>{ii'}) )</td>
</tr>
<tr>
<td>Owsiński - Zadrożny (1986)</td>
<td>( F_{ZOZ}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} ((1 - \alpha) a_{ii'} x_{ii'} + \alpha \bar{a}<em>{ii'} \bar{x}</em>{ii'}) ) with ( 0 &lt; \alpha &lt; 1 )</td>
</tr>
<tr>
<td>Newman-Girvan (2004)</td>
<td>( 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'} )</td>
</tr>
</tbody>
</table>

Table: Relational notation of linear modularity functions.
### Linear criteria

The following table presents some linear criteria in relational notation, along with their respective formulas.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Relational notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation to Uniformity (2013)</td>
<td>( F_{UNIF}(X) = \frac{1}{2M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( a_{ii'} - \frac{2M}{N^2} \right) x_{ii'} )</td>
</tr>
<tr>
<td>Deviation to Indetermination (2013)</td>
<td>( 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'} )</td>
</tr>
<tr>
<td>The Balanced Modularity (2013)</td>
<td>( F_{BM}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( (a_{ii'} - P_{ii'}) x_{ii'} + (\bar{a}<em>{ii'} - \bar{P}</em>{ii'}) \bar{x}_{ii'} \right) )</td>
</tr>
</tbody>
</table>

where \( P_{ii'} = \frac{a_i a_{i'}}{2M} \) and \( \bar{P}_{ii'} = \left( \frac{\bar{a}_{ii'} - (N-a_i)(N-a_{i'})}{N^2-2M} \right) \).

**Table:** Relational notation of linear modularity functions.
Existing linear criteria: Deviation to a particular structure

Principe: A *random graph does not have community structure.*

**Uniformity structure**

**Independence structure**
- Multiplicative model:
  - Geometric mean

**Indetermination structure**
- Additive model:
  - Arithmetic mean
Properties of these linear criteria

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

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Global balance</th>
<th>Local Balance</th>
<th>Null model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zahn-Condorcet</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Owsiński-Zadrożny</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Newman-Girvan</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Deviation to Uniformity</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Deviation to Indetermination</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Balanced modularity</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

**Table:** Balance Property for Linear criteria
The number of clusters

The number of clusters found by the generalized Louvain algorithm for two real networks.

<table>
<thead>
<tr>
<th>Network</th>
<th>Jazz</th>
<th>Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N = 198)</td>
<td>38</td>
<td>40,123</td>
</tr>
<tr>
<td>(M = 2,742)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N = 69,949)</td>
<td>6</td>
<td>456</td>
</tr>
<tr>
<td>(M = 351,380)</td>
<td>(\alpha = 2%)</td>
<td>(\alpha &lt; 1%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criterion</th>
<th>(\kappa)</th>
<th>(\kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zahn-Condorcet</td>
<td>38</td>
<td>40,123</td>
</tr>
<tr>
<td>Owsiński-Zadrożny</td>
<td>6 (\alpha = 2%)</td>
<td>456 (\alpha &lt; 1%)</td>
</tr>
<tr>
<td>Deviation to Uniformity</td>
<td>20</td>
<td>173</td>
</tr>
<tr>
<td>Newman-Girvan</td>
<td>4</td>
<td>46</td>
</tr>
<tr>
<td>Deviation to Indetermination</td>
<td>6</td>
<td>39</td>
</tr>
<tr>
<td>Balanced Modularity</td>
<td>5</td>
<td>41</td>
</tr>
</tbody>
</table>

How to explain these differences?
### First approach: the deviation form notation

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zahn-Condorcet</td>
<td>$F_{ZC}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( a_{ii'} - \frac{1}{2} \right) x_{ii'}$</td>
</tr>
<tr>
<td>Owsiński-Zadrożny</td>
<td>$F_{OZ}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} (a_{ii'} - \alpha)x_{ii'}$</td>
</tr>
<tr>
<td>Deviation to uniformity</td>
<td>$F_{UNIF}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( a_{ii'} - \frac{2M}{N^2} \right) x_{ii'}$</td>
</tr>
<tr>
<td>Newman-Girvan</td>
<td>$F_{NG}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( a_{ii'} - \frac{a_{i.i'}^{2}}{2M} \right) x_{ii'}$</td>
</tr>
<tr>
<td>Deviation to indetermination</td>
<td>$F_{DI}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left( a_{ii'} - \left( \frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2} \right) \right) x_{ii'}$</td>
</tr>
</tbody>
</table>
Comparison of linear criteria

