Logical indetermination coupling:a method to minimize drawing matches and its applications

Pierre Bertrand, Michel Broniatowski, Jean-François Marcotorchino

Logical *indetermination* coupling:a method to minimize drawing matches and its applications

Pierre Bertrand*[,1], Michel Broniatowski[1,2], and Jean-François Marcotorchino[3]

[1]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France
[2]CNRS UMR 8001, Sorbonne Université, Paris, France
[3]Institut de statistique, Sorbonne Université, Paris, France

December 22, 2020

Abstract

While justifying that *independence* is a canonic coupling, the authors show the existence of a second equilibrium to reduce the information conveyed from the margins to the joined distribution: the so-called *indetermination*. They use this null information property to apply *indetermination* to graph clustering. Furthermore, they break down a drawing under *indetermination* to emphasis it is the best construction to reduce couple matchings, meaning, the expected number of equal couples drawn in a row. Using this property, they notice that *indetermination* appears in two problems (Guessing and Task Partitioning) where couple matchings reduction is a key objective.

Keywords: *Correlation Clustering, Mathematical Relational Analysis, Optimal Transport, Logical indetermination, Coupling Functions, Graph Theory, Task Partitioning Problem, Guessing Problem*

1 Introduction

In a precedent paper [3], we highlighted a list of structural analogies between two discrete couplings namely *independence* and *indetermination* together with an application to graph clustering that we will recall hereafter in section 4.

A discrete coupling is a function C operating on two discrete marginal laws $\mu = \mu_1 \dots \mu_p$ and $\nu = \nu_1 \dots \nu_q$ and which defines a probability law π on the product space:

$$\pi_{u,v} = C(\mu_u, \nu_v), \ \forall\ 1 \leq u \leq p, \ 1 \leq v \leq q$$

We respectively quote both those above mentioned couplings $C^{\times}$ (*independence*) and C^{+} (*indetermination*); this last notion has been initially introduced by J.-F. Marcotorchino in his seminal papers [21] and [24]) while their formula will be reintroduced and rediscussed later on in section (2).

Their usefulness arises in statistical applications: namely, most of our usual statistical deviation criteria for contingency analysis are expressed in terms of deviations from one of the two couplings ([8] gathers a classification of them, deviation to *independence* or deviation to

*Corresponding author: pierre.bertrand@ens-cachan.fr

indetermination). The most famous example for *independence* is the χ^2, widely used in practice which computes nothing but a deviation to the *independence* coupling of the induced margins. Symmetrically the Janson-Vegelius coefficient, initially introduced in [16] as a contingency association index, measures also a deviation no longer to *independence* but rather to *indetermination*; we shall detail this point in subsection 4.1.

Furthermore, purely theoretical considerations lead to consider *independence* and *indetermination* as the only two possible "natural" equilibria: this fact being based upon a work of Csizar [9], a summarized version of which is expressed in [3]; we will briefly explain, in this present document, the rationale behind.

While *independence* is commonly used and studied in the scientific literature, *indetermination* appears as a lesser known coupling, whose properties have been rarely presented in an explicit way. After introducing both coupling functions in section (2), the next section, section 3 of this paper, is precisely dedicated to the properties implied by *indetermination*. Notably, we show that it consists in reducing couple matching occurrences: drawing two couples under *indetermination*, the probability to have both couples equal will be particularly small. In addition, we decompose an *indetermination* coupling into an arbitrary margin and a uniform one leading to a constructive drawing.

Section 4 provides several information problems where the logical *indetermination* coupling appears. First, since it is an equilibrium similar to independance, we coin a new local graph criteria as the Girvan-Newman modularity ([30]) but based on the *indetermination* coupling. Besides, leveraging on the constructive drawing as well as on the reducing "couple matchings" property, we show that *indetermination* naturally occurs in solving the so called "guessing problem" ([27]) as well as the "task partitioning problem" ([4]). The guessing problem appears in cryptography, when a spy screening a communication session tries to determine which message was sent, using to do so, some partial information. The task partitioning problem occurs in manufacturing process when we want to optimize the way to assign tasks to production teams, or to machines in job-shop scheduling.

Eventually, having deeply studied the discrete version of *indetermination*, section 5 extends the *indetermination* notion towards a continuous version, providing straightforwardly associated density and cumulative distribution functions.

2 From discrete transport problem to coupling function

When we want to couple two marginal laws, the most common and straightforward way to proceed, consists in assuming *independence* and keep on computations. It is so well integrated in our mindset, that it naturally appears in real life applications, as soon as we want to build fast models up. In our scientific work, the approach is quite the same: when we use a very classical and usual criterion like the χ^2 index, we are measuring nothing but a deviation to *independence*.

Thinking about how we first introduced *independence*, we immediately suggest empirical experiments: let us say if we roll a dice twice, how should we derive the resulting probabilities from a unique dice? Most of us will naturally apply *independence* coupling: it really relies on empirical experiments.

Although being the most natural, it is not, by far, the only existing available coupling method; actually, as introduced by Sklar in [34], any copula function will lead to a coupling function behaving on two cumulative distribution functions. In [3] we defined a notion called "coupling function" resulting from the resolution of a fixed transport problem. Throughout the first part of this precedent paper we considered discrete optimal transport problem: how can we efficiently move a pile of sand into a corresponding hole to fill? Using Kantorovitch's theorem formulation, how can we respect given margins (in terms of masses or quantities deposited in

the "origins – destinations" schema) while minimizing a cost function splitting optimally those quantities on the resulting coupling.

Formally, through modern notations, two probability measures $\mu = \mu_1 \dots \mu_p$ and $\nu = \nu_1 \dots \nu_q$ represent the initial constraints (respectively masses at origins and destinations). Typically, for instance in national economy planning: μ expresses the quantity produced by a sector and ν the consumed quantity by an industry. Then, given a cost function M which associates a positive value to any element in $[0, 1]$, the discrete "Monge Kantorovitch Problem" (MKP) can be expressed as follows:

Problem 1 (Discrete Version of MKP)**.**

$$\begin{aligned} \min_{\pi} \quad & \sum_{u=1}^{p}\sum_{v=1}^{q} \mathbf{M}(\pi_{u,v}) \\ subject\ to: \quad & \\ & \sum_{v=1}^{q} \pi_{u,v} = \mu_u; \forall u \in \{1, ..., p\} \\ & \sum_{u=1}^{p} \pi_{u,v} = \nu_v; \forall v \in \{1, ..., q\} \\ & \pi_{u,v} \geq 0; \forall (u, v) \in \{1, ..., p\} \times \{1, ..., q\} \end{aligned}$$

The choice of a cost function $\mathbf{M}$ depends upon the applications we want to address. For instance, we can force the result π to concentrate as little information as possible; this means, we shall constraint it to be as close as possible to the uniform law, referring to the product space given the margins. Other choice: we can, as well, minimize the entropy of π given the margins. Both those cases are usual approaches. They expect the global assignment to be as smooth as possible.

A MKP problem is essentially given by its cost function M, while margins (μ, ν) may vary. As such the solution to Problem 1 is written as $C^P(\mu, \nu)$

Definition 1 (MKP Problem Associated with Coupling function)**.**
For a given MKP problem P, we define a coupling function C^P by: $C^P(\mu, \nu) = \pi^*(P)$ provided that π^* exists as a unique solution of P with margins μ and ν.

The existence of a solution depends on the cost function $\mathbf{M}$ as well as on the margins. It has been discussed in numerous papers and we will not further discuss this point here; neither did we in [3].

In paper [3] nevertheless, we expressed Problem 1 using the projections introduced in [9]. Doing so, we were able to justify the two canonic discrete costs defined below (Equations (1) and (2)) as the only two ways leading to a structured coupling function.

$$\mathbf{M}^{\times} : x \mapsto x \log(pqx) \tag{1}$$

$$\mathbf{M}^{+} : x \mapsto pq\left(x - \frac{1}{pq}\right)^2 \tag{2}$$

Under the projection formalism, let us unify the costs provided beforehand. As it will be mentioned later again, when we associate each problem to its coupling function, both share the same solution if we remove the constraints on margins, that is to say: the uniform law ($\pi_{u,v} = \frac{1}{pq}$). Actually the cost function can be formally expressed as a *divergence* (see the definition 2 below) to the uniform law (consequently to the optimal solution without constraints).

Definition 2 (Divergence with parameter)**.**
Given, a positive function $\phi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ with $\phi(1) = 0$, and given two discrete probabilities m and n on pq elements, we define the divergence D_ϕ as:

$$D_\phi(m|n) = \sum_{u=1}^{p}\sum_{v=1}^{q} n_{u,v}\phi\left(\frac{m_{u,v}}{n_{u,v}}\right) = \mathbb{E}_n\left[\phi\left(\frac{m_{u,v}}{n_{u,v}}\right)\right]$$

additionally we set $0 * \phi(x) = 0$ for all x.

Denote by $\mathcal{U}$ the uniform distribution on $(1, .., p) \times (1, .., q)$, ie $\mathcal{U}(u, v) := 1/pq$ for all u and v. Now, for $\phi : x \mapsto x\log(x) - x + 1$ it holds:

$$\begin{aligned} D_{KL}(\pi\,|\mathcal{U}) &= \sum_{u=1}^{p}\sum_{v=1}^{q}\frac{1}{pq}\left[pq\pi_{u,v}\cdot\log(pq\pi_{u,v}) - pq\pi_{u,v} + 1\right] \\ &= \sum_{u=1}^{p}\sum_{v=1}^{q}\pi_{u,v}\log(pq\pi_{u,v}) \\ &= \sum_{u=1}^{p}\sum_{v=1}^{q}\mathbf{M}^{\times}(\pi_{u,v}) \end{aligned}$$

so that Problem 1 associated with the cost function $\mathbf{M}^{\times}$ is rewritten as a minimization of D_{KL} which is the usual Kullback-Leibler divergence.

