Optimal Transport and Minimal Trade Problem, Impacts on Relational Metrics and Applications to Large Graphs and Networks Modularity

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

[1] Thales Communications et Sécurité, TCS, Gennevilliers, France and (LSTA)
Université Pierre et Marie Curie, Paris, France
jeanfrancois.marcotorchino@thalesgroup.com

[2] Laboratoire de Statistique théorique et Appliquée (LSTA), Université Pierre et Marie Curie, Paris, France
patricia.conde_cespedes@upmc.fr

Abstract. This article presents a summary of the principal results found in [MAR13]. Starting with the seminal works on transportation theory of G. Monge and L. Kantorovich, while revisiting the works of Maurice Fréchet, we will introduce direct derivations of the optimal transport problem such as the so-called *Alan Wilson's Entropy Model* and the *Minimal Trade Problem*. We will show that optimal solutions of those models are mainly based in two dual principles: the independance and the indetermination structure between two categorical variables. Thanks to Mathematical Relational Analysis representation and the Antoine Caritat's (Condorcet) works on Relational Consensus, we will give an interesting interpretation to the indeterminaion structure and underline the duality Relationship between deviation to independence and deviation to indetermination structures. Finally, these results will lead us to the elaboration of a new criterion of modularization for large networks.

Keywords. Optimal Transport, Monge-Kantorovich problem, Minimal Trade Problem, independance structure, indetermination structure, Relational Analysis, Condorcet, Modularity, Large Graphs, Complex Networks.

1 Introduction

The main purpose of this article is to link the optimal transport problem to problems originated in different fields, such as the Minimal Trade Problem and the exchange Entropy Model. This allows to elaborate a new correlation measure between discrete structures.

2 The optimal transport problem: Monge and Monge-Kantorovich Problems

Gaspard Monge studied first the optimal transport problem in his article "Mémoire sur la théorie des déblais et des remblais" (1781) [MON81]. Later in 1942, Leonid

Kantorovich [KAN42] proposed a relaxation of the original problem, leading to the famous MKP Problem "Monge Kantorovich Problem" and mainly the "Kantorovich duality" (for more details the interested reader can see [VIL04], [VIL03], [CAR10] and [EVA97]).

Given two separable metric subsets of $\mathbb{R}^n$: X and Y, the generalized Monge's problem consists in finding a transport map $T : X \to Y$ of mass initially located at $x \in X$) to the destination $T(x) \in Y$ at a minimum cost $c(x, y)$. The distribution of the initial mass and of the destination mass are represented by the probability measures μ and ν respectively (where the mass conservation constraint entails: $0 < \mu(X) = \nu(Y) \leq 1)$[3]. This problem is very complex to solve and quite rigid.

In 1942, Leonid Kantorovich proposed a relaxation of the original problem leading to *the famous Monge-Kantorovich's problem* (MKP); this re-formulation is given by the following minimization problem:

$$P[\pi^*] = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X\times Y} c(x, y) d\pi(x, y) \tag{1}$$

where $\Pi(\mu, \nu)$ is the set of Borel probability measures on the product set: $X \times Y$ which have μ, and ν as marginal probabilities.

The linear Monge-Kantorovich problem has a dual formulation, essential to establish the existence of optimal maps for certain cost functions, given by:

$$D[\varphi, \psi] = \sup_{(\varphi,\psi)} \{ \int_X \varphi d\mu + \int_Y \psi d\nu \, : \, c(x, y) \geq \varphi(x) + \psi(y) \, \text{on} \, X \times Y \} \tag{2}$$

Let us define $\pounds = \{(\varphi, \psi) | \varphi, \psi : X \to \mathbb{R}, Y \to \mathbb{R}$ as the set of continuous mappings of class $\mathcal{C}^1$ such as $\varphi(x) + \psi(y) \leq c(x, y)\}$. Now, the primal and dual formulation (formula (1) and (2) respectively) lead us to the following theorem:

