Hamming distance between partitions, clustering comparison and information

Giovanni Rossi

***Abstract*—Measuring the distance between partitions is useful for clustering comparison in different fields. For example, in bioinformatics the measuring mostly obtains through a *maximum matching* distance MMD, although this is algorithmically demanding and hardly fits certain instances. In fact, another distance measure is being tested, namely one based on information theory and termed *variation of information* VI. Alternatively, this paper proposes the *Hamming* distance HD, displaying large range and great measurement sensitivity, while also relying on a neat binary string representation of partitions. Novel distance HD is computationally handy and shares with VI important characterizing axioms. Developing from the combinatorial concern to translate the traditional Hamming distance from subset to partition lattices, HD constitutes a valuable computational tool for clustering and information processing where a distance between partitions is to be measured.**

***Index Terms*—Combinatorial problem, partition lattice, Hamming distance, clustering, information theory, bioinformatics.**

I. Introduction

A Partition or clustering of a (finite) set N divides this latter into disjoint subsets whose union yields the whole (partitioned) set N. There exists a massive literature on cluster analysis and clustering algorithms, which is important in a variety of fields at the interface of computer science, artificial intelligence and engineering, including pattern recognition/learning, web mining and bioinformatics. In particular, measuring the distance between partitions is useful for information processing in terms of comparisons between clustering results. For example, a local-search clustering algorithm generally provides different outputs for varying initial candidate solutions (or inputs), and it may be relevant to measure the distance between any two of them, or between each of them and a given one (see [22, section 1]). Specifically, in bioinformatics a great deal of attention is being paid to measuring the distance between clusterings of populations, either natural or experimental, for sibling relationship reconstruction. In practice, thus far the focus has been placed almost exclusively on a unique distance measure (see [17], [6], [31], [19], [18], [30], [2], [8], [5], [10], [3]), namely one relying on maximum matching and denoted MMD in the sequel. After its first appearance [4], this measure was subsequently shown [15] to be computable via the assignment problem (see [20, p. 236] and [12, subsection 4.1]). In this view, a notable exception relies on using an alternative distance measure [9], called *variation of information* VI and derived from information theory [22].

The present paper provides a further solution to the problem of measuring a distance between partitions. This proposed measure is of the Hamming type (see section 2), indeed obtained by translating the traditional Hamming distance between subsets [7, p. 3] in terms of partitions. As long as lattice theory is concerned, the result is thus combinatorially consistent. Furthermore, it shares with VI important characterizing axioms. On the other hand, it also displays two features that appear interesting for biologists and practitioners in general. Firstly, as the Hamming distance HD between partitions provided here obtains simply by means of scalar products between (input) vectors, no proper computational issue arises, while determining MMD and VI is algorithmically demanding (see [15], [19], [18] and [12, chapter 4]). Secondly, HD has a large range and hence great measurement sensitivity.

Giovanni Rossi is with the Department of Informatica Scienza Ingegneria (DISI), University of Bologna, via Mura Anteo Zamboni 7, Bologna 40127, Italy; e-mail: giovanni.rossi6@unibo.it

II. Hamming and partition distances

FOR a finite set $N = \{1, \ldots, n\}$, let $(2^N, \cap, \cup)$ and $(\mathcal{P}^N, \wedge, \vee)$ denote the corresponding subset and partition lattices, with inclusion $\supseteq$ and coarsening $\geqslant$ as order relations, respectively. Both are atomic (and atomistic), but the fomer is distributive while the latter is geometric [1], [32]. For populations or data sets, n is the number of individuals or data points being partitioned.