Similarly, if we pose $\phi : x \mapsto (x-1)^2$ (the so-called Chi-square divergence function)

$$\begin{aligned} D_2(\pi\,|\mathcal{U}) &= \sum_{u=1}^{p}\sum_{v=1}^{q}\frac{1}{pq}(1 - pq\pi_{u,v})^2 \\ &= pq\sum_{u=1}^{p}\sum_{v=1}^{q}\left(\pi_{u,v} - \frac{1}{pq}\right)^2 \\ &= \sum_{u=1}^{p}\sum_{v=1}^{q}\mathbf{M}^{+}(\pi_{u,v}) \\ &= -1 + pq\sum_{u=1}^{p}\sum_{v=1}^{q}\pi_{u,v}^2 \end{aligned} \tag{3}$$

Finally, the two costs $\mathbf{M}^{\times}$ and $\mathbf{M}^{+}$ differ from each other in the way the divergence from $\mathcal{U}$ to the uniform law π is taken into account and amounts to optimizing a divergence.

Although it seems arbitrary, the restriction to the two previous costs, is anything but a fortuitous decision: in [9], Csiszar actually shows that, provided we verify additional intuitive properties, we must restrict ourselves to use either least square or maximum entropy in Problem 1. Our transport problem aims at projecting the uniform $\mathcal{U}$ into the eligible space of probability law respecting two margins.

A general question is how to adapt a "prior guess" u_0 to verify a list of constraints. Let us say u_0 lives in the simplex S_n while the given constraints define a subspace $L \in \mathcal{L}$ ($\mathcal{L}$ is the space of subspaces of S_n tuned by a finite list of affine constraints, see [9] for more details). To formalize it, Csiszar defines a *projection rule* Π as a function whose input is a set $L \in \mathcal{L}$ and which generates a method Π_L to project any prior guess u_0 to a vector in L:

$$\begin{aligned} \Pi : \quad & \mathcal{L} \to (S_n \to S_n) \\ & L \to \Pi_L : (u_0 \to \Pi_L(u_0)) \in L) \end{aligned}$$

The article then introduces a collection of "natural" properties that we gather hereafter.

- *consistency*: if $L' \subset L$ and $\Pi_L(S_n) \subset L'$ then $\Pi_{L'} = \Pi_L$; basically, if the result of a projection to a bigger space is always inside a smaller, then the projection on the two spaces are equivalent.
- *distinctness*: if L and L' are defined by a unique constraint and they are not equal, then $\Pi_L \neq \Pi_{L'}$ (unless they both contains the initial prior guess). Typically, in $\mathbb{R}^2$, minimizing $||\cdot||$ on two lines returns a different result as soon as they do not both contain 0.
- *continuity*: Π is continuous with regards to $L \in \mathcal{L}$; it has a continuous relation with constraints.
- *scale invariant*: $\Pi_{\lambda L}(\lambda u) = \lambda u$ for any positive λ and any $u \in S_n$.
- *local*: for any subset $J \subset \{1, \ldots, n\}$, $(\Pi_L)_J = (\Pi_{L'})_J$ as soon as $L_J = L'_J$ where L_J means we only keep constraints dealing with coordinates in J and $(\Pi_L)_J$ is the restriction of the resulting vector of Π_L to the J coordinates. This property indicates that the results of Π on a set of coordinates, only depends on constraints applied to those coordinates.
- *transitive*: for any $L' \subset L$, $\Pi_{L'} = \Pi'_L \circ \Pi_L$. We can first project on a bigger space without affecting the result.

Adding a last property which guarantees that the "no interaction" solution in case we omit constraints respects a proportional behavior finally lead to restrict ourselves to the two problems we treat in this document.

2.1 The Alan Wilson's entropy model: role of "independence"

First introduced by Sir Alan Wilson in 1969 for "Spatial Interaction Modeling" the "Flows Entropy Model" of Alan Wilson, can be found in his various publications; originated in [38], developed in [39], and refined in his book [40]. A fundamental justification of his approach corresponds to the following contextual situation: in a theoretical system it is advisable to determine the distribution $\pi_{u,v}$ (normalized frequency flows), supposing $\pi \geq 0$, which maximizes the entropy of the system under some given constraints.

Without any marginal constraints, the optimal solution is nothing but the uniform law $\pi_{u,v} = \frac{1}{pq}, \forall\ 1 \leq u \leq p, \forall\ 1 \leq v \leq q$. By using margins, that is to say information about total exports (origins flows) and total imports (destination flows), degree of disorder of the system is drastically reduced. We are led to the following Problem 2; solution of which is given by Theorem 1.

Problem 2 (Balanced PSIS)**.**

$$
\begin{aligned}
\min_{\pi} \quad & \sum_{u=1}^{p}\sum_{v=1}^{q} \pi_{u,v} \log(pq\pi_{u,v}) = D_{KL}\left(\pi \,|\mathcal{U}\right) \\
\text{subject to:} \quad & \\
& \sum_{v=1}^{q} \pi_{u,v} = \mu_u, \ \forall 1 \leq u \leq p \\
& \sum_{u=1}^{p} \pi_{u,v} = \nu_v, \ \forall 1 \leq v \leq q \\
& 0 \leq \pi_{u,v} \leq 1, \ \forall 1 \leq u \leq p, \ 1 \leq v \leq q
\end{aligned}
$$

Theorem 1.

The solution of Problem 2 is $\pi^{\times}(u,v) = \mu_u \nu_v,\ \forall 1 \leq u \leq p,\ 1 \leq v \leq q.$
Hence the coupling function associated to Problem 2 is nothing but independence:

$$C^{Problem\ 2}(\mu,\nu)_{u,v} = C^{\times}(\mu,\nu)_{u,v} = (\mu \otimes \nu)_{u,v} = \mu_u \nu_v$$

As a conclusion, from the direct maximization of entropy, we get the solution expressed in terms of probability and remark that the associated coupling function is nothing but *independence* (denoted by $\otimes$ in the sequel). We also note that the degree of disorder is not total: flows possess an intensity which is proportional to the weights of the partners in the world trade exchanges matrix in case of an economic application.

2.2 The minimal trade model: role of *indetermination*

In the "Minimal Trade Model" (see [36], [21] and [24]), we still impose the objective function to respect the marginal distributions and positivity constraints; but we change the structure of the cost function. In that case the criterion is a quadratic function measuring squared deviation of the cells values from the "no information" situation (the uniform joint distribution law). As expected, in case of free margins, the solution remains the uniform law $\mathcal{U}$. But, adding usual pre-conditioned constraints on margins leads to the least squares problem is Problem 3; solution of which is given by Theorem 2.

Problem 3 (Minimal Trade Model)**.**

$$\min_{\pi} \quad pq \sum_{u,v} \left(\pi_{u,v} - \frac{1}{pq} \right)^2 = D_2\left(\pi \,|\mathcal{U}\right)$$

subject to:

$$\sum_{v=1}^{q} \pi_{u,v} = \mu_u,\ \forall 1 \leq u \leq p$$

$$\sum_{u=1}^{p} \pi_{u,v} = \nu_v,\ \forall 1 \leq v \leq q$$

$$0 \leq \pi_{u,v} \leq 1,\ \forall 1 \leq u,v \leq q$$

Theorem 2.

The solution of Problem 3 is $\pi^{+}(u,v) = \frac{\mu_u}{q} + \frac{\nu_v}{p} - \frac{1}{pq}.$
Hence the coupling function associated to Problem 3 is nothing but indetermination:

$$C^{Problem\ 3}(\mu,\nu)_{u,v} = C^{+}(\mu,\nu)_{u,v} = (\mu \oplus \nu)_{u,v} = \frac{\mu_u}{q} + \frac{\nu_v}{p} - \frac{1}{pq}$$

A supplementary condition, which is exogenous with regard to the previous model, must be added on the margins (which are, by the way, constant values given *a priori*), this condition (see [21]) is a the following inequality which guarantees the positivity of the frequency matrix $\pi^{+}(u,v) = C^{+}(\mu,\nu)_{u,v}$:

$$p \min_{u} \mu_u + q \min_{v} \nu_v \geq 1 \tag{4}$$

From now on, we shall consider that Condition (4) applies whatever the values of the μ_u and ν_v are.

At that stage we introduced the two coupling functions in the discrete version used in graph clustering criteria. We have proposed in [3] some few properties of the *indetermination* notion; we now consider a constructive approach to such coupling. It appears that this construction aims at reducing some matching problem. We will present two specific applications in Section 4.

3 Discrete *indetermination* and associated properties

3.1 Monge property

The class of matrices we define here is attributed to Gaspard Monge, from a basic idea appearing in his 1781 paper. In fact, to introduce Monge's properties, we follow the exhaustive work of Rainer Burkard, Bettina Klinz and Rüdiger Rudolf exposed in their 66-pages-long article [5]. We begin with Definition 3.

Definition 3 (Monge and Anti-Monge matrix)**.**
A $p \times q$ real matrix $c_{u,v}$ is said to be a Monge matrix if it satisfies:

$$c_{u,v} + c_{u',v'} \leq c_{u',v} + c_{u,v'} \ \forall \ 1 \leq u \leq u' \leq p, \ 1 \leq v \leq v' \leq q$$

and an Anti-Monge matrix if:

$$c_{u,v} + c_{u',v'} \geq c_{u',v} + c_{u,v'} \ \forall \ 1 \leq u \leq u' \leq p, \ 1 \leq v \leq v' \leq q$$

Remark 1 (Full-Monge matrix)**.**
The important case for our purpose is the case of equality when a matrix is both Monge and Anti-Monge, we will call this situation "Full-Monge" matrix.

$$c_{u,v} + c_{u',v'} = c_{u',v} + c_{u,v'} \ \forall \ 1 \leq u \leq u' \leq p, \ 1 \leq v \leq v' \leq q$$

Although it was seldomly studied (see namely [24]), this last equality fits perfectly our purpose. The inequalities on the contrary, are common and can be met in various situations such as cumulative distribution functions, or copula theory.

Remark 2 (Adjacent cells)**.**
A straightforward but important derived property is the local adjacency cells equality: it is sufficient to satisfy the property of the Remark 1 on adjacent cells, to ensure the obtainment of a "Full-Monge" matrix behavior for the global set of cells i.e.:

$$c_{u,v} + c_{u+1,v+1} = c_{u+1,v} + c_{u,v+1} \ \forall \ 1 \leq u \leq p, \ 1 \leq v \leq q$$

Remark 2 is a key property to study Monge matrices since it gives a direct $\mathcal{O}(pq)$ algorithm to verify if a matrix is full-Monge.

A fact remains that a question arises: which density functions verify the Full Monge property? The following Proposition 1 gives an interesting answer: all full Monge matrices derive from the density of an *indetermination* structure.

