A Conditional Information Inequality and its Combinatorial Applications

Tarik Kaced, Andrei Romashchenko, and Nikolay Vereshchagin

***Abstract*—We show that the inequality $H(A|B,X) + H(A|B,Y) \leqslant H(A|B)$ for jointly distributed random variables A,B,X,Y, which does not hold in general case, holds under some natural condition on the support of the probability distribution of A,B,X,Y. This result generalizes a version of the conditional Ingleton inequality: if for some distribution $I(X:Y|A) = H(A|X,Y) = 0$, then $I(A:B) \leqslant I(A:B|X) + I(A:B|Y) + I(X:Y)$.**

We present two applications of our result. The first one is the following easy-to-formulate theorem on edge colorings of bipartite graphs: assume that the edges of a bipartite graph are colored in K colors so that each two edges sharing a vertex have different colors and for each pair (left vertex x, right vertex y) there is at most one color a such both x and y are incident to edges with color a; assume further that the degree of each left vertex is at least L and the degree of each right vertex is at least R. Then $K \geqslant LR$. The second application is a new method to prove lower bounds for biclique cover of bipartite graphs.

Keywords*—*Shannon entropy, conditional information inequalities, non Shannon type information inequalities, biclique cover, edge coloring

I. Introduction

The most general and fundamental properties of Shannon's entropy can be expressed in the language of linear inequalities. The usual universal information inequalities (the linear inequalities that hold for Shannon's entropies of jointly distributed tuples of random variables for every distribution) have many equivalent characterizations and interpretations in very different areas — these inequalities can be equivalently reformulated in the settings of Kolmogorov complexity and group theory; they give characterizations of the network coding capacity rates, of the cardinalities of projections of finite sets, etc., see the surveys in [11], [22], [23]. The parallel and interplay between different "incarnations" of information inequalities lead to their better understanding and to more efficient applications of this technique. However, there exists a class of less common information inequalities that still lack a satisfactory explanation and have no clear combinatorial interpretation. These are the *conditional linear information inequalities*, which hold only for distributions that satisfy some constraints. The first nontrivial example of a conditional linear information inequality was proven in the seminal paper [6]; see a survey of other similar results in [25]. Until now, these inequalities looked like artifacts without practical or theoretical application. In this paper, we argue that some conditional inequalities can be naturally interpreted in purely combinatorial terms. We propose a new "conditional information inequality," discuss its combinatorial meaning, and show how it can be employed in purely combinatorial proofs.

Let A, X, Y be jointly distributed discrete random variables. In this paper, we consider the inequality

$$H(A|X) + H(A|Y) \quad \leqslant \quad H(A), \tag{1}$$

where $H(\cdot)$ stands for Shannon's entropy. For some A, X, Y this inequality is false, e.g., for constant X, Y and non-constant A. We provide a natural condition on the distribution of A, X, Y implying inequality (1). Then we provide two combinatorial applications of the resulting conditional inequality and show that it implies the conditional inequality from [21].

More specifically we consider the following condition:

$$\begin{array}{l} \text{for each quadruple of values } a, a', x, y, \\ \text{if the probabilities of all four events} \\ \quad [A=a, X=x], \quad [A=a, Y=y], \\ \quad [A=a', X=x], \ [A=a', Y=y] \\ \text{are positive, then } a = a'. \end{array} \tag{2}$$

Theorem 1. *The inequality* (1) *holds for all random variables* A, X, Y *satisfying* (2).

We first prove this theorem and then show its combinatorial applications.

II. Notation

To simplify formulas, we use the following notation for the marginal distributions (conditional and unconditional): $p(a)$ denotes $\Pr[A = a]$, $p(a,x) = \Pr[A = a,\ X = x]$, $p(a|x) = \Pr[A = a|X = x]$, $p(a,y) = \Pr[A = a,\ Y = y]$, and so on.

If X is a random variable and $\mathcal{E}$ is an event in the same probabilistic space (and $\Pr[\mathcal{E}] > 0$), we denote by $X|\mathcal{E}$ the conditional distribution of X, i.e., the restriction of X on the subspace corresponding to the event $\mathcal{E}$. For example, for jointly distributed random variables (X,Y) we denote by $X|(Y = y)$ the conditional distribution of X under the assumption $Y = y$.

While working on the paper the authors were in part supported by RFBR grant 16-01-00362, by ANR grant RaCAF ANR-15-CE40-0016-01, and by the Russian Academic Excellence Project '5-100'.