The distance between elements of a ordered set is to be measured in terms of the order relation. Also, measures of the distance between elements of any set are called *Hamming distances* when these elements are represented as arrays and the distance between two of them is the number of entries where their array representations differ. The Hamming distance $d(A, B)$ between two subsets $A, B \in 2^N$ is $d(A, B) =$

$$= |A \Delta B| = |A \backslash B| + |B \backslash A| = r(A \cup B) - r(A \cap B), \quad (1)$$

$r : 2^N \to \mathbb{Z}_+$ being the rank function: $r(A') = |A'|$ for all $A' \in 2^N$. In words, it counts how many $i \in N$ are included in either A or else B, but not in both. Elements $i \in N$, when regarded as 1-cardinal subsets $\{i\} \in 2^N$, are *atoms* in subset lattice $(2^N, \cap, \cup)$. Expression (1) defines indeed a Hamming distance in that subsets $A \in 2^N$ are firstly represented as binary n-vectors $\chi_A \in \{0, 1\}^n$ (or vertexes of the n-dimensional unit hypercube $[0, 1]^n$) through their *characteristic function* $\chi_A : N \to \{0, 1\}$, defined by $\chi_A(i) = 1$ if $i \in A$ and $\chi_A(i) = 0$ if $i \in N \backslash A$. Next, the distance between any two of them $A, B \in 2^N$ is the number of entries where χ_A and χ_B differ. That is, the cardinality of their symmetric difference $A \Delta B$ [7], [1].

A partition $P = \{A_1, \ldots, A_{|P|}\} \subset 2^N$ of N is a collection of pairwise disjoint subsets, called blocks (or clusters), whose union yields N. Any subset $A \in 2^N$ has a unique complement $A^c = N \backslash A$. For all partitions $P \in \mathcal{P}^N$ and all non-empty

ISBN: 978-1-61804-323-8

subsets $\emptyset \subset A \subseteq N$, denote the partition of A induced by P by $P^A = \{B \cap A : B \in P, \emptyset \neq B \cap A\}$. The maximum matching distance $MMD : \mathcal{P}^N \times \mathcal{P}^N \to \{0, 1, \ldots, n-1\}$ between any two partitions P, Q is

$$MMD(P, Q) = \min\{|A^c| : \emptyset \subset A \subseteq N, P^A = Q^A\}. \quad (2)$$

This is the minimum number of elements $i \in N$ that must be deleted in order for the two residual induced partitions to coincide. Also, $MMD(P, Q)$ *is the minimum number of elements that must be moved between clusters of* P *so that the resulting partition equals* Q ([15, p. 160]). It is computable as a maximum matching (or assignment) problem. Firstly recall that in a (simple) graph a matching is a set of pairwise disjoint edges (i.e. the endpoints are all different vertexes). Now consider the bipartite graph $G = (P \dot\cup Q, E)$ with $|P| + |Q|$ vertexes ($\dot\cup$ is the union of disjoint sets), one for each block of each partition, and join any two of them $A \in P$ and $B \in Q$ with an edge $(A, B) \in E$ if $A \cap B \neq \emptyset$. In addition, let $|A \cap B|$ be the weight of the edge. Then, determining $MMD(P, Q)$ amounts to find a maximum weighted matching E^* in G, that is one where the sum $\sum_{(A,B)\in E^*} |A \cap B|$ of edge weights is maximal. In fact, the (minimal) number $MMD(P, Q)$ of elements that must be removed for the two residual partitions to coincide is the sum $\sum_{(A,B)\in E^*} |A \Delta B|$ over all selected edges of the cardinality of the symmetric difference between the associated endpoints.

Another important measure of the distance between two partitions P, Q is the variation of information $VI(P, Q)$. Firstly obtained axiomatically from information theory [22], this distance is now useful in bioinformatics as well [9]. Entropy $H(P) = -\sum_{A\in P} \frac{|A|}{n} \log\left(\frac{|A|}{n}\right)$ of a partition P and *mutual information* $I(P, Q) = \sum_{A\in P, B\in Q} \frac{|A\cap B|}{n} \log\left(\frac{n|A\cap B|}{|A||B|}\right)$ between P and Q (binary logarithm) enable to measure the distance between these latter as variation of information

$$VI(P, Q) = H(P) + H(Q) - 2I(P, Q). \quad (3)$$

While MMD ranges over all integer values $\{0, 1, \ldots, n-1\}$ in $[0, n-1] \subset \mathbb{R}_+$, VI ranges in $[0, \log n]$.