Property 1 (Full-Monge matrix is equivalent to *indetermination*)**.**
A "full Monge matrix" necessarily represents an "indetermination coupling".

Proof.
Summing on u' and v' the equality of Remark 1 we straightforwardly obtain:

$$\sum_{u'} \sum_{v'} \left(c_{u,v} + c_{u',v'} - c_{u',v} - c_{u,v'} \right) = pqc_{u,v} + c_{.,.} - qc_{.,v} - pc_{u,.} = 0 \rightarrow c_{u,v} = \frac{c_{u,.}}{q} + \frac{c_{.,v}}{p} - \frac{c_{.,.}}{pq}$$

□

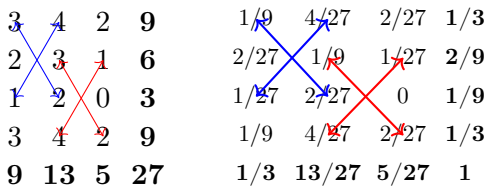

Figure 1: Example of an *indetermination* coupling (Statistical counting vs Probability forms)

By summarizing properties of Full-Monge Matrices we get the following Theorem 3.

Theorem 3 (Full-Monge matrices)**.**
Let π be a probability distribution on $(1,\dots,p)\times(1,\dots,q)$ then the following properties are equivalent.

1. *π is a Full-Monge matrix*
2. *$\pi_{u,v} = \pi^{+}_{u,v} = \frac{\mu_u}{q} + \frac{\nu_v}{p} - \frac{1}{pq}$*
3. *π optimizes Problem 3 for some given margins*
4. *All 2×2 sub-tables $\{u, v, u', v'\}$ extracted from π have the same sum on their diagonal and anti-diagonal*

Figure 1 features the last assertion in Theorem 3 and justifies the $\oplus$ notation assigned to *indetermination*. Indeed, if we take blue and red arrows we get the same resulting value. Using the contingency form:

$$\begin{aligned} \text{blue arrows} &: 3+2-1-4=0 \\ \text{red arrows} &: 3+2-4-1=0 \end{aligned}$$

Equality remains true for the probability form since we just have to divide the cell values by the total sum of the matrix (27 here).

3.2 A dependence to avoid matches

In the discrete case, Problem 3 is written as a projection of a prior guess $\mathcal{U} = \frac{1}{pq}$ (as already mentioned: the optimal solution without any initial information). As previously noticed $\mathbb{L}^2$ (Equation (1)) is the distance we project $\mathcal{U}$ with, on the space of probability measures respecting the two margins μ and ν. Theorem 2 provides the form of the optimal solution $\pi^{+} = C^{+}(\mu,\nu)$.

We notice that π^{+} introduces a dependence between its margins, different from $C^{\times}$, the *independence* coupling.

Moreover, we notice that, unlike $\pi^{+} = C^{+}(\mu,\nu)$, the Kullback-Leibler (or entropic) approach leads to *independence*. Indeed, Problem 2 can be formulated as a projection of $\mathcal{U}$ on the same space of probability measures respecting the two margins μ and ν but using Kullback-Leibler. The solution of which is given by Theorem 1: $\pi^{\times} = C^{\times}(\mu,\nu)$), generating no dependence between the margins, by definition.

To summarize, if $\mathcal{L}(\mu,\nu)$ is the space of probability measures respecting the two margins μ and ν we have:

$$\mathcal{U} \overset{KL}{\Longmapsto} \mathcal{L}(\mu,\nu) = \pi^{\times} \;\Big|\; \mathcal{U} \overset{\mathbb{L}^2}{\Longmapsto} \mathcal{L}(\mu,\nu) = \pi^{+}$$

Let us focus now on the *indetermination* coupling using an interpretation of the problem it solves. Indeed, removing constants, the cost function

$$\mathbf{M}^+ = pq \sum_{u,v} \left\{ \pi_{u,v} - \frac{1}{pq} \right\}^2$$

can be simply written:

$$\mathbf{M}^+ = \sum_{u=1}^{p} \sum_{v=1}^{q} \pi_{u,v}^2.$$

As we are minimizing $\mathbf{M}^+$, the formulation of the problem precisely aims at reducing couple matchings. Indeed, if we independently draw two occurrences of $W = (U, V) \sim \pi$, the probability of getting a matching for a couple, that is to say $U_1 = U_2$ and $V_1 = V_2$ simultaneously, is nothing but $\pi_{U_1,V_1} \times \pi_{U_2,V_2} = \pi^2_{U_1,V_1}$.

3.3 Constructive definition of an *indetermination* matrix

Let us unfold hereafter the dependence introduced by the *indetermination*, associated to the $\mathbb{L}^2$ projection. We use the Full Monge property that characterizes *indetermination* (see Theorem 3 and which can be rewritten as:

$$\pi^+_{u,v} - \pi^+_{u',v} = \pi^+_{u,v'} - \pi^+_{u',v'}$$

This last equality shows that the difference between two lines is the same whatever the column. We deduce a constructive definition of an *indetermination* law on $(1, \ldots, p) \times (1, \ldots, q)$, for this sake:

1. For the first line we fix an arbitrary distribution on the q columns: $(\Pi^+_{1,v})_{1 \leq v \leq q}$.
2. For any other line u, we define $\Pi^+_{u,v} = \Pi^+_{1,v} + \frac{\Delta_u}{q}$, $\forall 1 \leq \ldots \leq q$ where Δ_u is any real number such that $\Pi^+_{u,v}$ is always positive
3. To eventually obtain a probability matrix (summing up to 1), we set $T = \sum_{u=1}^{p} \sum_{v=1}^{q} \Pi^+_{u,v}$ and define: $\pi^+_{u,v} = \frac{\Pi^+_{u,v}}{T}$ for any (u, v) together with $\delta_u = \frac{\Delta_u}{T}$ for any u.

There are as many *indetermination* matrices on pq elements as choices of a first line $(\pi^+_{1,v})_{1 \leq v \leq q}$ and of increments $(\delta_1 = 0, \delta_2, \ldots, \delta_p)$.

Remark 3 (From lines to columns)**.**
In the precedent construction, lines and columns played a different role. Obviously, any property is symmetric so that it can be easily transferred from u to v .

Remark 4 (δ_u expresses the differences between margins)**.**
From the precedent construction, the deduced margins are:

$$\pi^+_{u,\cdot} = \pi^+_{1,\cdot} + \delta_u, \forall 1 \leq u \leq p$$

where for any u, $\pi^+_{u,\cdot} = \sum_{v=1}^{q} \pi^+_{u,v}$.

The division by q we introduced in the definition allows us to interpret δ_u *as the difference between any margin to the first one.*

We suppose the δ_u are positive; following Remark 4 it amounts to saying that the first line corresponds to the minimal margin.

The more a δ_u is chosen high, the more probable the line u will be and the more the drawings of v on line u will be close to uniform on $1, \ldots, q$. This behavior is compliant with the couple matching limiting property exposed in section 3.2. It decreases couple matchings since when a first frequent u_1 is drawn, a conflict is probable with the second drawn u_2. To avoid a couple conflict the *indetermination* π^+ draws v_1 and v_2 as uniformly as possible.

A uniform margin on u returns to say $\delta = 0$. In that case, any line equals the first one and we have for any (u, v):

$$\begin{aligned} \pi^+_{u,v} &= \pi^+_{1,v} \\ &= \frac{\pi_{\cdot,v}}{q} \\ &= \pi_{u,\cdot}\pi_{\cdot,v} \end{aligned}$$

meaning that we constructed *independence*. One can actually easily check that if μ is uniform and whatever ν is, $C^+(\mu, \nu) = C^\times(\mu, \nu)$, which explains the result.

On the contrary, the more both margins μ and ν differ from an uniform law and the more $\pi^+ = C^+(\mu, \nu)$ and $\pi^\times = C^\times(\mu, \nu)$ will differ; this was already spotted in [3]. Using our new formalism, if μ is far from the uniform, the associated array δ will obviously take high values. On the corresponding lines with a high weight δ_u the influence of $\pi_{1,v}$ will disappear. As a consequence, the second variable v will be close to a uniform drawn on $1 \ldots q$.

3.4 Drawing under *indetermination*

Given a probability law $\pi^+ = C^+(\mu, \nu)$ for two margins μ and ν we use the above constructive process of *indetermination* to propose a two steps method for drawing under π^+. Besides, we enforce $\mu_1, \leq \mu_2 \leq \ldots \leq \mu_p$. It amounts to order modalities according to their increasing probability and ends up being a renaming.

First, an unbalanced distribution $\pi^+_{1,v}$ adding up to μ_1 is extracted from π by reading its first line (we already supposed μ_1 is the smallest margin thanks to our order hypothesis). Then, we extract a list of corresponding computed $\delta_u = \mu_u - \mu_1$ to ensure that the differences between margins are respected. Eventually, a drawing is computed as follows:

1. We draw u through μ
2. We roll a loaded dice with the Bernoulli's skew: $I = \text{Be}\left(\frac{\delta_u}{\mu_u}\right)$
3. If we get 0, we draw v under the distribution $\left(\pi^+_{1,1}, \ldots, \pi^+_{1,q}\right)$ else, having 1, we draw v uniformly hence under $\left(\frac{1}{q}, \ldots, \frac{1}{q}\right)$.

In Proposition 2 we formally demonstrate that our new drawing realizes the *indetermination* law π^+.

Property 2 (Drawing under *indetermination*)**.**
The method we just built up does realize indetermination.

Proof.