Tarik Kaced is with Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, F-75005 Paris, France.

Andrei Romashchenko is with LIRMM, Univ Montpellier, CNRS, Montpellier, France; on leave from IITP RAS.

Nikolay Vereshchagin is with National Research University Higher School of Economics.

III. The proof of Theorem 1

We apply the method of [6], [8]. The crucial property of inequality (1) is that no term contains both X and Y. The inequality (1) can be re-written in terms of unconditional entropies as follows:

$$H(A,X)+H(A,Y)\leqslant H(X)+H(Y)+H(A).$$

Thus it means that the average value of the logarithm of the ratio

$$\frac{p(x)p(y)p(a)}{p(a,x)p(a,y)} \tag{3}$$

is less than or equal to 0. The average is computed with respect to the distribution $p(a,x,y)$. Computing the average, we take into account only the triples (a,x,y) with positive probability. For such triples, both the numerator and denominator of ratio (3) are positive and hence its logarithm is well defined.

Now consider a new distribution p' where

$$p'(a,x,y)=\begin{cases}\frac{p(a,x)p(a,y)}{p(a)} & \text{if } p(a)>0,\\ 0 & \text{otherwise.}\end{cases}$$

Random variables distributed according to p' can be generated by the following process: First generate a using the original distribution of A, then generate independently x using the conditional distribution $x|a$ and y using the conditional distribution $y|a$.

Notice that $p'(a,x,y)$ is positive if so is $p(a,x,y)$ but not the other way around. However, ratio (3) is still well defined and positive for all triples a,x,y with positive $p'(a,x,y)$. Therefore we can compute the average value of the logarithm of (3) using the distribution p' in place of p. Moreover, changing the distribution does not affect the average. Indeed, the logarithm of (3) is the sum of logarithms of its factors. Thus it suffices to show that the average of the logarithm of each factor is not affected when p is replaced by p'. Let us prove this, say, for the factor $1/p(a,x)$.

This factor does not depend on y. Therefore the average of its logarithm does not depend on how $p(a,x)$ is split among $p(a,x,y)$ for different values y: we just sum up $\log 1/p(a,x)$ over all a,x with weights $p(a,x)$. As $p(a,x)=p'(a,x)$, summing with weights $p'(a,x)$ will yield the same result.

By Jensen's inequality[1] the average value of the logarithm of the ratio (3) with respect to the distribution p' is at most

$$\log\Bigl(\sum_{a,x,y:p'(a,x,y)>0} p(x)p(y)\Bigr).$$

The condition (2) guarantees that for each x,y there is at most one a with $p(a,x)>0, p(a,y)>0$ and hence

$$\log\Bigl(\sum_{a,x,y:p'(a,x,y)>0} p(x)p(y)\Bigr)\leqslant \log\Bigl(\sum_{x,y} p(x)p(y)\Bigr) = \log 1 = 0.$$

[1]We need Jensen's inequality for the logarithmic function: let $p_1,\dots,p_n$ be positive numbers that sum up to 1; then $p_1\log x_1+\cdots+p_n\log x_n\leqslant \log(p_1x_1+\cdots+p_nx_n)$.

IV. Combinatorial applications of Theorem 1

A. A lower bound for the number of colors in edge colorings of bipartite graphs

An *edge coloring of a graph* is an assignment of colors to its edges so that each two edges sharing a node have different colors. Finding the *edge coloring number* (the minimum possible number of colors in an edge coloring) of a given graph is a classic problem of graph theory. The study of edge coloring is motivated by theoretical aspects of graph theory as well as by numerous applications in information theory and computer science (mostly by different types of scheduling problems, see a survey in [24]).

Vizing's theorem [1] claims that the edge coloring number of a graph is either its maximum degree d or $d+1$; for bipartite graphs the number of colors is always d. From Theorem 1 we can derive a much stronger lower bound for edge colorings of bipartite graphs satisfying the following constraint:

Definition 1. Call an edge coloring of a bipartite graph *rich* if for each pair

$$\langle \text{left vertex } x,\ \text{right vertex } y\rangle$$

there is at most one color a touching both x and y (the latter means that there is an edge with color a incident to x and an edge, maybe a different one, with color a incident to y).