Theorem 1 (Kantorovich duality).
If there exists $\pi^* \in \Pi(\mu, \nu)$ and an admissible pair $(\varphi^*, \psi^*) \in \pounds$ such that:

$$\int_{X\times Y} c(x, y) d\pi^*(x, y) = \int_X \varphi^*(x) d\mu(x) + \int_Y \psi^*(y) d\nu(y)$$

then π^* is an Optimal Transport Plan and the pair (φ^*, ψ^*) solves the problem (2). So there is no gap between the values: $\inf_{\pi} P[\pi] = \sup_{(\varphi,\psi)} D[\varphi, \psi]$

In the next sections we will consider mostly the discrete version of the Monge-Kantorovich problem, this will allow us to link the works of Gaspard Monge, Maurice Fréchet and Antoine Caritat de Condorcet.

[3] To satisfy this constraint ν must be the *forward Measure* of μ by the transport map T, which is written $\nu = T\#\mu$.

3 Extensions and variants of the MKP problem

In this section we will introduce direct derivations from the Monge and Monge-Kantorovich problems. Most of them originated from contexts having a priori nothing to do with the pure transport problem. First, we present the discrete version.

Let $X = \{1, ..., p\}$ be a set of p *origins* and let $Y = \{1, ..., q\}$ represent a set of q *destinations*. Our objective is to transport a fixed total quantity of goods located initially at X origins to Y destinations. The available data are the *unit transportation cost* associated with the transfer from an origin u to a destination v: $c(u, v)$; the discrete probability distribution of the quantities located at the p *origins*: $\{\mu_1, ..., \mu_p\}$, and the discrete probability distribution of the quantities to be delivered to the q *destinations*: $\{\nu_1, ..., \nu_q\}$. Those quantities can be identified as marginal probability distributions relative to the unkown bi-dimensional probability distribution : π_{uv}. In a balanced situation μ and ν satisfy the *mass preserving* constraint (normalized to 1):

$$\sum_{u=1}^{p} \mu_u = \sum_{v=1}^{q} \nu_v = 1 \tag{3}$$

Finally, our problem is to find π^*_{uv} minimizing the total cost:

$$\min_{\pi} \sum_{u=1}^{p} \sum_{v=1}^{q} c(u, v)\pi_{uv} \tag{4}$$

subject to:

$$\sum_{v=1}^{q} \pi_{uv} = \mu_u \quad \forall u \in \{1, 2, ..., p\} \tag{5}$$

$$\sum_{u=1}^{p} \pi_{uv} = \nu_v \quad \forall v \in \{1, 2, ..., q\} \tag{6}$$

$$\pi_{uv} \geq 0 \quad \forall u \in \{1, ..., p\}; v \in \{1, ..., q\} \tag{7}$$

Indeed, this is the discrete version of the the MKP (see [EVA97] where π_{uv} plays the role of $\pi(x, y)$ in (1) for the continuous case. We can find this problem in various contexts: in the International Trade Exchange Models, we deal with monetary exchanges between countries or geographical zones; in the ”Spatial Interaction Models”, the figures represent people travelling from a region to another one.

These exchanges system can be described by a rectangular table crossing the p origins (in lines) and the q destinations (in columns). Besides, by multiplying each value of such a table by the total quantity exchanged N and if all the figures are integer we get a *Contingency table* where:

$n_{uv} = N\pi_{uv}$: quantity of mass transported from $u \in X$ to $v \in Y$; $n_{u.} = N\pi_{u.}$: total mass located originaly at u; $n_{.v} = N\pi_{.v}$: total mass transported to v and $n_{..} = N$ Total exchange mass.

In this case, the inequalities (3), (5), (6) and (7) are the basic constraints of the so called *Contingency Cells Adjustment to fixed margins*, first studied by [DES40] and later by Maurice Fréchet (see [FRE51] and [FRE60]).

4 Optimizing transport problem

In this section we consider two particular cost functions of the discrete transport problem (4): the *Alain Wilson's Entropy Model* and the *Minimal Trade Model*. In both cases, the cost is a function of the unkown joint distribution $h(\pi)$. Their optimal solution will be deeply studied in the next sections due to the important interpretation and duality they carry on in multivariate statistics and theory of contingency.