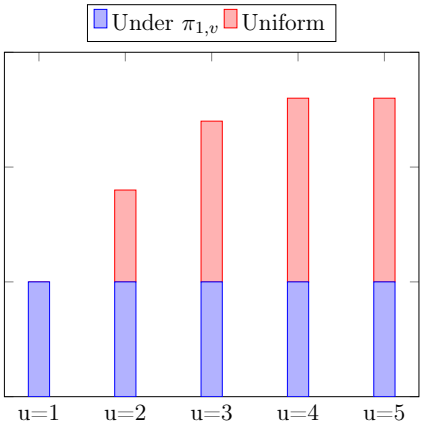

Figure 2: Distribution of v according to the margin u

We show that the resulting probabilities follow *indetermination*.

$$\begin{aligned}
\mathbb{P}(U=u,V=v) &= \mathbb{P}(V=v|U=u)\mathbb{P}(U=u) \\
&= \mathbb{P}(V=v|U=u,I=0)\mathbb{P}(U=u)\mathbb{P}(I=0) \\
&\quad +\mathbb{P}(V=v|U=u,I=1)\mathbb{P}(U=u)\mathbb{P}(I=1) \\
&= \pi^{+}_{1,v}\mu_u\frac{\mu_1}{\mu_u}+\frac{1}{q}\mu_u\frac{\mu_u-\mu_1}{\mu_u} \\
&= \pi^{+}_{1,v}\mu_1+\frac{\mu_u-\mu_1}{q}
\end{aligned}$$

Now, we notice that the sum on u of those probabilities equals $\mathbb{P}(V=v)$ so that:

$$\begin{aligned}
\mathbb{P}(V=v) &= \nu_v = p\pi^{+}_{1,v}\mu_1+\frac{1-p\mu_1}{q} \\
\text{or:} &\qquad \pi^{+}_{1,v}\mu_1-\frac{\mu_1}{q}=\frac{\nu_v}{p}-\frac{1}{pq}
\end{aligned}$$

Replacing, we obtain the expected formula:

$$\mathbb{P}(U=u,V=v)=\frac{\mu_u}{q}+\frac{\nu_v}{p}-\frac{1}{pq}$$

□

The constructive method we just described creates a three steps process to realize *indetermination*; it leads to the Histogram 2 presented just above.

This Histogram 2 focuses on the distribution evolution. Hence, we observe that the higher μ_u is, the more uniform $V|U=u$ is. Indeed, $(\pi_{1,1},\dots,\pi_{1,q})$ brings up a relative disequilibrium (it is arbitrary) but the complementary part is uniformly drawn. The proportion featured in Histogram 2 is, obviously, dummy but it correctly features the evolution of the conditional probability depending on the margin. Furthermore, the sorting on μ_u, coherent with the uniform proportion, insists on the property that the important masses are at close-to-uniform.

Actually, this constructive process always behaves as a strategy to avoid "couple matchings". When having drawn a first margin u_1 with a small weight, u_2 will probably differ. Remember margins are fixed, the *indetermination* structure hides in those "small lines" any disequilibrium

on v: $\left(\pi^+_{1,1}, \dots, \pi^+_{1,q}\right)$ that the margins introduced. On the contrary, having drawn a first margin u_1 with a high weight, u_2 has a high probability to be equal to u_1; there the close-to-uniform distribution of $V|u = u_1$ protects the couple from a matching ($v_1 = v_2$ is rare).

In the next section, dedicated to applications, we will leverage this property that *indetermination* ends up being a strategy to avoid couple matching while respecting margins.

4 Applications of *indetermination* structure

4.1 Deviation from *indetermination*: Janson Vegelius coefficient

In statistical analysis, given the values of two descriptive variables on a number n of individuals, an usual and important problem is to use a coefficient or index, measuring the correlation between the two variables.

Formally, U represents a first variable which characterizes individuals among p modalities (for instance the city where they are living, their socio-professional category, their ages, ...); a second variable V classifies them among q categories (or split them into q categories or classes).

Given realizations $(U_1, \dots, U_p)$ and $(V_1, \dots, V_q)$, the categorization of n individuals, how do we measure the correlation between U and V? Correlation typically means that the value of V is dependent on the value of U. Expressing it with "dependence" notion, we naturally define a deviation-to-independence coefficient (*i.e.* a departure from *independence* index), for instance: the χ_2.

To do so, from the n realizations, of U we deduce an empirical margin μ counting the proportion of individuals in each modality:

$$\mu_u = \frac{\#\{i \ / \ U_i = u\}}{n}$$

similarly, an empiric margin ν out of the realizations of V and eventually an empiric margin π out of (U, V).

The usual χ^2 can be defined as a square deviation to *independence*:

$$\chi^2(U,V) = \sum_{u=1}^{p}\sum_{v=1}^{q} \frac{(\pi_{u,v} - (\mu \otimes \nu)_{u,v})^2}{(\mu \otimes \nu)_{u,v}} \tag{5}$$

which obviously happens to be null if and only if the empirical distribution π of the observed data is an indepedence coupling of the empiric margins.

Using a symmetric idea, a less used criterion, called "Janson-Vegelius Index", after the name of the inventors of this coefficient, who coined it in [16], [17] or [18] writes as a deviation to *indetermination*:

$$JV(U,V) = \sum_{u=1}^{p}\sum_{v=1}^{q} \frac{(\pi_{u,v} - (\mu \oplus \nu)_{u,v})^2}{\sqrt{\frac{p-2}{p}\left(\sum_{u=1}^{q}\mu_u^2 + 1\right)}\sqrt{\frac{q-2}{q}\left(\sum_{v=1}^{q}\nu_v^2 + 1\right)}} \tag{6}$$

and obviously is equal to zero if and only if the empirical π is an *indetermination* coupling of the empirical margins.

JV index, although its formulation, using contingency notations appears as non trivial, is actually just a classical cosine, or a Pearson's like correlation coefficient when rewritten in the "Mathematical Relational Analysis" Space. A list of papers which gathers some of the most important key features about the subject is [26], [21], [28], [31], [22], [23], [1].

The relational analysis space no longer encodes modalities but links between individuals. Two matrices X and Y of size $n \times n$ respectively associated to variables U and V are introduced as shown in Definition 4.

Definition 4 (Mathematical Relational Analysis notations)**.**
Let $(U_1, \ldots, U_n)$ *and* $(V_1, \ldots, V_n)$ *be two* n *probabilistic draws of* $U \sim \mu$ *and* $V \sim \nu$ *respectively. We define two associated symmetric* $n \times n$ *matrices* X *and* Y *by*

$$X_{i,j} = 1_{u_i = u_j}, \ \forall 1 \leq i, j \leq n$$
$$Y_{i,j} = 1_{v_i = v_j}, \ \forall 1 \leq i, j \leq n$$

Or in literal form:

- $X_{i,j} = 1$*, if* i *and* j *share the same modality of variable* U*,* $X_{i,j} = 0$ *if not*
- $Y_{i,j} = 1$*, if* i *and* j *share the same modality of variable* V*,* $Y_{i,j} = 0$ *if not*

To understand the notation, let us begin with some remarks about Definition 4. Basically, the two $\{0, 1\}$ matrices X and Y (which correspond in fact to two binary equivalence relations based on the drawn modalities) represent agreements and disagreements between the two variables on a same draw of size n; they are symmetric with 1 values on their diagonal.

As expected, one can pass from the relational encoding to the usual contingency encoding as well as in the reciprocal way; those transfer formulas are demonstrated in the mentioned articles. Coming back to the JV index , those formulas enable us to write JV as a cosine in the relational space:

$$JV(U,V) = JV(X,Y) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_{i,j} - \frac{1}{p}\right)\left(Y_{i,j} - \frac{1}{q}\right)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_{i,j} - \frac{1}{p}\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_{i,j} - \frac{1}{q}\right)^2}} \tag{7}$$

Calculations leading to Equation (7) from Equation (6) can be found in [26] or [25].

Instead to study in depth the differences between the two coefficients, $\chi^2(U,V)$ and $JV(U,V)$ we recall how they apply in clustering and refer to [3] for a more precise study.

4.2 Clustering problem dedicated to graphs

Clustering algorithms aim at getting a decomposition of graphs into classes, maximizing internal similarities as well as minimizing external ones. It is motivated by the fact that in any application, the connections are not random. Precisely, a triangle has a high chance of being a clique, as expressed in [14]: "two of your friends will have a greater probability of knowing one another than will two people chosen at random from the population, on account of their common acquaintance with you".

Various methods were proposed during the 20th. Due to the huge amount of papers available, we provide here an arbitrary selection while trying to highlight most of the subjects the scientific community is facing. A first option is to take as a fixed input the number K of classes we are looking for (although it is quite unrealistic for huge graphs), together with an associated distance (or dissimilarity index) and come up with a list of best representatives or "means" for each class. The output "means" tend to optimize the sum of distances from all vertices to their nearest mean. K-means algorithm whose idea goes back to the fifties (originated in [35] and rewritten by E.W. Forgy in [11]) typically illustrates this option. A second option, is to construct a local criterion w which assigns a weight $w_{i,j}$ to each (i, j) couple of vertices based on their similarity; the more similar they are, the higher the criterion is. For instance, [19] affects the number of paths from i to j in the graph with an exponentially-decreasing factor according to their length. One then build up a global criterion W by summing up the local values $w_{i,j}$ if and only if i and j are in the same class as proposed in Problem 4.

When [14] is published, the community detection problem, though it is applied in various domains and contexts, still lacks a method to impartially evaluate and compare two results. Two years passed before the authors Girvan and Newman published a method to measure the global quality of a network clustering through the so-called modularity ([30]): $M^{\times}$ (Definition 7). The idea is to compare the number of edges in a community with the expected number in case of edges distributed without regard to the community they belong to. Having expressed a global measure for the quality of a clustering enabled the community to objectively compare their performances.

In fact, they were faced with several problems. First, finding a clustering which maximizes $M^{\times}$ is a NP-hard problem and therefore forces the specialists to use heuristics like the original one in [7]. We can find out a list of such heuristics in [33], for instance among others: the use of spectral methods [10] as well as data structure tricks [37]. Some of those methods are gathered in the book written by Fortunato [12]. Furthermore $M^{\times}$ is as plain as arbitrary since it introduces a resolution limit concern, first mentioned in [13] and which led to a research gate for improving the modularity (see in particular [8]). Either the authors decided to trick the modularity index such as in [32] or with the help of the recently introduced modularity density ([6]) or they completely changed the criteria using typically a so-called clustering coefficient [29]. Eventually, the definition itself of clustering is subject to discussion, shall communities be allowed to overlap? some articles allow that and define fuzzy communities (see [15]).

Basically, out of the literature, three subjects stand out: algorithms to optimize a criterion, the definition of a criterion and the extension of the original problem itself.

This section both justifies the construction $M^{\times}$ as a comparison to one of the two canonical equilibria which reduces the available information and introduces the second canonical construction M^{+} (see Definition 9). Let us first start with some usual definitions for a graph:

Definition 5 (Weighted graph)**.**
A weighted graph G contains n vertices $1 \leq i \leq n$, which are connected each other through edges (i, j) linked with weights $a_{i,j}$ (representing a weighted incidence matrix). We also introduce the total weight $2M = \sum_{i,j} a_{i,j}$.