Apart from MMD and VI, there exist several other partition distance measures (see [13, sections 10.2 and 10.3, pp. 191-193] and [11], [16], [33], [23]). One was proposed as the *Hamming distance between (matrices representing) partitions* [22], [24], [25], and thus shall be briefly distinguished from the object of this paper. A binary relation R on N is a subset $R \subseteq N \times N$ of ordered pairs (i, j) of elements $i, j \in N$. The collection of all such binary relations is subset lattice $(2^{N\times N}, \cap, \cup)$. If symmetry $(i, j) \in R \Rightarrow (j, i) \in R$ and *transitivity* $(i, j), (j, h) \in R \Rightarrow (i, h) \in R$ hold, then R is an *equivalence* relation, or a partition of N into equivalence classes: maximal subsets $A \in 2^N$ such that $(i, j), (j, i) \in R$ for all $i, j \in A$ are precisely its blocks. A binary relation R may be represented as a $n \times n$ matrix $M^R \in \{0, 1\}^{n\times n}$ with entries $M^R_{ij} = 1$ if $(i, j) \in R$ and $M^R_{ij} = 0$ if $(i, j) \notin R$. Now let two equivalence relations R, R' have associated partitions P, P' and representing matrices $M^R, M^{R'}$. The distance $d(R, R')$ between subsets $R, R' \in 2^{N\times N}$ can be computed through expression (1) above: $d(R, R') = |R\Delta R'| = |R \cup R'| - |R \cap R'|$. While providing a distance $\delta(P, P')$ between partitions, this factually is a Hamming distance between subsets. In particular, $|R\Delta R'|$ is the number of 1s in matrix $M^{R\Delta R'} = M^R + M^{R'}$ modulo 2 (see [1, p. 338]). The point is that $2^{N\times N}$ contains many lattice elements (or binary relations) that do not correspond to partitions (or equivalence relations).

Roughly speaking, partition lattice $(\mathcal{P}^N, \wedge, \vee)$ is compressed into a larger subset lattice, with which some elements are shared while some others are not. Apart from binary relations just described, another example comes from noticing that partitions P are collections of subsets, i.e. $P \in 2^{2^N}$, and thus the distance between P and Q may be computed as the Hamming distance $|P\Delta Q|$ between elements of subset lattice $(2^{2^N}, \cap, \cup)$, i.e. the number of subsets $A \in 2^N$ that are blocks of either one but not both. Again, there are many set systems (or collections $S \in 2^{2^N}$ of subsets) that do not correspond to partitions. This feature is maintained even when P and Q are represented as joins of atoms, for they generally admit several such representations. The Hamming distance HD between partitions proposed below relies on representing partitions as $\{0, 1\}$-valued and $\binom{n}{2}$-dimensional arrays, because $\binom{n}{2}$ is the number of atoms. But not all $2^{\binom{n}{2}}$ such arrays correspond to partitions.

III. Atoms, rank and size

The rank function $r : \mathcal{P}^N \to \mathbb{Z}_+$ for the partition lattice is $r(P) = n - |P|$, that is, the maximum conceivable number of blocks for a partition of a n-set (which is n, of course) minus the actual number of blocks of any partition P. The unique partition $P_\perp = \{\{1\}, \ldots, \{n\}\}$ with rank $r(P_\perp) = 0$ is the bottom element. By definition, atoms are immediately above, with rank 1. That is, atoms populate level 1 (above level 0) of the associated *Hasse diagram* (see [1], [32], [22]), with coarsening $\geqslant$ as the order relation (i.e. coarser partitions in upper levels). This means that an atom, in the partition lattice, is any P with $n-1$ blocks, out of which $n-2$ must be singletons, while the remaining one is a pair, and there are $\binom{n}{2}$ unordered pairs.