First approach: the deviation form notation

The (+) agreements term is the same for the 5 criteria:

$$\phi_{ii'}^{ZC} = \phi_{ii'}^{OZ} = \phi_{ii'}^{NG} = \phi_{ii'}^{DI} = \phi_{ii'}^{UNIF} = a_{ii'} \forall (i, i').$$

So we have to analyse the (-) agreements term. We distinguish 2 cases:
First approach: the deviation form notation

The (+) agreements term is the same for the 5 criteria:

\[
\phi^{ZC}_{ii'} = \phi^{OZ}_{ii'} = \phi^{NG}_{ii'} = \phi^{DI}_{ii'} = \phi^{UNIF}_{ii'} = a_{ii'} \forall (i, i').
\]

So we have to analyse the (-) agreements term. We distinguish 2 cases:

**Case 1:** \(\bar{\phi}\) is a constant not depending on each pair of nodes.
Thas is the case of Zahn-Condorcet, Owsiński-Zadrożny and Deviation to Uniformity criteria.
First approach: The degree distribution

Case 2: $\phi$ depends on the degree of each pair of nodes as for Newman-Girvan and Deviation to Indetermination criteria.
First approach: The degree distribution

**Case 2:** \( \overline{\phi} \) depends on the degree of each pair of nodes as for Newman-Girvan and Deviation to Indetermination criteria.

Many real networks fall into the class of *scale-free networks*, meaning that their degree distribution follows a **power-law** which distinguish them from random networks [Barabasi and Albert (1999)].

In a scale-free network a few nodes called **hubs** have many connexions whereas most nodes have few connexions.

![Random network](image1)

![Scale-free network](image2)
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 $C_1$ and $C_2$ 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:

$$C = \sum_{i \in C_1}^{n_1} \sum_{i' \in C_2}^{n_2} (\phi_{ii'} - \bar{\phi}_{ii'})$$  \hspace{1cm} (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:

$$C = \sum_{i \in C_1}^{n_1} \sum_{i' \in C_2}^{n_2} (\phi_{ii'} - \bar{\phi}_{ii'})$$

(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.

<table>
<thead>
<tr>
<th>Criterion: $F$</th>
<th>$C_F = \sum_{i \in C_1} \sum_{i' \in C_2} (\phi_{ii'} - \bar{\phi}_{ii'})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zahn-Condorcet</td>
<td>$C_{ZC} = (l - \frac{n_1 n_2}{2})$</td>
</tr>
<tr>
<td>Owsiński-Zadrożny</td>
<td>$C_{OZ} = (l - n_1 n_2 \alpha)$ for $0 &lt; \alpha &lt; 1$</td>
</tr>
<tr>
<td>Deviation to Uniformity</td>
<td>$C_{UNIF} = (l - n_1 n_2 \delta)$</td>
</tr>
<tr>
<td>Newman-Girvan</td>
<td>$C_{NG} = (l - n_1 n_2 \frac{d_{av}^1 d_{av}^2}{2M})$</td>
</tr>
<tr>
<td>Deviation to Indetermination</td>
<td>$C_{DI} = (l - n_1 n_2 \left( \frac{d_{av}^1}{N} + \frac{d_{av}^2}{N} - \frac{2M}{N^2} \right))$</td>
</tr>
</tbody>
</table>
### Comparison of linear criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Characteristics of the clustering</th>
</tr>
</thead>
</table>
| 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-Zadrożyński       | - It gives the choice to define the minimum required within-cluster **density**, $\alpha$.  
|                            | - For $\alpha = 0.5$ the Owsiński-Zadrożyński criterion $\equiv$ the Zahn-Condorcet criterion.  
|                            | - No resolution limit.                                                                         |
| Deviation to Uniformity    | - A particular case of Owsiński-Zadrożyński criterion with $\alpha = \delta$.  
|                            | - The **density** of within cluster edges of each cluster is at least $\delta$.  
|                            | - It has a resolution limit.                                                                   |
### Summary by criterion