Graph clustering is devoted to the research of classes, groupings, modules or cliques (whatever we call them) within a graph. They are defined through an equivalence relation as specified in Definition 6:

Definition 6 (Graph clustering)**.**
Let us call X, a matrix representation of a binary equivalence relation, the result of the clustering of a graph G. Then $X_{i,j}$ equals 0 or 1 and equals 1 if and only if the two vertices i and j are in the same class for X, and 0 if not.

Problem 4 (Generic clustering problem)**.**

$$\begin{aligned} \max_{X} \quad & W(w, X) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} X_{i,j} \\ \text{subject to:} & \\ X \quad & \text{is an equivalence relation} \end{aligned}$$

First let us remark that, as notably spotted in [22], [26], [31] an equivalence relation constraint can be written as :

- $X_{i,i} = 1, \ \forall 1 \leq i \leq n$ (reflexivity)
- $X_{i,j} = X_{j,i}, \ \forall 1 \leq i, j \leq n$ (symmetry)

- $X_{i,j} + X_{j,k} - X_{j,k} \leq 1,\ \forall 1 \leq i,j,k \leq n$ (transitivity)

Remark 5.
Thanks to this relational formulation, appearing in the model presented just above, it is not mandatory to fix in advance the number K of clusters which is expected, this latter is got as a direct and final result of the computing process.

4.2.1 Modularity criterion of Newman-Girvan based upon "Independence"

The original, famous, and well known Newman-Girvan's presentation of a global criterion for graphs clustering, see [14] or [30], has been introduced inside the generic Louvain's algorithm together with a global cost called "Modularity" defined by:

Definition 7 (Modularity)**.**
Given an equivalence relation $X_{i,j}$ (also called partition) and a graph G with weighted function $a_{i,j}$ on its edges, the global modularity N.G. criterion amounts to:

$$W^{\times}(G,X) = \frac{1}{2M} \sum_{i,j} \left[a_{i,j} - \frac{a_{i,\cdot} a_{\cdot,j}}{2M} \right] X_{i,j} \tag{8}$$

Let us first remark that the original modularity $W^{\times}$ is nothing but our generic global cost function defined through Problem 4 with:

$$w_{i,j} = w^{\times}(G)_{i,j} = \frac{a_{i,j}}{2M} - \frac{a_{i,\cdot} a_{\cdot,j}}{(2M)^2}$$

and the local gain $w^{\times}(G)_{i,j}$ to put two vertices in the same class is the local deviation to *independence*. Indeed, using Definition 5, we know that $\pi_{i,j} = \frac{a_{i,j}}{2M}$ can be seen as a probability measure on $\{1 \dots n\}^2$ with margins $\mu_i = \frac{a_{i,\cdot}}{2M}$ so that $w^{\times}$ rewrites:

$$w^{\times}(G)_{i,j} = 2M \left(\pi_{i,j} - \mu_i \mu_j \right)$$

and does express itself as a canonic *deviation to independence* criterion.

4.2.2 Extended logical modularity based upon "*indetermination*"

Problem 4 basically represents an extension of the already introduced "Modularity criterion" towards a generic criterion based on a local input one.

We suggest an expression $w^{+}(G)_{i,j}$ which represents a deviation to *indetermination*. It will be used as a local cost function in Problem 4 leading to a slightly different global formula $W^{+}(G,X)$ to optimize locally:

$$w^{+}(G)_{i,j} = a_{i,j} - \frac{a_{i,\cdot}}{n} - \frac{a_{\cdot,j}}{n} + \frac{2M}{n^2}$$

Symmetrically as $w^{\times}$, it ends up being a canonic *deviation to indetermination* criterion. Indeed, keeping $\pi_{i,j} = \frac{a_{i,j}}{2M}$, w^{+} rewrites:

$$w^{+}(G)_{i,j} = 2M * \left(\pi_{i,j} - \frac{\mu_i}{n} - \frac{\mu_j}{n} + \frac{1}{n^2} \right)$$

The global criterion being:

$$W^{+}(G,X) = \sum_{i,j} \left[a_{i,j} - \frac{a_{i,\cdot}}{n} - \frac{a_{\cdot,j}}{n} + \frac{2M}{n^2} \right] X_{i,j} \tag{9}$$

It has been shown in [9] that those two costs play a specific role as deviations from canonic couplings. Inserted in the generic Louvain algorithm, they allow to get clustering solutions $X_{i,j}$ which slightly differ from each other, but not that much (see for instance [8], where some comparisons of the criteria of Definition 7 and Definition 7 are provided, in practical contexts). We will not specifically insist here on it, but rather propose two new applications of the indtermination structure.

4.3 Guessing or spy problem

4.3.1 Original problem

In cryptography, a message u in a finite alphabet $\mathcal{U}$ of size p is typically sent from Alice to Bob while a spy whose name is Charlie tries to intercept it. A common strategy for Alice to communicate efficiently and secretly with Bob consists in encoding the message using a couple of keys (public, private) for each character or a symmetric encryption which only requires one shared key between Alice and Bob. The literature concerned with the encryption method to choose according to the situation is diverse, the most-used standard is Advanced Encryption Standard described in various articles. Possibly, Charlie observes an encrypted message V in a second finite alphabet $\mathcal{V}$ of size q which is a function of the message u.

Related to the cryptography situation, the guessing problem quoted hereafter as Problem 5 was first introduced in the article [27]. While in cryptography Charlie tries to decode a sequence of messages, the guessing problem focuses on decoding a unique message. Furthermore, the initial version of Problem 5 is limited due to the lack of access to any prior knowledge by the spy. A second version described in subsection 4.3.2 will introduce a variable V correlated to the message, this second variable will code but will not be limited to code the encrypted message conveyed from Alice to Bob.

Though, the original version provides a collection of results that easily transpose themselves to the more realistic one. Let us formalize this simplest situation: U is a random variable which takes its values in a finite alphabet $\mathcal{U}$ and follows the probability law $\mathbb{P}_U = \mu$. A sender "Alice" generates a sequence of independent messages under μ.

Problem 5 (Original Guessing Problem or Spy Problem)**.**
When Alice sends a message $U = u$ to Bob, the spy Charlie must find out the value u of the realization. He has access to a sequence of formatted questions for any guess $\tilde{u}$ he may have: "Does u equal $\tilde{u}$?" for which the binary answer is limited to "yes/no".

Definition 8 (Original Strategy)**.**
A strategy $S = \sigma$ of Charlie is defined by an order on $\mathcal{U}$ representing the first try, the second and so on until number p. It can be deterministic or random: we quote $\mathbb{P}_S$ its probability law.

Besides, for a given position $i \in [1, p]$, $\sigma[i]$ is the element in $\mathcal{U}$ corresponding to the i-th try.

In [27], a measure of performance is associated to any fixed strategy σ of Charlie. It basically computes the ρ moment of G which counts the number of trials needed by Charlie to find out which message u was sent. We shall add another performance measure later on.

Definition 9 (Performance measure)**.**
The function $G(\sigma, u)$ is defined as the number of questions required to eventually obtain a "yes" in Problem 5 when Charlie proposed the order $S = \sigma$ and Alice generated the message $U = u$. It can be a random variable even for a fixed u as soon as S is. $G(S, U)$ is a random variable and whose formal definition is:

$$G(\sigma, u) = \sum_{i=1}^{p} i 1_{\sigma[i]=u}$$

We eventually define the efficiency of a strategy S by a measure of the $\rho-$moment of $G(S,U)$ under the independent coupling of $S \sim P_S$ and $U \sim P_U$.

$$||G(S,U)||_\rho = \left[\mathbb{E}_{(S,U)\sim \mathbb{P}_S \otimes \mathbb{P}_U} \left(G(S,U)^\rho\right)\right]$$

The definition of $G(\sigma, u)$ precisely codes the number of trials before Charlie discovers the message u. For instance, with an alphabet $\mathcal{U} = \{a, b, c, d\}$, if the message is $u = c$ and the strategy σ of the spy consists in the order (b, c, a, d) (meaning he first proposes message b then c,...) we have:

$$\begin{aligned} G(\sigma, u) &= \sum_{i=1}^{p} i 1_{u=\sigma[i]} \\ &= 1 \cdot 1_{u=b} + 2 \cdot 1_{u=c} + 3 \cdot 1_{u=a} + 4 \cdot 1_{u=d} \\ &= 2 \cdot 1_{u=c} \\ &= 2 \end{aligned}$$

It has been proven in the same article [27] a natural result: provided $\mathbb{P}_U = \mu$ is known, the best strategy consists in proposing answers under the deterministic order σ of decreasing probabilities. That is to say we first propose the message which appears most often, then the second most probable and so on:

$$\mu_{\sigma[p]} \leq \ldots \leq \mu_{\sigma[1]}$$

Besides they demonstrated a lower bound on the average number of questions which no strategy can break as it is specified in Theorem 4.

Theorem 4 (Lower bound on the efficiency)**.**
The minimal expected number of questions to solve Problem 5 verifies the inequality:

$$\min_S ||G(S,U)||_\rho \geq (1 + \log(p))^{-\rho} \left[\sum_{u\in\mathcal{U}} \mathbb{P}(U = u)^{\frac{1}{1+\rho}}\right]^{1+\rho}$$

Proof.
We won't gather the whole demonstration but a glimpse. In [2], $||G(S,U)||_\rho$ is written as the integral of some concave function with respect to some probability measure Q. By Jensen's inequality, a lower bound is obtained whatever Q is. The multiplicative factor comes from the inequality

$$\sum_{u\in\mathcal{U}} \frac{1}{G(\sigma, u)} = \sum_{i=1}^{p} \frac{1}{i} \leq 1 + \log(p)$$

Eventually, selecting a special value for Q leads to the result. □

A practical application of Theorem 4 is to provide a guarantee on the average time a spy will take to guess a message. The sender, on its side, is motivated by maximizing the lower bound.

4.3.2 Extended problems

As announced beforehand, Charlie has now access to an observed random variable V correlated with the sent message U. In the common cryptography problem it would be the encrypted message that Charlie observes when Alice sends a message, hence a deterministic function of the message U. Here, we generalize and suppose it can also contain, for instance, the size of the message, the frequency channel used, the sender's location, the receiver, or any physical information a spy can have access to. Finally, the added information, more or less useful, is encoded into a random variable V whose values belong to a finite alphabet $\mathcal{V}$ of size q. Obviously, V is correlated with the message U but we do not suppose their link is deterministic as it would be for an encryption.