From Theorem 1 we can derive the following bound for rich colorings of bipartite graphs:

Corollary 1. *Assume that the degree of each left vertex in a given bipartite graph is at least L and the degree of each right vertex is at least R. Then the number of colors in every rich edge coloring of the graph is at least LR.*

Proof: Consider the uniform distribution on the set of edges of the graph. Denote by (A,X,Y) the following triple of jointly distributed random variables:
$X=$ [the left end of the edge],
$Y=$ [the right end of the edge],
$A=$ [the color of the edge].
As the coloring is rich, the triple (A,X,Y) satisfies (2): if both events $[A=a, X=x]$ and $[A=a, Y=y]$ have positive probabilities, then both x and y are touched by a, and hence such a is unique. Therefore by Theorem 1 we have $H(A|X)+H(A|Y)\leqslant H(A)$.

By construction, the distribution on the edges is uniform. Hence, for each vertex x, the conditional distribution of edges incident to this x is also uniform. All edges incident to one and the same vertex must have different colors. So for every fixed vertex x, all colors touching this x are equiprobable. In other words, conditional on $X=x$, the value of A is uniformly distributed on the set of colors touching x. Thus, $H(A|X)\geqslant \log L$. Similarly, we have $H(A|Y)\geqslant \log R$. By Theorem 1 we have $H(A)\geqslant \log L+\log R$. It follows that the range of A is at least LR. ■

Remark 1. In fact this proof gives a stronger result. Let us call by *the left* and *the right degrees* of an edge (in a bipartite graph) the degrees of its left and right ends. Denote by $\tilde L$ and $\tilde R$ the geometric means of the left and the right degrees of the graph's edges. That is, if the degrees of the vertices in the left part of the graph are $l_1, \ldots, l_n$ and the degrees of the vertices in right part of the graph are $r_1, \ldots, r_m$ ($l_1 + \ldots + l_n = r_1 + \cdots + r_m = e$, where e is the number of edges), then

$$\tilde{L} := \left(l_1^{l_1} \cdots l_n^{l_n}\right)^{1/e}, \ \tilde{R} := \left(r_1^{r_1} \cdots r_m^{r_m}\right)^{1/e}.$$

The proof of Corollary 1 explained above implies that the number of colors in every rich edge coloring of the graph is at least $\tilde{L}\tilde{R}$. Notice that $\tilde{L} \geqslant L$ and $\tilde{R} \geqslant R$ (these inequalities become equalities, if and only if the graph is uniform on the left or on the right respectively).

In what follows we exhibit three examples of rich colorings. The first two examples are pretty trivial; in the third example, Corollary 1 provides a non-trivial lower bound.

Example 1. For some bipartite graphs the lower bound LR proven in Corollary 1 is tight. Consider the simplest example: let $K_{R,L}$ be the complete bipartite graph with R left and L right vertices so that the degree of each left vertex is exactly L and the degree of each right vertex is exactly R (see in Fig. 1 an example for $R = 3$ and $L = 4$). This graph has LR edges. We may color them into LR colors, each edge having its unique color. This coloring is rich, as for each pair $\langle$left vertex x, right vertex $y\rangle$ only the color of that edge touches both x and y.

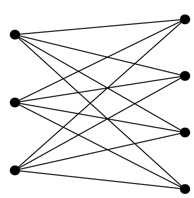

Fig. 1. Complete bipartite graph $K_{3,4}$.

For this example it is easy to compute directly the minimum number of colors in a rich coloring. Indeed, no different edges (x_1, y_1), (x_2, y_2) can share a color, as in that case the pair x_1, y_2 would violate the condition: the color of the edge (x_1, y_2) also touches both x_1 and y_2 and is different from the shared color of (x_1, y_1), (x_2, y_2).

Example 2. In general, the lower bound from Corollary 1 is not optimal. To construct the simplest example, consider the complete bipartite graph $K_{3,3}$ and delete from this graph three edges forming a perfect matching, see Fig. 2. In this graph the

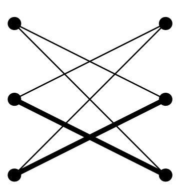

Fig. 2. Complete bipartite graph $K_{3,3}$ minus a perfect matching.

degree of each vertex is 2, so Corollary 1 claims that every rich coloring has at least $2 \cdot 2 = 4$ colors. However, it is easy to verify that the optimal rich coloring has 5 colors: two edges (e.g., the pair of edges shown in bold in Fig. 2) may share a color, while each of the remaining edge must have a unique color.

The next example is less obvious and exhibits a series of rich colorings for which Corollary 1 provides a non-trivial lower bound.