1. **Alan Wilson's Entropy Model:** $h(\pi_{uv}) = \ln \pi_{uv}$: The *Flows Entropy Model* of Alan Wilson was introduced in [WIL67] (see [WIL69] and [WIL70]) for *Spatial Interaction Modeling*. In his approach he considers a system whose elements do not maintain affinities, the purpose is to determine the distribution of the normalized frequency flows π_{uv} (supposing $\pi_{uv} > 0 \, \forall \, u, v$) which maximizes the entropy of the system. The objective function to be maximized is based upon the Boltzmann's or Shannon's Entropies:

$$\max_{\pi} - \sum_{u=1}^{p} \sum_{v=1}^{q} \pi_{uv} \ln \pi_{uv} \tag{8}$$

   Such a problem is called *Program of Spatial Interaction System* (PSIS). The optimal solution is obtained by using the Lagrange's multipliers to maximize (8) subject to the contraints (5), (6) and (7). The explicit expression of the optimal solution is shown in table 1, the degree of disorder has drastically reduced. The flow maximizing entropy reveals statistical independence between the p suppliers and the q clients. Indeed, without specific affinities between row and columns, fixing any column v the flow from a given line to this column is proportional to its total marginals in the whole population.

2. **The Minimal Trade Model:** $h(\pi_{uv}) = \pi_{uv}$
   In the Minimal Trade Model, the criterion is a quadratic function measuring the squared deviation of the cells values from the *no information* situation (the uniform joint distribution) in order to get a smooth ventilation of the origins-destinations n_{uv} values subject to the balanced marginals and mass preserving constraints (5), (6) and (7) (see [STE77], [MAR84]). We are dealing, then, with a least squared problem in order to get a smooth ventilation of the origins-destinations n_{uv} values (this explains the term *Minimal Trade*).

We solve this problem by using the Lagrange multipliers, since the function to optimize is convex, we are looking for a minimum. The optimal solution [4],[5] is shown in table 1. The following inequality for the marginal values (see [MAR84]) garantees the positivity of the optimal values π^*_{uv}:

$$p \min_u \mu_u + q \min_v \nu_v \geq 1 \Longrightarrow p \min_u n_{u.} + q \min_v n_{.v} \geq N \tag{9}$$

From now on we assume this condition is true. Furthermore this inequality guarantees $0 \leq \pi_{uv} \leq 1$. The optimal solution reveals ”indetermination structure” between the p suppliers and the q clients. This concept of ”indetermination structure” will be studied in details later on.

Table 1. Variants of the MKP problem

Model	**Objective function**	**Subject to**	**Optimal solution**
Alan Wilson's Entropy Model	$\max_\pi - \sum_{u=1}^{p} \sum_{v=1}^{q} \pi_{uv} \ln \pi_{uv}$	Contraints (5),(6) and (7)	$\pi^*_{uv} = \mu_u \nu_v \forall (u,v)$ $n^*_{uv} = \frac{n_{u.} n_{.v}}{N}$
The Minimal Trade Model	$\min_\pi \sum_{u=1}^{p} \sum_{v=1}^{q} \left(\pi_{uv} - \frac{1}{pq}\right)^2$	Contraints (5),(6) and (7)	$\pi^*_{uv} = \frac{\mu_u}{q} + \frac{\nu_v}{p} - \frac{1}{pq}$ $n^*_{uv} = \frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq}$

5 Monge and Anti-Monge matrices and some related structural properties

Monge's condition for matrices was originally studied by Gaspard Monge[6] in his well known *Mémoire sur la théorie des déblais et des remblais*[MON81]. From this condition, it is possible to derive Anti-Monge's condition, both conditions are defined as follows:

Definition 1 *A $p \times q$ real matrix $\{c_{uv}\}$ is called a Monge matrix, if $\boldsymbol{C}$ satisfies the so called Monge's property:*

$$c_{uv} + c_{u'v'} \leq c_{uv'} + c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q \tag{10}$$

[4] Notice that for *the Contingency Adjustment to Fixed Margins* case, the associated values n_{uv} must be integer, the interested reader can find in [MAR84] a specific paragraph on the conditions to add to get pure integers.