As mentioned in the article [2], Charlie now chooses its strategy according to the value taken by the observed second variable V: he typically adapts himself to the conveyed encryption. The probability law of the couple (U, V) is quoted $\mathbb{P}_{U,V} = \pi$ while its margins are $\mathbb{P}_U = \mu$ and $\mathbb{P}_V = \nu$.

The gain function now expresses as $G(S, U|V)$: we purposely use the notation symbol "knowing V" to insist on the fact that V is known when the spy decides the strategy he uses. Eventually, for any observed value $V = v$, an original strategy S_v (see Definition 8) is built up leading to an original gain function $G(S_v, U)$ that is to say:

$$G(S, U|V) = \sum_{v \in \mathcal{V}} G(S_v, U) 1_{V=v}$$

The same article comes up with a generalization of Proposition 4 that we report here:

Theorem 5 (Generalized lower bound on the efficiency)**.**
For any strategy, the average time to reconstruct the message always respects the lower bound:

$$\mathbb{E}_{(S,U,V)\sim\mathbb{P}_{S,U,V}}[G(S, U|V)^\rho] \geq (1 + \log(p))^{-\rho} \sum_{v \in \mathcal{V}} \left[\sum_{u \in \mathcal{U}} (\pi_{u,v})^{\frac{1}{1+\rho}} \right]^{1+\rho}$$

Proof.
The result is plain given that, as we already noticed, S decomposes into original strategies S_v for any fixed v. Hence, for any v, the local or original assigned strategy obeys Proposition 4 which directly leads to the result. □

4.3.3 Logical *indetermination* as a lower bound

Let us move away from the literature and measure Charlie's performance by its probability to find out after one trial the message u Alice sent. It is a reasonable measure as, if a sequence of messages is sent, we may have to jump from one to the following after only one trial.

Definition 10 (one-shot performance)**.**
For a given strategy S, we define the following performance measure as the probability to find out the value u after one trial, formally:

$$M(S, U, V) = \mathbb{P}_{(S,U,V)\sim\mathbb{P}_{S,U,V}}(S[1] = U)$$

Remark 6 (Generalized one-shot performance)**.**
One could easily introduce a measure whose name could be "k shots performance" evaluating the probability to guess after up to k trials. We would hence notice that if $k \geq p$ then the "k shots performance" equals 1 for any sensitive strategy. We will not detail it further here.

We suppose as for the original optimal strategy that the spy has access to the distribution $\mathbb{P}_{U,V} = \pi$. We can imagine he previously observed the non-encrypted messages in a preliminary step.

Two strategies immediately stand out:

1. S_{max}: systematically returns at v fixed (observed by hypothesis), the u associated with the maximal probability on the margin $\mathbb{P}_{U|V=v}$
2. S_{margin}: returns at v fixed a random realization of x under the law $\mathbb{P}_{U|V=v}$

While we know S_{max} is the best strategy in case the performance measure is $||G(S,U)||_\rho$, we have no guarantee it maximizes the one-shot performance. Furthermore, S_{margin} is by far harder to cope with for the sender who cannot easily prevent random conflicts. Consequently we come back to the reduction of couple matchings, whose "*indetermination* coupling", we know, prevents us against. Let us unfold this remark hereafter.

The one-shot performance of the two strategies is given by:

$$M(S_{max}, U, V) = \sum_{v\in\mathcal{V}} \nu_v \left[\max_{u\in\mathcal{U}} \pi_{u|V=v}\right] \tag{10}$$

$$M(S_{marge}, U, V) = \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} \nu_v (\pi_{u|V=v})^2 \tag{11}$$

Let us suppose, commendable task if any, that Alice wants to minimize Charlie's one-shot performance. We also suppose that the margins μ on U and ν on V are fixed. It is a common hypothesis: the alphabet $\mathcal{U}$ in which the messages are composed typically respects a distribution on letters; variable V on its own, if it represents frequencies for instance may have to satisfy occupation weights on each channel.

Concerning the strategy S_{marge} we have the two bounds:

$$\frac{||\pi||_2^2}{\min_{v\in\mathcal{V}} \nu_v} \geq M(S_{marge}, U, V) \geq \frac{||\pi||_2^2}{\max_{v\in\mathcal{V}} \nu_v} \tag{12}$$

with

$$||\pi||_2 = \sqrt{\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} (\pi_{u,v})^2}$$

We notice using Equation (3) that

$$\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} \mathbf{M}^+(\pi_{u,v}) = pq||\pi||_2^2 - 1$$

It shows that studying the guessing problem brings us back to Problem 1 associated with cost $\mathbf{M}^+$ whose solution is given by the *indetermination* coupling of the margins.

Formally, using the coupling

$$\mathbb{P}^+_{U,V} = C^+(\mathbb{P}_U, \mathbb{P}_V) = \frac{\mu_u}{q} + \frac{\nu_v}{p} - \frac{1}{pq} = \pi^+_{u,v}$$

guarantees an efficient reduction of conflicts (see section 3.2) and eventually a weak one-shot performance as expressed in Equation (12).

A coupling to fight Charlie who uses the strategy S_{marge} consists for Alice in dispatching the messages U among V according to the logical *indetermination*. Typically she would use the constructive drawing exposed in subsection 3.4.

Furthermore, the lower bound of the one-shot performance given by Equation 12 is minimal under the *indetermination* π^+ so that regardless of the coupling π actually used, we always have:

$$M(S_{marge}, U, V) \geq \frac{||\pi^+||_2^2}{\max_{v \in \mathcal{V}} \nu_v} \tag{13}$$

This optimality of *indetermination* fits the observation provided on Figure 2: on any visible information V for Charlie, messages U are as evenly distributed as possible which reduces the probability of matching the its first guess.

Eventually, a classic norm inequality allows us to deduce that the one-shot performance of S_{marge} limits from below as well as from above the one-shot performance of S_{max} according to the inequality valid for any couple probability law on (U, V):

$$M(S_{marge}, U, V) \leq M(S_{max}, U, V) \leq \sqrt{M(S_{marge}, U, V)}$$

4.4 Tasks partitioning

Task partitioning problem is originally introduced in [4] where the authors provide a lower bound on the moment of the number of tasks to perform. Let us follow the gathering work of [20] where they also coin a generalized task partitioning problem basically adapting it as a special case of the guessing problem.

Formally, we begin with the original problem of tasks partitioning: a finite set $\mathcal{U}$ of tasks size of which is quoted p is given together with an integer $q \leq p$. The problem consists in creating a partition $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_q)$ of $\mathcal{U}$ in q classes to minimize the number of tasks to perform, knowing that if one needs to perform a task $u \in \mathcal{A}_i$, it is mandatory to launch simultaneously the whole subset of tasks included within $\mathcal{A}_i$.

Practically, a task $U = u$ to perform is randomly drawn from $\mathcal{U}$ under a probability distribution $\mathbb{P}_U = \mu$ representing the tasks frequencies. As any task, the task u to perform is assigned to a unique class $\mathcal{A}_{i(u)}$ of the arbitrary partition. Hence, $A(u) = |\mathcal{A}_{i(u)}|$ counts the number of tasks to perform. Precisely, one plays on the partition knowledge to perform, in average, as few tasks as possible.

Similarly to the guessing problem, the performance of a partition $\mathcal{A}$ is estimated through the $\rho-$moment of $A(U)$, formally $\mathbb{E}_U\left[A(U)^\rho\right]$. Moreover, the authors show in [20], quite similarly as for Theorem 4 that we have:

$$\min_{\mathcal{A}} \mathbb{E}_{U \sim \mu}\left[A(U)^\rho\right] \geq \frac{1}{q}\left[\sum_{u \in \mathcal{U}} (\mu_u)^{\frac{1}{1+\rho}}\right]^{1+\rho} \tag{14}$$

which expresses a minimum average number of tasks to perform whatever the partition is.

Inspired by the general guessing problem, they extend the task partitioning problem. Let us introduce here this generalized version, in which we are no longer interested in minimizing the number of tasks to perform but rather in reducing the number of tasks before a selected (or a chosen) task u.

Indeed, in the first version, as soon as u is drawn, an arbitrary rule imposes to perform the whole subset $\mathcal{A}_{i(u)}$ leading to realize $A(u)$ tasks. In the new version, tasks are performed sequentially in $\mathcal{A}_{i(u)}$ according to a global strategy S that can be deterministic or random.

Typically, tasks may consists in a signatures flow which an administration requires while q would be the number of workers dedicated to perform those signatures on incoming documents. A worker can be given the entitlement to perform several signatures, assistants usually do. In that case, the partition encodes the assignments of tasks to workers. When a worker $V = v$

is assigned a document, the depositor waits until the signature. Then the worker follows his own strategy S_v to sign his assigned documents, meaning he can always follow the same order leading to a deterministic strategy or change every day leading to a random strategy.

With a global strategy S which gathers the workers' strategies S_v, $\forall 1 \leq v \leq q$ and for a task u to perform, the performance of a partition $\mathcal{A}$ is measured using

$$N_{S,\mathcal{A}}(u)$$

which represents the number of tasks performed before the intended task u (u included). A lower bound is provided in the paper [20].

Let us now suppose the keys $1 \leq v \leq q$ are associated with offices that must perform a proportion ν_v of the incoming tasks which still follow a distribution μ. It actually appears as a sensible problem where a manager would have to distribute in advance tasks among teams according to the usual observed distribution of tasks and a list of available teams with their capacities.

Besides, we suppose each team uses the strategy S_{marge} to perform tasks, meaning they randomly perform one according to their margin theoretical distribution; for a document signing, they randomly sign one.

Remark 7 (Concrete estimated distribution)**.**
In any of the previous applications, for spy as well as for tasks, we are dealing with probabilities. Actually, we send a finite and integer number n of messages and we similarly distribute a finite number n of tasks.

Moreover, $\mathcal{U}$ and $\mathcal{V}$ are finite. Eventually, for any $u \in \mathcal{U}$ and $v \in \mathcal{V}$, an integer number $n_{u,v}$ of tasks is associated corresponding (in the spy problem) to the number of same letters u sent using channel v.