Example 3. *Assume that a finite family F of pair-wise disjoint squares inside the square $[0;1)^2$ in the Euclidean plane is given, each square having the form $[a;b) \times [c,d)$. Assume that all vertices of those squares have rational coordinates* [2]. *Assume further for each $x \in [0;1)$ there are at least L squares in F whose first projection includes x and similarly for each $y \in [0;1)$ there are at least R squares in F whose second projection includes y (see Fig. 3). Then $|F| \geqslant LR$.*

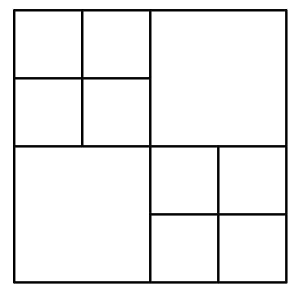

Fig. 3. The square is partitioned into 10 disjoint squares so that each vertical or horizontal line intersects 3 squares.

Proof: Obviously, there is a natural N such that each of the given squares has the form $[i/N; j/N) \times [k/N; l/N)$ for integer $i, j, k, l \leqslant N$. Consider the graph whose left and right nodes are rational numbers of the form i/N with $0 \leqslant i < N$. For each given square $[a; b) \times [c, d)$ consider the diagonal $\{(a{+}t, c{+}t) | 0 \leqslant t < b{-}a = c{-}d\}$, see Fig 4. The edges of the graph are those pairs (x, y) of nodes that lie on such diagonals. The edges of the resulting graph can be naturally colored into $|F|$ colors: the edges obtained from each diagonal are colored in a unique color. This coloring is rich: if the diagonal of a square has a point of the form $(x, *)$ and a point of the form $(*, y)$, then that square includes the point (x, y) and there is at most one such square (the squares are disjoint). The left degree of the graph is at least L. Indeed, for each $x_0 \in [0; 1)$ of the form i/N there are at least L squares whose first projection includes x_0 (see the dashed vertical in Fig 4). For each such square $[a; b) \times [c, d)$ its diagonal intersects the vertical segment $\{(x_0, y) | 0 \leqslant y < 1\}$, say, in the point (x_0, y_0). Since both a, c have the form i/N, the number $y_0 = c + (x_0 - a)$ also has this form and hence (x_0, y_0) is an edge of the graph. Similarly, the right degree of the graph is at least R (see the dashed horizontal line in Fig 4) and by Corollary 1 we have $|F| \geqslant LR$. ■

Historical remark: The study of edge colorings with additional constraints is by all means not new, see, e.g., strong edge colorings [5], complete edge colorings [3], the Thue number of a graph [13], etc. Some versions of constraint edge

[2] This assumption is added for technical simplicity and may be dropped.

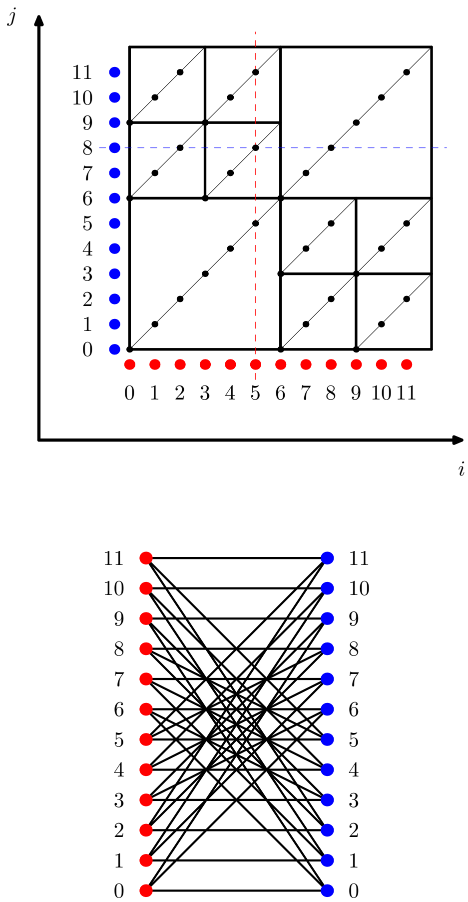

Fig. 4. Diagonals of the squares of the partition (points with rational coordinates $(i/N, j/N)$ for $N = 12$ and the corresponding bipartite graph).

coloring have found direct applications in Information theory (e.g., [7], [15]). The problem concerned in Corollary 1 looks quite natural in the context of the variety of edge coloring problems investigated in graph theory, though we are not aware of any earlier studies or applications of this specific variant of edge coloring.