[5] The Continuous version of the Minimal Trade Problem is treated in [MAR13]. The optimal solution, obtained by considering the Kantorovich duality (2), is given by:

$$\pi^*(x,y) = \frac{f(x)}{B} + \frac{g(y)}{A} - \frac{1}{AB} \quad \forall\,(x,y) \in [a,b] \times [c,d]$$

where $\pi : [a,b] \times [c,d] \longrightarrow [0,1]$ is defined on the product of two closed intervals of the cartesian plan; $A = (b-a)$ and $B = (d-c)$ are respective the lengths of these intervals; μ and ν (the marginals of π) have densities f and g respectively.

[6] This notion of Monge matrix has been coined by [HOF63].

Reciprocally, an "Inverse Monge Matrix" (or Anti Monge matrix) $\boldsymbol{C}$ *satisfies the following inequality:*

$$c_{uv} + c_{u'v'} \geq c_{uv'} + c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q \tag{11}$$

In case both inequalities (10) and (11) hold:

$$c_{uv} + c_{u'v'} = c_{uv'} + c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q \tag{12}$$

The last property (12) is very important since it corresponds to the so called *Indetermination or indecision structure*, which plays an important role in Relational Analysis Theory when applied to Condorcet's Voting Theory, or Central Partition Clustering.
A frequency matrix (or a contingency table if we multiply every entry by N) satisfaying both Monge and Anti-Monge conditions verifies interesting properties formulated in the following theorem:

Theorem 2. *Let* $\{\pi_{uv}\}$ *be a* $p \times q$ *real nonnegative frequency Matrix, then the following properties hold and are equivalent:*

i) If $\{\pi_{uv}\}$ *is a Monge and Anti-Monge Matrix then:*
$\pi_{uv} + \pi_{u'v'} = \pi_{uv'} + \pi_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q$

ii) $\pi_{uv} = \left(\dfrac{\mu_u}{q} + \dfrac{\nu_v}{p} - \dfrac{1}{pq}\right)$ *is a minimizer of the* Minimal Trade Model.

iii) All the sub tables $\{u, v, u, v\}$ *of size* 2×2 *with* $1 \leq u < u' \leq p$, $1 \leq v < v' \leq q$ *have the sum of their diagonals equal to the sum of their anti-diagonals.*

Other interesting properties can be derived from those Monge and Anti Monge conditions, which concern positive matrices only (i.e. $c_{uv} > 0 \forall u, v$):

Definition 2 *A* $p \times q$ *positive real matrix* $\{c_{uv}\}$ *is called a Log Monge matrix, if* $\boldsymbol{C}$ *satisfies the Log-Monge's property:*

$$\ln c_{uv} + \ln c_{u'v'} \leq \ln c_{uv'} + \ln c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q \tag{13}$$

Reciprocally, an "Inverse Log-Monge Matrix" (or Log-Anti-Monge matrix) $\boldsymbol{C}$ *satisfies the following inequality:*

$$\ln c_{uv} + \ln c_{u'v'} \geq \ln c_{uv'} + \ln c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q \tag{14}$$

In case both inequalities (13) and (14) hold:

$$\ln c_{uv} + \ln c_{u'v'} = \ln c_{uv'} + \ln c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q \tag{15}$$

Property (15) corresponds to the situation of statistical independence, because from it we can derive: $c_{uv}c_{u'v'} = c_{uv'}c_{u'v} \,\forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q$.

From the previous explanations, we can derive the following theorem concerning Log-Monge and Log Anti Monge matrices:

Theorem 3. *Let $\{\pi_{uv}\}$ be a $p \times q$ real positive frequency Matrix, then the following properties hold and are equivalent:*

i) If $\{\pi_{uv}\}$ is a Log-Monge and Log-Anti-Monge Matrix then:
$\ln \pi_{uv} + \ln \pi_{u'v'} = \ln \pi_{uv'} + \ln \pi_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q$

ii) $\pi_{uv} = \mu_u \nu_v \quad \forall\, 1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q$ is a minimizer of the Alan Wilson's Program of Spatial Interaction System based upon Entropy Model, with fixed Margins.

iii) All the sub tables $\{u, v, u, v\}$ of size 2×2 with $1 \leq u < u' \leq p,\, 1 \leq v < v' \leq q$ have the product of their diagonal terms equal to the product of their anti-diagonals terms.