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Characteristics of the clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newman-Girvan</td>
<td>• It has a resolution limit.</td>
</tr>
<tr>
<td></td>
<td>• The contribution depends on the <strong>degree distribution</strong> of the clusters.</td>
</tr>
<tr>
<td></td>
<td>• The optimal partition has no single nodes.</td>
</tr>
<tr>
<td>Deviation to Indetermination</td>
<td>• It has a resolution limit.</td>
</tr>
<tr>
<td></td>
<td>• The contribution depends on the <strong>degree distribution</strong> of the clusters.</td>
</tr>
<tr>
<td>Balanced modularity</td>
<td>• It has a resolution limit.</td>
</tr>
<tr>
<td></td>
<td>• The contribution depends on the <strong>degree distribution</strong> of the clusters.</td>
</tr>
<tr>
<td></td>
<td>• Depending upon $\delta$ and $d_{av}$ this criterion behaves like a <strong>regulator</strong> between the NG criterion and the DI criterion.</td>
</tr>
</tbody>
</table>
Table of contents

1. Introduction and objective

2. Relational approach

3. Contributions

4. Applications
   - The generalized Louvain algorithm
   - Examples: real large graphs

5. Conclusions
Modularizing large graphs with the generalized Louvain algorithm

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

<table>
<thead>
<tr>
<th></th>
<th>Web nd.edu</th>
<th>Amazon</th>
<th>Youtube</th>
<th>WebBase01</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N \sim$</td>
<td>325k</td>
<td>334k</td>
<td>1M</td>
<td>118M</td>
</tr>
<tr>
<td>$M \sim$</td>
<td>1M</td>
<td>925k</td>
<td>3M</td>
<td>1B</td>
</tr>
<tr>
<td>$\delta$</td>
<td>$2.77 \times 10^{-05}$</td>
<td>$1.65 \times 10^{-05}$</td>
<td>$4.64 \times 10^{-06}$</td>
<td>$1.46 \times 10^{-07}$</td>
</tr>
<tr>
<td>Criterion</td>
<td>$\kappa$</td>
<td>$\kappa$</td>
<td>$\kappa$</td>
<td>$\kappa$</td>
</tr>
<tr>
<td>ZC</td>
<td>201,647</td>
<td>161,439</td>
<td>878,849</td>
<td>71,806,729</td>
</tr>
<tr>
<td>UNIF</td>
<td>711</td>
<td>265</td>
<td>51,584</td>
<td>2,777,580</td>
</tr>
<tr>
<td>NG</td>
<td>511</td>
<td>250</td>
<td>5,567</td>
<td>2,759,248</td>
</tr>
<tr>
<td>DI</td>
<td>324</td>
<td>246</td>
<td>13,985</td>
<td>2,770,647</td>
</tr>
<tr>
<td>BM</td>
<td>333</td>
<td>230</td>
<td>6,410</td>
<td>2,736,808</td>
</tr>
</tbody>
</table>

Ref: Zahn-Condorcet (ZC), Deviation to Uniformity (UNIF), Newman-Girvan (NG), Deviation to Indetermination(DI) and Balanced Modularity (BM)
**Running times obtained for each criterion**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>M</th>
<th>NG</th>
<th>ZC</th>
<th>DI</th>
<th>UNIF</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>karate</td>
<td>34</td>
<td>77</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>arxiv</td>
<td>9k</td>
<td>48k</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>internet</td>
<td>70k</td>
<td>351k</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>webndu</td>
<td>325k</td>
<td>2M</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>webuk-2005</td>
<td>39M</td>
<td>783M</td>
<td>232</td>
<td>468</td>
<td>201</td>
<td>192</td>
<td>178</td>
</tr>
<tr>
<td>webbase-2001</td>
<td>118M</td>
<td>1B</td>
<td>430</td>
<td>1,043</td>
<td>446</td>
<td>394</td>
<td>324</td>
</tr>
</tbody>
</table>

**Table:** Running times (in seconds) obtained for each criterion
Table of contents

1. Introduction and objective
2. Relational approach
3. Contributions
4. Applications
5. Conclusions
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.
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!