B. A lower bound for the biclique cover number of bipartite graphs

Definition 2. For any bipartite graph $G = (V_1, V_2, E)$ (with the set of vertices $V_1 \cup V_2$ and the set of edges $E \subset V_1 \times V_2$) its *biclique cover number* $bcc(G)$ is defined as the minimal number of bicliques (complete bipartite subgraphs) that cover all edges of G.

Biclique coverings play an important role in communication complexity. Specifically, the *non-deterministic communication complexity* (see [14]) of a predicate

$$P : U \times U \to \{0, 1\}$$

can be defined as $\log bcc(G)$ for the bipartite graph $G = (V_1, V_2, E)$, where $V_1 = V_2 = U$, and E is the set of all pairs $(x, y) \in U \times U$ such that $P(x, y) = 1$.

Corollary 2. *Assume that the edges of a bipartite graph $G = (V_1, V_2, E)$ are colored in such a way that*

(*) *if edges (x, y') and (x', y) of the graph have the same color a, and the vertices x and y, as well as vertices x' and y', are also connected by edges, then the latter two edges also have color a.*

Assume further that a probability distribution over the edges of the graph is given. Denote by (X, Y, A) the random variables where

- $X =$ *[the left end of the edge],*
- $Y =$ *[the right end of the edge],*
- $A =$ *[the color of the edge].*

Then $bcc(G) \geqslant 2^{\frac{1}{2}(H(A|X)+H(A|Y)-H(A))}$.

Proof: Assume that this graph G can be covered by t bicliques $C_1, \dots, C_t$. Extend the distribution (X, Y, A) and add another random variable: we define Z as the index of a biclique C_i that covers the edge (X, Y). (If an edge belongs to several bicliques C_i, then we choose any of them.) Notice that Z ranges over $\{1, \dots, t\}$, so $H(Z) \leqslant \log t$.

The crucial point is that for a fixed value i of Z the condition (2) is satisfied. Indeed, assume that both $p(a, x|Z = i)$ and $p(a, y|Z = i)$ are positive. Then the biclique C_i has edges (x, y') and (x', y), both with color a. By property $(*)$ the color of the edge (x, y) also equals a and hence such a is unique. By Theorem 1 for each conditional distribution $(A, X, Y)|Z = i$ the inequality (1) holds. Hence we get

$$H(A|X, Z) + H(A|Y, Z) \leqslant H(A|Z).$$

It follows that

$$H(A|X) - H(Z) + H(A|Y) - H(Z) \leqslant H(A).$$

Thus, we obtain $t \geqslant 2^{H(Z)} \geqslant 2^{\frac{1}{2}[H(A|X)+H(A|Y)-H(A)]}$. ■

Example 4. Let us apply this corollary to a specific bipartite graph. Consider the bipartite Kneser graph $KG_{n,k} = (V_1, V_2, E)$, where both parts V_1 and V_2 consist of k-elements subsets of $\{1, \dots, n\}$, and the set of edges $E \subset V_1 \times V_2$ consists of all pairs of disjoint sets. Let us color the edge (x, y) in color $x \cup y$ and consider the uniform probability distribution over the edges of this graph. The condition $(*)$ is fulfilled. Indeed, assume we are given three pairs of disjoint k-element subsets: (x, y), (x, y') and (x', y). Assume further that $x \cup y' = x' \cup y = a$. Then $x = x'$ and $y = y'$ and hence $x \cup y = a$ as well. Hence

$$bcc(KG_{n,k}) \geqslant 2^{\frac{1}{2}[H(A|X)+H(A|Y)-H(A)]}.$$

We have $\binom{n}{2k}$ equiprobable colors and hence $H(A) = \log_2 \binom{n}{2k}$. On the other hand, $H(A|X) = H(A|Y) = \log_2 \binom{n-k}{k}$. Thus

$$bcc(KG_{n,k}) \geqslant \sqrt{\binom{n-k}{k}^2 \Big/ \binom{n}{2k}}.$$

If $n \gg k$ then $\binom{n-k}{k}^2/\binom{n}{2k}$ is close to $\binom{2k}{k} \approx 2^{2k}$ and we obtain a lower bound about 2^k for $bcc(KG_{n,k})$. On the other hand, it is known that $bcc(KG_{n,k}) \leqslant 2^{O(k+\log\log n)}$ (see [14, Section 2.3]), so in the case $\Omega(\log\log n) \leqslant k \ll n$ these lower and upper bounds are pretty close.