6 Duality related to "Independence" and "Indetermination" structures

An important number of statistical indexes and criteria have been proposed in the scientific literature for measuring the relationships between two categorical variables. Among those indexes, some (very well known) are built up basically by measuring their deviation to the situation of "independence" and some (lesser known) are built by measuring their deviation to the situation of "indetermination". The reader will find the explicit expressions of those indexes as well as their behavour in case of independence or in case of indetermination structures in [MAR84] and [MAR13]. In this summary we will focus on the Mutual Information index and the Deviation to indetermination Index. First of all, we define the quantities:

- **The Mutual Information index (MI):** it compares the Alan Wilson's Entropy $S(X,Y) = -\sum_{u=1}^{p}\sum_{v=1}^{q} \pi_{uv} \ln \pi_{uv}$ to $S(X) = -\sum_{u=1}^{p} \mu_u \ln \mu_u$ and $S(Y) = -\sum_{v=1}^{q} \nu_v \ln \nu_v$. The explicit expression of the *mutual information index* ρ_{MI} is given in table (2). It represents the quantity of information which is present (in duplication) into X and Y simultaneously. Clearly this index measures the departure from independence and behaves nearly as the χ^2 criterion does in the neighborhood of independence.

- **The Deviation to indetermination Index (IND):** it is null if and only if the variables verify the indetermination structure. It compares the Minimal Trade Criterion multiplied by pq: $K(X,Y) = pq \sum_{u=1}^{p}\sum_{v=1}^{q} \left(\pi_{uv} - \frac{1}{pq}\right)^2$ to the quantities $K(X) = p \sum_{u=1}^{p} \left(\mu_u - \frac{1}{p}\right)^2$ and $K(Y) = q \sum_{v=1}^{q} \left(\nu_v - \frac{1}{q}\right)^2$. In fact, $K(X,Y)$ plays a role analogue to that of the Entropy $S(X,Y)$ for the Independence Structure. The explicit expression of the *deviation to indetermination index* ρ_{IND} is given in table (2)

Table 2. Duality between independence and indetermination structures

The Independence case	**The Indetermination case**
$S(X,Y) \leq S(X) + S(Y)$ $\forall X,Y \vert X \sim \mu\,, Y \sim \nu\,, (X,Y) \sim \pi$ with equality in case of independence $\rho_{\mathrm{MI}}(X,Y) = S(X) + S(Y) - S(X,Y)$ $\rho_{\mathrm{MI}}[\pi] = \sum_{u=1}^{p}\sum_{v=1}^{q} \pi_{uv} \ln\left(\frac{\pi_{uv}}{\mu_u \nu_v}\right)$	$K(X,Y) \geq K(X) + K(Y)$ $\forall X,Y \vert X \sim \mu\,, Y \sim \nu\,, (X,Y) \sim \pi$ with equality in case of indetermination $\rho_{\mathrm{IND}}(X,Y) = K(X,Y) - K(X) - K(Y)$ $\rho_{\mathrm{IND}}[\pi] = pq \sum_{u=1}^{p}\sum_{v=1}^{q} \left(\pi_{uv} - \frac{\mu_u}{q} + \frac{\nu_v}{p} + \frac{1}{pq}\right)^2$

7 Relational Analysis Approach

We assume the reader is familiar with Relational Analysis theory (the unfamiliar reader can see [MAM79], [MAR84], [MIC87], [MAY91], [AHM07], etc...). The principle on which Relational Analysis (RA) is based consists in representing relations between objects by binary coding. For the discrete transport problem, the variables *origins* and *destinations* split the set of objects in p and q clusters respectively of sizes defined by marginal distributions. In RA a partition is nothing but an equivalence relation on the set of objects, which is represented by a relational $N \times N$ matrix $\mathbf{X}$, whose entries are defined as follows:

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ belong to the same cluster.} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

As X is an equivalence Relation, it must be Reflexive, Symmetric and Transitive, those properties can be turned into linear constraints on the general terms of the relational matrix X. We define, as well, the inverse relation of X by: $\bar{x}_{ij} = 1 - x_{ij}\ \forall (i,j)$.