For notational convenience, let $[ij] \in \mathcal{P}^N$ denote the atom where the unique 2-cardinal block is (unordered) pair $\{i, j\} \in [ij]$, noting that $2^{\binom{n}{2}}$ is the number of simple graphs with n vertexes. In fact, partitions are the *transitive closure* of such graphs: the former obtain from the latter by adding all edges needed to render complete (or fully connected) each component [1, Point 2.31, p. 54]. The order relation among partitions is coarsening $\geqslant$, where $P \geqslant Q$ means that every block of Q is included in some block of P. Let $\mathbf{1} \in \{0, 1\}^n$ denote the n-vector with all entries equal to 1 and $\langle x, y\rangle$ be the scalar product of any two vectors x, y. For any subset $A \in 2^N$, clearly $|A| = r(A) = \langle \chi_A, \mathbf{1}\rangle$. Also, given any $A, B \in 2^N$,

$$d(A, B) = |A\Delta B| = |A| + |B| - 2|A \cap B| = \quad (4)$$

$$= \langle \chi_A, \mathbf{1}\rangle + \langle \chi_B, \mathbf{1}\rangle - 2\langle \chi_A, \chi_B\rangle. \quad (5)$$

Let $\mathcal{P}^N_{(1)} = \{[ij] : 1 \leq i < j \leq n\}$ denote the $\binom{n}{2}$-set of atoms of the partition lattice. The analog of the characteristic

function χ_A of subsets $A \in 2^N$ for partitions $P \in \mathcal{P}^N$ is the *indicator function* $I_P : \mathcal{P}^N_{(1)} \to \{0,1\}$ defined by

$$I_P([ij]) = \begin{cases} 1 \text{ if } P \geqslant [ij] \\ 0 \text{ if } P \not\geqslant [ij] \end{cases} \quad \text{for all } P \in \mathcal{P}^N, [ij] \in \mathcal{P}^N_{(1)}.$$

In words, if pair $\{i,j\}$ is included in some block A of P (i.e. $\{i,j\} \subseteq A \in P$), then partition P is coarser than atom $[ij]$, and the corresponding position $I_P([ij])$ of indicator array I_P has entry 1. Otherwise, that position is 0.

By reversing the order relation, $P \geqslant [ij]$ turns into $[ij] \leqslant P$, i.e. atom $[ij]$ is finer than P. The number of atoms finer than $P = \{A_1, \ldots, A_{|P|}\}$ is the *size* $s : \mathcal{P}^N \to \mathbb{Z}_+$ (see [27]), i.e.

$$s^P = \sum_{1 \leq k \leq |P|} \binom{|A_k|}{2} = \langle I_P, \mathbf{1} \rangle, \tag{6}$$

with $s^P = s(P)$ and $\mathbf{1} \in \{0,1\}^{\binom{n}{2}}$ now denoting the unitary $\binom{n}{2}$-vector. That is to say, $s^P = |\{[ij] \in \mathcal{P}^N_{(1)} : [ij] \leqslant P\}|$.

TABLE I
Available sizes of partitions of n-sets, $1 \leq n \leq 7$.

$\lvert N\rvert = n$	$\{s^P : P \in \mathcal{P}^N\}$ Available sizes
1	$\{0\}$
2	$\{0,1\}$
3	$\{0,1,3\}$
4	$\{0,1,2,3,6\}$
5	$\{0,1,2,3,4,6,10\}$
6	$\{0,1,2,3,4,6,7,10,15\}$
7	$\{0,1,2,3,4,5,6,7,9,10,11,15,21\}$