The proven bound in itself is of no interest; the simple and standard fooling set technique (see [14]) proves for this graph the bound $bcc(KG_{n,k}) \geqslant \binom{2k}{k}$ that holds for all $n \geqslant 2k$. However, this simple example illustrates the connection between biclique cover and conditional information inequalities. It remains unknown whether a similar technique can surpass the fooling set method for other examples of graphs.

Historical remark: In graph theory the minimum number of bicliques (complete bipartite subgraphs) needed to cover all edges of a given graph is known as *the biclique cover number* or *the bipartite dimension* of a graph. The problem of computing the bipartite dimension appears in different areas of computer science. In particular, the notions of bipartite partition and bipartite cover play the central role in communication complexity, [14].

The problem of determining the bipartite dimension is NP-hard even for bipartite graphs, [4]. A good approximation or a nontrivial lower bound for the bipartite dimension of some particular classes of graphs may imply substantial progress in various problems of computational complexity, see [14], [16], [19], [20]. In Corollary 2 we proposed a new technique of lower bounds for the bipartite dimension. Establishing formal relations between our method and previous approaches to biclique cover remains an open problem.

V. A GENERALIZATION OF A CONDITIONAL INEQUALITY FROM [21]

In this section we show that Theorem 1 implies some conditional version of Ingleton's inequality for entropies. So-called *Ingleton's inequality* was originally formulated and proven for ranks of linear subspaces, [2]. It turns out that a counterpart of this inequality reformulated in terms of Shannon's entropy (for random variables) has many nontrivial applications. Though in general this inequality is not valid for entropies (see [10]), it holds for distributions that satisfy some special properties (e.g., for random variables that enjoy the property of extracting the mutual information, or for variables with some properties of independence, see [6], [9], [12], [17], [18]). In particular, in [21] it was shown that Ingleton's inequality for entropies holds for all distributions where the entropies satisfy some linear constraints:

Theorem 2 ([21])**.** *If random variables X, Y, A, B satisfy the the constraints*

$$I(X : Y|A) = H(A|X, Y) = 0, \tag{4}$$

then Ingleton's inequality

$$I(A : B) \leqslant I(A : B|X) + I(A : B|Y) + I(X : Y) \tag{5}$$

holds for this distribution.

A noteworthy fact is that this result cannot be obtained as a direct implication of any unconditional linear inequality for Shannon's entropy. More precisely, whatever pair of reals λ_1, λ_2 we take, the inequality

$$\begin{aligned} I(A : B) \leqslant I(A : B|X) + I(A : B|Y) + I(X : Y)+ \\ +\lambda_1 I(X : Y|A) + \lambda_2 H(A|X, Y) \end{aligned}$$

does not hold for some distribution, see [25].

We claim that Ingleton's inequality holds also under condition (2), which is weaker than (4). Moreover, even a stronger inequality than Ingleton's inequality (namely the inequality (6) below), holds under condition (2).

Theorem 3. *(i) Ingleton's inequality* (5) *follows from the inequality*

$$H(A|X, B) + H(A|Y, B) \leqslant H(A|B). \tag{6}$$

(ii) Inequality (6) *holds for all random variables A, B, X, Y satisfying condition* (2)*.*

(iii) Condition (2) *is implied by condition* (4)*.*

In brief, Theorem 3 states that (4) $\Rightarrow$ (2) $\Rightarrow$ (6) $\Rightarrow$ (5). The main novelty is the middle implication (2) $\Rightarrow$ (6), while the implications (4) $\Rightarrow$ (2) and (6) $\Rightarrow$ (5) are almost straightforward (see the proof below) and the implication (4) $\Rightarrow$ (5) was known (Theorem 2).

Proof: (i) It is easy to verify that Ingleton's inequality (5) can be equivalently rewritten as

$$\begin{aligned} H(A|X, B) + H(A|Y, B) \leqslant H(A|B) + I(X : Y|A) \\ + H(A|X, Y). \end{aligned} \tag{7}$$

and hence follows from (6). Notice that under the constraints (4), Ingleton's inequality is equivalent to (6).

(ii) Note that inequality (6) is a relativized version of (1) (the word *relativization* here means that we insert a new condition in all entropy expressions). This similarity between inequalities (1) and (6) suggests that Theorem 2 can be deduced from Theorem 1. The key observation is that condition (2) is "relativizable": property (2) remains true if we restrict the initial probabilistic space to some subspace.