Using the "Relational Transfer Principle" (see [KEN70] and [MAR84]) we can write the deviation to indetermination index in relational notation (where X and Y represent the Relational matrices of the two variables):

$$N^2 \rho_{\mathrm{IND}}(X,Y) = pq \sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij} y_{ij} - p \sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij} - q \sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij} + N^2 \tag{17}$$

The quantities $\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij} y_{ij}$ and $\sum_{i=1}^{N}\sum_{j=1}^{N} \bar{x}_{ij}\bar{y}_{ij}$ represent the situations where X and Y are in *agreement* with respect to the population, whereas, $\sum_{i=1}^{N}\sum_{j=1}^{N} \bar{x}_{ij} y_{ij}$ and $\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}\bar{y}_{ij}$ represent the configurations where X and Y are in *disagreement*. It has been shown in [MAR85] that if X and Y are in an "indeterminate" cross relationship, the mean (cell by cell) of their Agreements is equal to the mean of their disagreements (and ρ_{IND} is null in that case, see (17), therefore we are faced with a situation of complete *indetermination* or *indecision*. If $p = q = 2$ the ρ_{IND} index becomes $(2\rho_{\mathrm{RAND}} - 1)$, that is, an affine function of the well known *Rand index*, which, in relational notation is equal to the famous Condorcet's Criterion $C(X,Y)$ normalized by N^2. Therefore, in the situation of indetermination $(2\rho_{\text{RAND}} - 1) = 0 \Rightarrow \rho_{\text{RAND}} = \frac{C(X,Y)}{N^2} = \frac{1}{2}$. In voting theory, getting a value of the Condorcet Criterion equal to $\frac{1}{2}$ implies the existence of 50% of votes *in favor* and 50% of votes *against*, therefore, the indetermination or indecision situation.

There is an interesting duality relationship between the *"deviation to independence"* (numerator of the χ^2 index), and the *"deviation to indetermination"* (numerator of the Janson-Vegelius's index) (see [JAV82] and [AHM07]), which induces a reverse formalism while transforming them from contingency notations into relational ones using the following transfer formulas:

$$\boxed{\begin{gathered}\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}y_{ij}\ ;\ \sum_{u=1}^{p} n_{u.}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij};\ \sum_{v=1}^{q} n_{v.}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij};\\ \sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}n_{u.}n_{.v} = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{x_{i.}+x_{.j}}{2}\right)y_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_{i.}+y_{.j}}{2}\right)x_{ij};\\ \sum_{u=1}^{p}\sum_{v=1}^{q}\frac{n_{uv}^2}{n_{u.}n_{.v}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{x_{ij}}{x_{i.}}\frac{y_{ij}}{y_{.j}};\ \text{where } x_{i.} = \sum_{i=1}^{N} x_{ij};\ y_{.j} = \sum_{j=1}^{N} y_{ij}\end{gathered}}$$

we get the figures shown in table (3):

Table 3. Duality between deviation to independence and deviation to indetermination

	Contingency coding	Relational coding
numerator of the χ^2 index *Measuring the deviation to independence*	$\sum_{u=1}^{p}\sum_{v=1}^{q}\left(n_{uv} - \frac{n_{u.}n_{.v}}{N}\right)^2$	$\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_{ij} - \frac{x_{i.}}{N} - \frac{x_{.j}}{N} + \frac{x_{..}}{N^2}\right)\left(y_{ij} - \frac{y_{i.}}{N} - \frac{y_{.j}}{N} + \frac{y_{..}}{N^2}\right)$
Numerator of Janson-Vegelius's index *Measuring the deviation to indetermination*	$\sum_{u=1}^{p}\sum_{v=1}^{q}\left(n_{uv} - \frac{n_{u.}}{q} - \frac{n_{.v}}{p} + \frac{N}{pq}\right)^2$	$\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_{ij} - \frac{1}{p}\right)\left(y_{ij} - \frac{1}{q}\right)$

It appears clearly that the relational formalism induces a reverse structure for the indexes. The representation of independence in contingency space is translated into a representation of indetermination in relational space and vice versa.