To convert $n_{u,v}$ into a probability measure, one only needs to divide by n and define:

$$\pi_{u,v} = \frac{n_{u,v}}{n} \tag{15}$$

Reciprocally, given a probability measure, one will draw n messages according to π. As n increases, it will approximate the theoretical distribution better and better.

Eventually, we notice that we do not have an explicit formula to optimize any of the two transport problems (Problem 2 or Problem 3) with an additional hypothesis of π respecting the fraction-of-integers-form given in Equation (15). Furthermore, indetermination or independence optimizes with a relaxed hypothesis (values in $\mathbb{R}$) and thus provides a better solution; to be approached by drawings.

From now on, we can rewrite our task partitioning problem under the form of a guessing problem:

- $V = v$, formerly corresponds to a worker, now it represents the information the spy has access to
- $U = u$, formerly represents a task to perform, now it represents a sent message
- $S = \sigma$, formerly represents the order in which tasks are performed, now it represents the order in which Charlie proposes his guesses

Under this formalism, we are interested in measuring the probability $M(S, U, V)$ of executing u first as an extended application of the one-shot performance of Definition 10 and we have:

$$M_(S, U, V) \geq \frac{||\pi||_2^2}{\max_{1 \leq v \leq q} \nu_v} \geq \frac{||\pi^+||_2^2}{\max_{1 \leq v \leq q} \nu_v} \tag{16}$$

This inequality provides a lower bound for any distribution of the tasks among the team, no distribution can generate a worst "one-shot probability" of satisfying the intended task.

In task partitioning actually, each u is uniquely associated to a worker $v = i(u)$ so that the random variable representing the worker is deterministic conditionally to U. Yet, it is a reducing case of the guessing problem where V is random.

Remark 8 (Splitting mass)**.**
Eventually, we notice that having V random conditionally on U is a generalization of task partitioning along the same lines the Monge-Kantorovith problem was an extension of the one dimension Monge problem: we allow the mass splitting possibility, since a task may be randomly assigned among several workers.

In task partitioning problem using a partition $\mathcal{A}$ instead of V (hence allowing no mass splitting), we notice $\mathbb{P}_Y = \nu$ is not properly defined. Let us extract it from the partition $\mathcal{A}$ by providing each worker with a probability which sums up the probabilities of the tasks he has to perform. Formally, we define $V(\mathcal{A})$ to be a random variable whose probability is:

$$\nu_v^{\mathcal{A}} = \mathbb{P}_{V(\mathcal{A})}(V(\mathcal{A}) = v) = \sum_{u \in \mathcal{A}_v} \mu_u$$

together with the couple probability:

$$\pi_{u,v}^{\mathcal{A}} = \mathbb{P}_{U,V(\mathcal{A})}\left(U = u, V(\mathcal{A}) = v\right) = \mu_u 1_{u \in \mathcal{A}_v}$$

It enables us to deduce that the one-shot probability of satisfying the intended task for a partitioning problem accepts as a lower bound:

$$M(S,U,\mathcal{A}) = M\left(S,U,V(\mathcal{A})\right) \geq \frac{||\pi^{\mathcal{A}}||_2^2}{\max_{1 \leq v \leq q} \nu_v^{\mathcal{A}}} \geq \frac{||C^+(\mu,\nu^{\mathcal{A}})||_2^2}{\max_{1 \leq v \leq q} \nu_v^{\mathcal{A}}}$$

Indeed, a partition problem appears as a particular coupling of U and $V(\mathcal{A})$ (where $V(\mathcal{A})|U$ is deterministic) and no coupling can be worse than $C^+\left(U, V(\mathcal{A})\right)$.

A direct application is that no office affectation should provide a worse one-shot performance...

The efficiency of *indetermination* coupling in guessing problem as well as in task partitioning directly comes from its ability to reduce couple matchings. Either it prevents the spy from discovering the message or it provides a worst strategy by preventing a task from being performed.

As mentioned in the introduction, we shall see in the next section that a continuous *indetermination* notion exists and we will construct it using both prior guess and computations.

5 Continuous *indetermination*

5.1 Coupling function

This chapter introduces a continuous *indetermination* coupling that, to the extent of our knowledge, was introduced in [24], but never studied as such. We start with two marginals probability measures μ and ν on two segments $[a, A]$, $[b, B]$ and our goal is to define a probability measure on $[a, A] \times [b, B]$ which extends *indetermination* in a continuous space. Introducing some notations, we suppose that μ and ν respectively have a density measure quoted f and g together with an associated cumulative distribution function F and G.

The coupling function notion of Definition 1 operates on the discrete weights μ_u and ν_v whose transposition appears to be $f(u)\,\mathrm{d}u$ and $g(v)\,\mathrm{d}v$ for any $u \in [a, A]$ and $v \in [b, B]$. Using that transposition, we extend the definition of a coupling function in a continuous space

Definition 11 (Coupling functions (continuous))**.**
μ and ν being probability laws on segments $[a, A]$ and $[b, B]$, a coupling function C operates on their density (f, g).

$C(f, g)$ represents a density for a probability measure on the product space whose margins precisely are μ and ν.

A typical continuous (as well as discrete) coupling function is the independance quoted $C^\times$ (we extend here the discrete notation) which generates an eligible density under the formula:

$$C^\times(f, g)(u, v) = f(u)g(v), \ \forall u \in [a, A], \ \forall v \in [b, B]$$

The $\otimes$ notation is on purpose since it corresponds to the usual continuous *independence* coupling. More precisely, the density of an *independence* coupling of the two margins μ and ν will be nothing but $C^\times(f, g)$.

We would like to define a continuous version of the *indetermination* coupling. As often, when transposing a concept, we can either obtain it through computations or using a prior guess. We will follow the two approaches. Let us first propose a prior guess.

Definition 12 (Continuous *indetermination* density)**.**
f and g being densities of two probability measures on $[a, A]$ and $[b, B]$ respectively, C^+ operates on the couple (f, g) with the formula:

$$C^+(f, g)(u, v) = \frac{f(u)}{B - b} + \frac{g(v)}{A - a} - \frac{1}{(A - a)(B - b)}$$

The equation comes from an adaptation of the discrete one, although this adaptation seems coherent, at that stage no guarantee is given on its truthfulness neither on its construction. Let us begin with computing the margin for a fixed $u \in [a, A]$ (it is obviously symmetric for v):

$$\begin{aligned}
\int_{v=b}^{B} C^+(f, g)(u, v) \, \mathrm{d}u \, \mathrm{d}v &= \int_{v=b}^{B} \left(\frac{f(u)}{B - b} + \frac{g(v)}{A - a} - \frac{1}{(A - a)(B - b)} \right) \mathrm{d}u \, \mathrm{d}v \\
&= f(u) \, \mathrm{d}u + \frac{1}{A - a} \, \mathrm{d}u - \frac{1}{(B - b)} \, \mathrm{d}u \\
&= f(u) \, \mathrm{d}u
\end{aligned}$$

It appears actually that margins are as expected. Yet, they must also be continuous densities and therefore positive, which is by the way, not always true for $C^+(f, g)$. As in the discrete case an additional hypothesis is therefore required:

$$\min_u f(u) + \min_v g(v) \geq 1 \tag{17}$$

this last equation being the continuous version of Hypothesis 4.

Eventually, applied on any couple of margins whose densities respect Equation (17), the continuous C^+ of Definition 12 generates an eligible density which is similar to the discrete optimal coupling function which solves Problem 3. Let us unfold this remark by defining a continuous version of Problem 3.

First, to simplify any future computation, we pose $a = b = 0$ and $A = B = 1$, converting any formula will be done using a dedicated affine transformation.

We use transport problem to validate our prior guess, transposing in the continuous space the discrete Minimal Transport problem using square deviation as a cost function:

Problem 6 (Minimal Trade Problem)**.**

$$\begin{aligned}
\min_{\pi} \quad & \int_0^1 \int_0^1 \pi^2(u,v)\,\mathrm{d}v\,\mathrm{d}v \\
\text{under constraints:} \quad & \\
& \int_{v=0}^1 \pi(u,v)\,\mathrm{d}v\,\mathrm{d}u = \mu_u\,\mathrm{d}u \\
& \int_{u=0}^1 \pi(u,v)\,\mathrm{d}v\,\mathrm{d}u = \nu_v\,\mathrm{d}v \\
& \int_0^1 \int_0^1 \pi(u,v)\,\mathrm{d}u\,\mathrm{d}v = 1 \\
& \pi \geq 0
\end{aligned}$$

We then add Inequality (17), which, as Hypothesis (4) guarantees that our prior guess function is the density of a probability law. Hence, as well as restricting ourselves to $[0,1]$ we restrict ourselves to margins respecting the condition.

Property 3.
Under the ad hoc Hypothesis (17), the solution of Problem 6 is nothing but our prior guess continuous coupling function of Definition 12 applied to the margins f and g, formally, the solution density function is:

$$C^+(f,g)(u,v) = c^+_{f,g}(u,v) = (f(u) + g(v) - 1)$$

Proof.
We use Definition 12 to efficiently solve Problem 6 by noticing that

$$\int_{u=0}^1 \int_{v=0}^1 \left[\pi(u,v) - (f(u) - g(v) - 1)\right]^2 \mathrm{d}u\,\mathrm{d}v \geq 0$$

rewrites (using margins constraints):

$$\begin{aligned}
\int_{u=0}^1 \int_{v=0}^1 \pi^2(u,v)\,\mathrm{d}u\,\mathrm{d}v &\geq \int_{u=0}^1 \int_{v=0}^1 \left(-f^2 - g^2 - 1 - 2\pi - 2fg + 2f\pi + 2g\pi + 2f + 2g\right)\mathrm{d}u\,\mathrm{d}v \\
&= \int_{u=0}^1 \int_{v=0}^1 \left(-f^2 - g^2 - 1 - 2 - 2fg + 2f^2 + 2g^2 + 2 + 2\right)\mathrm{d}u\,\mathrm{d}v \\
&= \int_{u=0}^1 \int_{v=0}^1 \left(f^2 + g^2 - 2fg + 1\right)\mathrm{d}u\,\mathrm{d}v \\
&= \int_{u=0}^1 \int_{v=0}^1 (f + g - 1)^2\,\mathrm{d}u\,\mathrm{d}v
\end{aligned}$$

At the end, using Hypothesis (17) we notice that $c^+_{f,g}(u,v) = (f(u) + g(v) - 1)$ is eligible as a density function since always positive. Margins constraints are also satisfied as computed right after the introduction of continuous C^+. □

Eventually, following our computations, and given two random variables on $[0,1]$: $U \simeq \mu$, (f,F) and $V \simeq \nu$, (g,G) we define a third variable W on the product space $[0,1] \times [0,1]$. W respects the two margins probability laws μ, ν and is called *indetermination* coupling of U and V quoted $U \oplus V$. Its density is $h^+_{f,g} = C^+(f,h)$ while its cumulative distribution function is quoted $H^+_{F,G}$ and will be computed hereafter.