Lemma. *If a tuple of random variables (A, X, Y) satisfies* (2)*, then for each event $\mathcal{E}$ having positive probability the conditional random variables of $(A, X, Y)|\mathcal{E}$ satisfy* (2)*.*

Proof: Assume that the four probabilities

$$\Pr[X = x,\ A = a\ |\mathcal{E}], \Pr[Y = y,\ A = a\ |\mathcal{E}],$$
$$\Pr[X = x,\ A = a'|\mathcal{E}], \Pr[Y = y,\ A = a'|\mathcal{E}]$$

are positive. Then the unconditional probability of each of these events is positive as well and hence $a = a'$ by (2). ■

Now we can show that (6) follows from condition (2). Indeed, for every possible value b of B the lemma guarantees that (2) remains valid conditional on the event $B = b$. By Theorem 1 this implies

$$H(A|X, B = b) + H(A|Y, B = b) \leqslant H(A|B = b),$$

and taking the average over all values b we get (6).

(iii) Inequality $I(X : Y|A) = 0$ means that

$$p(a, x, y)p(a) = p(a, x)p(a, y)$$

for all triples a, x, y. Thus it implies that for each triple a, x, y of values of A, X, Y, if both probabilities $p(a, x)$ and $p(a, y)$ are positive, then $p(a, x, y)$ is also positive. Hence, if for some a, a', x, y all the four probabilities

$$p(a, x), p(a, y), p(a', x), p(a', y)$$

are positive (the assumption of (2)), then it follows that the probabilities $p(a, x, y)$ and $p(a', x, y)$ must be also positive.

Now we employ the condition $H(A|X, Y) = 0$ (which means that the value of A is a deterministic function of (X, Y)). If both probabilities $p(a, x, y)$ and $p(a', x, y)$ are positive, then $a = a'$, and we obtain the conclusion of (2). ■

Note that in general inequality (6) is stronger than Ingleton's inequality (5). For instance, let B be constant, let X, Y be independent uniformly distributed random bits, and let $A = X \oplus Y$. Then inequality (6) specializes to $1 + 1 \leqslant 1$ and hence is wrong, while Ingleton's inequality (5) specializes to $0 \leqslant 0 + 0 + 0$ (or to $1 + 1 \leqslant 1 + 1 + 0$, if written in the form (7)) and hence is true.

Note also that in general condition (2) is weaker than condition (4). For instance, let A be constant and let X, Y be any dependent random variables.

Acknowledgment

The authors are grateful to Maxim Popov for valuable comments on a preliminary version of this paper. The authors thank anonymous referees for many comments and suggestions.

References

[1] V. G. Vizing, "On an estimate of the chromatic class of a p-graph," *Disagree. Analiz.*, vol. 3, pp. 25–30, 1964.

[2] A. W. Ingleton, "Conditions for representability and transversability of matroids," In *Proc. Fr. Br. Conf* 1970, pp. 62–67. Springer-Verlag, 1971.

[3] J. Bosák and J. Nešetřil, "Complete and pseudocomplete colorings of a graph," *Matematicky asopis Slovenskej Akadémie Vied*, vol. 26, no. 3, pp. 171–184, 1976.

[4] J. Orlin, "Contentment in graph theory: covering graphs with cliques," *Indagationes Mathematicae*, vol. 80, no. 5, pp. 406–424, 1977.

[5] J.-L. Fouquet and J.-L. Jolivet, "Strong edge-colorings of graphs and applications to multi-k-gons," *Ars Combinatorica*, vol. 16 (A), pp. 141–150, 1983.

[6] Z. Zhang, R. W. Yeung, "A non-Shannon-type conditional information inequality," *IEEE Trans. Inf. Theory*, vol. 43, pp. 1982–1986, 1997.

[7] C. Kaklamanis, P. Persiano, T. Erlebach, and K. Jansen, "Constrained bipartite edge coloring with applications to wavelength routing," *In Proceedings of 24th International Colloquium on Automata, Languages, and Programming (ICALP 97), Bologna, Italy*, pp. 493–504, 1997.

[8] Z. Zhang, R. W. Yeung. "On characterization of entropy function via information inequalities," *IEEE Trans. Inf. Theory*, vol. 44, no. 4, pp 1440–1452, 1998.

[9] F. Matúš, "Conditional independences among four random variables III: Final conclusion," *Combinatorics, Probability and Computing*, vol. 8, pp. 269–276, 1999.