8 Linear graph modularization criteria

Nowadays, the increasing use of social networks has considerably reinforced their complexity. Then, to analyze them it is necessary to decompose them in small homogeneous components. The process of splitting a network has received different names: graph clustering (in data analysis) or modularization. Different modularization criteria have been defined in the last few years. The most famous of them is the *Newman-Girvan modularity* (see [NEW04]). This criterion is based on the deviation to the independance structure. Considering the duality between *independance* and *indetermination* structures shown in the previous section, we introduce a new modularization criteria called the *deviation to indetermination index*. The expression of this two functions are given in the following table:

Table 4. Modularization criteria

Newman-Girvan modularity	**Deviation to indetermination index**
$\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}\right)x_{ii'}$	$\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{2M}{N^2}\right)x_{ii'}$

Where $a_{ii'}$ is the general term of the adjacency matrix of the graph $G(V, E)$; $N = |V|$ is the number of nodes, M is the number of edges and $x_{ii'}$ is the general term of the relational matrix defined by (16).

The behavior of this new criterion is close to that of Newman-Girvan modularity mainly due to the fact that they have many common properties: they are linear, separable and null models.

References

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

CAR10. **Carlier G.**: *"Optimal Transport and Economic Applications"*, Lecture Notes IMA, New Mathematical Models in Economics and Finance, pp:1-82, (2010).

DES40. **Deming W.E. , Stephan F.F.**: *On the least squares adjustment of a sampled frequency table when the expected marginal totals are known*, the Annals of Mathematical Statistics, Vol 11,pp: 427-444, (1940).

EVA97. **Evans L. C.**: *Partial Differential Equations and Monge-Kantorovich Mass Transfer*, Current Developments in Mathematics , S. T. Yau, Editor, (1997)

FRE51. **Fréchet M.**: *"Sur les Tableaux de Corrélations dont les Marges sont Données"*. Annales de l'Université de Lyon, Section. A, n 14, pp:53-77. (1951)

FRE60. **Fréchet M.**: "*Sur les Tableaux de Corrélations dont les Marges et les Bornes sont Données*". Revue de l'Institut de Statistique, n28, pp :10-32, (1960).

FUS04. **Fustier B.**: *Echanges commerciaux Euro-méditerranéens: essai d'analyse structurale*, Revue des Sciences Economiques et de Gestion n3, pp : 1-25, (2004).

HOF63. **Hoffman A.J.**: *On simple linear programming problems*, in Proceedings of Symposia in Pure Mathematics, V. Klee editor, Vol. VII, pp: 317-327, AMS, Providence, (1963).

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

KAN42. **Kantorovich L.**: "*On the translocation of masses*". Comptes Rendus (Doklady) Acad. Sci. URSS (N.S.), n37, pp:199201, (1942).

KEN70. **Kendall G.**: "*Rank correlation methods*". Griffin, Londres, 1970.

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

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

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

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

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

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

MON81. **Monge G.**: "*Mémoire sur la théorie des déblais et de remblais*". Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année, pp : 666704, (1781).

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

STE77. **Stemmelen E.**: *Tableaux d'Echanges, Description et Prévision*, Cahiers du Bureau Universitaire de Recherche Opérationnelle n28, Paris (1977).

VIL03. **Villani C.**: "*Topics in Optimal Transportation*", Graduate Studies in mathematics, Volume 58, The American Mathematical Society (2003).

VIL04. **Villani C.**: "*Transport Optimal de mesure : coup de neuf pour un très vieux problème*", Images des mathématiques, pp. 114-119, (2004)

WIL67. **Wilson A.G.**: *A statistical theory of spatial distribution models*, Transportation Research, Vol. 1, pp. 253-269, (1967).

WIL69. **Wilson A.G.**: *The use of entropy maximising models*, Journal of transport economies and policy, vol.3, pp. 108-126, (1969).

WIL70. **Wilson A.G.**: *Entropy in Urban and Regional Modelling*, Pion, London. (1970).