To compute the cumulative distribution function associated to the just expressed density $h^+_{f,g}$ we only have to apply a quick integration of the density. It leads to a second characterization of an *indetermination* coupling.

Property 4 (*indetermination* cumulative distribution function)**.**
Given two random variables U and V whose densities f, g satisfy Hypothesis (17), the cumulative distribution function associated to their indetermination coupling $U \oplus V$ is:

$$H^+_{F,G} = vF(u) + uG(v) - uv$$

.

5.2 Constructive method for adapted margins

Beforehand, we assumed margins are respecting Hypothesis (17), we propose here a method to construct adapted margins out of any couple.

Property 5 (Constructive margins)**.**
A couple (f, g) of densities fulfills Hypothesis (17) if and only if, it exists an $\alpha, 0 \leq \alpha \leq 1$ and a couple of densities (r, s) such that :

$$f = (1-\alpha)r + \alpha \quad \text{and} \quad g = \alpha s + 1 - \alpha \tag{18}$$

Proof.
Let us first suppose (f, g) is under this form, then we notice, $\min f \geq \alpha$ as well as $\min g \geq 1-\alpha$, hence, $\min f + \min g \geq 1$ so that the condition is satisfied.

Now, if the condition is respected, we define $\alpha = \min f$ and have $g \geq 1 - \alpha$. If $\alpha = 0$ or $\alpha = 1$ then, respectively, G or F is uniform so that the corresponding coupling is *independence* and condition is degenerated. If not, $0 < \alpha < 1$ and we can write:

$$f = (1-\alpha)\frac{f-\alpha}{1-\alpha} + \alpha$$

together with:

$$g = \alpha\frac{g-(1-\alpha)}{\alpha} + (1-\alpha)$$

Quoting $r = \frac{f-\alpha}{1-\alpha}$ and $s = \frac{g-(1-\alpha)}{\alpha}$, we have $0 \leq r, s \leq 1$ and $\int r = \int s = 1$. It precisely shows that (r, s) is a couple of densities on $[0, 1]$. □

Using the last proposition, we can easily construct a couple of margins on which an *indetermination* coupling is feasible. Though, as mentioned during the proof, if $\alpha \in \{0, 1\}$ then f or g is an uniform density for which *indetermination* and *independence* merge one with another.

6 Conclusion

This paper gathers a list of properties as well as applications to various problems of a canonical coupling called *indetermination*.

In section 2 we introduced the notion of coupling function together with two particular examples: *independence* and *indetermination*. The construction of both showed they minimize the information the fixed margins constraints convey into the joined law. This no-information property encouraged us to coin graph clustering criterion out of each equilibrium since a criterion usually compare the observed graph with the expected "null graph respecting vertices degrees".

Section 3 focuses on the study of *indetermination*. After the usual Monge property, we proposed a constructive decomposition of an *indetermination* matrix and explained how it helped to avoid couple matchings. Furthermore, we noticed this property was already expressed in the cost M^+ of the problem it solves by writing it again as the expected number of couple matchings.

Knowing it is the best construction to reduce couple matchings led us to section 4 in which that property was used in two applications (Guessing Problem and Task Partitioning Problem) in which one can interpret a couple matching: either spy right guesses or performed tasks. We respectively demonstrated that *indetermination* could protect a sender from a spy with a specific strategy and that it appeared in the worst task partitioning method.

The presence of this second equilibrium in at least three usual problems (graph clustering, guessing problem and task partition) reinforces the usefulness of section 3.

Finally, we extended the *indetermination* coupling in a continuous space expressing a prior guess formula as the optimal solution of a continuous version of the original Minimal Trade Model.

We repeatedly noticed in [3] that properties of *independence* usually have a symmetric transcript for *indetermination*. In the continuous case, we know we can define an *independence* copula, operating on margins cumulative distribution functions to generate the cumulative distribution function of an *independence* coupling. Yet, even if we built up a continuous *indetermination* notion in section 5, coining an *indetermination* copula is not straightforward. We will develop the computations it requires and the applications it may have in a future article to be published.

References

[1] Ah-Pine, J. On aggregating binary relations using 0-1 integer linear programming. *workshop ISAIM* (2009).

[2] Arikan, E. An inequality on guessing and its application to sequential decoding. *IEEE Trans. Inform. Theory 42* (1996).

[3] Bertrand, P., Broniatowski, M., and Marcotorchino, J.-F. "Statistical Independence" versus "Logical Indetermination", two ways of generating clustering criteria through couplings : Application to graphs modularization. working paper or preprint, July 2020.

[4] Bunte, C., and Lapidoth, A. Encoding tasks and Rényi entropy. *IEEE Trans. Inform. Theory 60*, 9 (Sep 2014), 5065–5076.

[5] Burkard, R. E., Klinz, B., and Rudolf, R. Perspectives of Monge properties in optimization. *Discrete App. Math. 70* (1996), 95–161.

[6] Chen, M., Nguyen, T., and Szymanski, B. K. A new metric for quality of network community structure. *arXiv preprint arXiv:1507.04308* (2015).

[7] Clauset, A., Newman, M. E., and Moore, C. Finding community structure in very large networks. *Phys. Rev. E 70*, 6 (2004), 066111.

[8] Conde-Cespedes, P. *Modélisations et extensions du formalisme de l'analyse relationnelle mathématique à la modularisation des grands graphes.* PhD thesis, Paris 6, 2013.

[9] Csiszar, I., et al. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *Ann. Statist. 19*, 4 (1991), 2032–2066.

[10] Fasino, D., and Tudisco, F. A modularity based spectral method for simultaneous community and anti-community detection. *Proc. Natl. Acad Sci USA 542* (2018), 605–623.

[11] Forgy, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics 21* (1965), 768–769.

[12] Fortunato, S. Community detection in graphs. *Phys. Rep. 486*, 3-5 (2010), 75–174.

[13] Fortunato, S., and Barthelemy, M. Resolution limit in community detection. *Proc. Natl. Acad Sci USA 104*, 1 (2007), 36–41.

[14] Girvan, M., and Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad Sci USA* (2002), 7821–7826.

[15] Gomez, D., Rodríguez, J. T., Yanez, J., and Montero, J. A new modularity measure for fuzzy community detection problems based on overlap and grouping functions. *Internat. J. Approx. Reason. 74* (2016), 88–107.

[16] Janson, S., and Vegelius, J. Correlation coefficients for nominal scales. *Uppsala: Department of Statistics* (1977).

[17] Janson, S., and Vegelius, J. On the applicability of truncated component analysis based on correlation coefficients for nominal scales. *Applied Psychological Measurement* (1978), 135–145.

[18] Janson, S., and Vegelius, J. The J-index as a measure of nominal scale response agreement. *Applied Psychological Measurement* (1982), 111–121.

[19] Katz, L. A new status index derived from sociometric analysis. *Psychometrika 18*, 1 (1953), 39–43.

[20] Kumar, A., Sunny, A., Thakre, A., Kumar, A., and Manohar, G. A general moment minimization problem concerning guessing, source coding and tasks partitioning. *TBD* (2021).

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

[22] Marcotorchino, J.-F. Maximal association theory as a tool of research. *Classification as a tool of research, W.Gaul and M. Schader editors, North Holland Amsterdam* (1986).

[23] Marcotorchino, J.-F. Seriation problems:an overview. *Applied Stochastic Models and Data Analysis 7* (1991), 139–151.

[24] Marcotorchino, J.-F., and Cespedes, P. C. Optimal transport and minimal trade problem, impacts on relational metrics and applications to large graphs and networks modularity. *Geometric Science of Information, Springer* (2013), 169–179.

[25] Marcotorchino, J.-F., and El Ayoubi, N. Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. *Revue de Statistique Appliquée 39* (1991), 25–46.

[26] Marcotorchino, J.-F., and Michaud, P. *Optimisation en Analyse Ordinale des Données*. Masson, 1979.

[27] Massey, J. L. Guessing and entropy. *IEEE Int. Symp. on Info. Th.* (1994), 204.

[28] Messatfa, H. Maximal association for the sum of squares of a contingency table. *Revue RAIRO, Recherche Opérationnelle 24* (1990), 29–47.

[29] Nascimento, M. C. Community detection in networks via a spectral heuristic based on the clustering coefficient. *Discrete App. Math. 176* (2014), 89–99.

[30] Newman, M. E. J., and Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E 69*, 2 (2004), 026113.

[31] Opitz, O., and Paul, H. Aggregation of ordinal judgements based on condorcet's majority rule. *Data Analysis and Decision Support. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg* (2005).

[32] Reichardt, J., and Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E 74*, 1 (2006), 016110.

[33] Santiago, R., and Lamb, L. C. Efficient modularity density heuristics for large graphs. *European J. Oper. Res. 258*, 3 (2017), 844–865.

[34] Sklar, A. Random variables, joint distribution functions and copulas. *Kybernetika* (1973), 449–460.

[35] Steinhaus, H. Sur la division des corps matériels en parties. *Bulletin de l'acadéemie polonaise des sciences, v. 4, no. 12* (1957), 801–804.

[36] Stemmelen, E. Tableaux d'échanges, description et prévision. *Cahiers du Bureau Universitaire de Recherche Opérationnelle 28* (1977).

[37] Wakita, K., and Tsurumi, T. Finding community structure in mega-scale social networks. In *Proceedings of the 16th international conference on World Wide Web* (2007), pp. 1275–1276.

[38] Wilson, A. G. A statistical theory of spatial distribution models. *Transportation Research 1* (1967), 253–269.

[39] Wilson, A. G. The use of entropy maximising models. *Journal of transport economies and policy 3* (1969), 108–126.

[40] Wilson, A. G. *Entropy in Urban and Regional Modelling.* Pion, London, 1970.