[10] D. Hammer, A. Romashchenko, A. Shen, N. Vereshchagin. "Inequalities for Shannon Entropy and Kolmogorov Complexity," *Journal of Computer and System Sciences*, vol. 60, pp. 442–464, 2000.

[11] R. W. Yeung, "A First Course in Information Theory," Norwood, MA: Kluwer, 2002.

[12] K. Makarychev, Yu. Makarychev, A. Romashchenko, N. Vereshchagin, "A New Class of non-Shannon-type Inequalities for Entropies," *Comm. Inform. Syst.*, vol. 2, no. 2, pp. 147–166, 2002.

[13] N. Alon, J. Grytczuk, M. Hałuszczak, O. Riordan, "Nonrepetitive colorings of graphs," *Random Structures & Algorithms.* vol. 21, no. 3–4, pp. 336–346, 2002.

[14] E. Kushilevitz, N. Nisan."Communication complexity," Cambridge University Press, 2006.

[15] C. L. Barrett, G. Istrate, V. S. A. Kumar, M. V. Marathe, S. Thite, and S. Thulasidasan, "Strong edge coloring for channel assignment in wireless radio networks," *Proc. Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, p. 106, 2006.

[16] S. Jukna, "On graph complexity," *Combinatorics, Probability and Computing*, vol. 15, no. 6, pp. 855–876, 2006.

[17] F. Matúš, "Infinitely many information inequalities," *Proc. IEEE Int. Symp. Inf. Theory*, Nice, France, pp. 41–44, June 2007.

[18] R. Dougherty, C. Freiling, K. Zeger, "Networks, matroids, and non-Shannon information inequalities," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 1949–1969, 2007.

[19] T. Lee and A. Shraibman, "Lower bounds in communication complexity," *Foundations and Trends in Theoretical Computer Science*, vol. 3, no. 4, pp. 263–399, 2009.

[20] S. Jukna and A. S. Kulikov, "On covering graphs by complete bipartite subgraphs." *Discrete Mathematics*, vol. 309, no. 10, pp. 3399–3403, 2009.

[21] T. Kaced, A. Romashchenko, "On essentially conditional information inequalities," *Proc. IEEE Int. Symp. Inf. Theory*, pp. 1935–1939, 2011.

[22] T. Chan. "Recent progresses in characterizing information inequalities," *Entropy*, vol. 13, no. 2, pp. 379–401, 2011.

[23] R. W. Yeung, "Facets of entropy," *IEEE Information Theory Society Newsletter*, vol. 62, no. 8, pp. 6–16, 2012.

[24] P. Formanowicz and K. Tanaś, "A survey of graph coloring-its types, methods and applications," *Foundations of Computing and Decision Sciences*, vol. 37, no. 3, pp. 223–238, 2012.

[25] T. Kaced, A. Romashchenko, "Conditional Information Inequalities for Entropic and Almost Entropic Points," *IEEE Trans. Inf. Theory*, vol. 59, issue 11, pp. 7149–7167, 2013.

Tarik Kaced received the B.Sc. in Fundamental Computer Science from École Normale Supérieure de Lyon in 2007, and the M.Sc. from Université de Nice Sophia-Antipolis in 2009. He completed the Ph.D. degree in Computer Science in 2012 at Université de Montpellier 2 in the ESCAPE team from the LIRMM.

He has been a post-doctoral fellow at the Chinese University of Hong Kong, at Université Paris-Est Créteil. He is now a post-doctoral fellow at Sorbonne Unversité. His research interests include information theory, Kolmogorov complexity, information inequalities, secret sharing and machine learning.

Andrei Romashchenko (M'11) graduated from the Dept. of Mathematical logic of Lomonosov Moscow State University in 1997 and received the Ph.D. degree from that university in 2001.

In 2001–2005 he was a Research Fellow at the Institute for Information Transmission Problems of the Russian Academy of Sciences; now he is an external member of this institute. Since 2005 he is a Research Fellow at the National Center for Scientific Research (CNRS), France. Now he is a CNRS researcher at the LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) and University of Montpellier. He is working in algorithmic information theory.

Nikolai Vereshchagin graduated from the Dept. of Mathematical logic of Lomonosov Moscow State University in 1981 and received the Ph.D. degree at that university in 1986. Since that time he has worked in the area of Computational complexity and Information theory.

Now he is a Professor in National Research University Higher School of Economics and Lomonosov Moscow State University. He is a member of Academia Europaea.