While the cardinality $|A|$ of subsets $A \in 2^N$ takes every integer value between 0 and n, the size $s^P = \langle I_P, \mathbf{1} \rangle$ of partitions $P \in \mathcal{P}^N$ does not the same between 0 and $\binom{n}{2}$. In fact, a main difference between characteristic function χ_A of subsets $A \in 2^N$ and indicator function I_P of partitions $P \in \mathcal{P}^N$, is that the former spans the whole vertex set of the unit n-dimensional hypercube (that is $\{\chi_A : A \in 2^N\} = \{0,1\}^n$), while the latter only spans a (rather small) proper subset of the vertex set of the unit $\binom{n}{2}$-dimensional hypercube (that is $\{I_P : P \in \mathcal{P}^N\} \subset \{0,1\}^{\binom{n}{2}}$, where $\subset$ denotes strict inclusion $\subsetneq$; again, see [1, Point 2.31, p. 54]). This is due to *linear dependence* [34], characterizing geometric lattices in general. The number of non-spanned vertexes is $2^{\binom{n}{2}} - \mathcal{B}_n$, where $\mathcal{B}_n = |\mathcal{P}^N|$ is the n-th *Bell number* or number of partitions of a n-set [29], [14] [18, Section 2.2]. The smallest value of n where linear dependence may be appreciated is $n = 3$, in that there are $\mathcal{B}_3 = 5$ partitions, namely the finest $\{\{1\},\{2\},\{3\}\}$ and coarsest $\{\{1,2,3\}\}$ ones, and $\binom{3}{2} = 3$ atoms: $[12] = \{\{1,2\},\{3\}\}$, $[13] = \{\{1,3\},\{2\}\}$ and $[23] = \{\{2,3\},\{1\}\}$. Thus, there is no partition with size equal to 2, as $[12] \vee [23] = [12] \vee [13] = [13] \vee [23] = \{\{1,2,3\}\} = [12] \vee [13] \vee [23]$, where for any two partitions $P, Q \in \mathcal{P}^N$, the join $P \vee Q$ is the finest partition coarser than both P and Q, while the meet $P \wedge Q$ is the coarsest partition finer than both P and Q (see [1], [32]). Available sizes for $1 \leq n \leq 7$ are in Table I above.

A. Representations as joins of atoms: rank and size

In atomic (or atomistic, see [1], [32]) lattices every element admits some representation as a join of atoms. Yet, while subsets $A \in 2^N$ admit a unique such a representation, namely $A = \bigcup_{i \in A} \{i\}$, partitions generally admit several such representations. Again, this results from linear dependence and is observable already when $n = 3$ (see above), in that the coarsest partition $\{\{1,2,3\}\}$ may be represented either as the join of any two atoms, or even as the join of all the three available atoms at once. In particular, the rank $r(P) = n - |P|$ of any partition P is the minimum number of atoms needed for representing P as a join of atoms, while the size $s^P = \sum_{A \in P} \binom{|A|}{2}$ is the maximum number of atoms available for representing P as a join of atoms. Hence, keeping the coarsest partition of a 3-cardinal set as the simplest example, the rank is $r(\{\{1,2,3\}\}) = 3 - 1 = 2$, while the size is $s^{\{\{1,2,3\}\}} = 3 = \binom{3}{2}$.

The rank or cardinality of subsets is a peculiar lattice function, namely a *valuation* of subset lattice $(2^N, \cap, \cup)$ [1, p. 187]), meaning that $|A \cup B| + |A \cap B| = |A| + |B|$ for all $A, B \in 2^N$. On the other hand, neither the rank nor the size of partitions are valuations of partition lattice $(\mathcal{P}^N, \wedge, \vee)$. In fact, valuations of the partition lattice are constant (partition functions), taking the same value on each and every partition [1, Exercise 12 (ii) p. 195]. The following definitions enable to appreciate the features displayed by the rank and the size of partitions, regarded as lattice functions $r, s : \mathcal{P}^N \to \mathbb{Z}_+$.

For any two partitions $P, Q \in \mathcal{P}^N$, the strict coarsening relation $>$ is $P > Q$, meaning $P \geqslant Q, P \neq Q$. The rank is easily seen to be a strictly monotone partition function, that is to say $P > Q$ entails $r(P) > r(Q)$. The same is established hereafter for the size.

Proposition 1: The size is a strictly monotone partition function:

$$s^P > s^Q \text{ for all } P, Q \in \mathcal{P}^N \text{ such that } P > Q.$$

Proof: If $P > Q$, then every block $A \in P$ is the union of some blocks $B_1, \ldots, B_{|Q^A|} \in Q$, with[1] $|Q^A| > 1$ for at least one block $A \in P$. The union of any two such blocks $B, B' \in Q$ increases the size (toward s^P, i.e. while reaching an higher level of the Hasse diagram toward P) by

$$\binom{|B| + |B'|}{2} - \left(\binom{|B|}{2} + \binom{|B'|}{2} \right) = |B||B'|,$$

which is strictly positive as blocks are non-empty. ■

If $f(P) + f(Q) \geq f(P \vee Q) + f(P \wedge Q)$ for all $P, Q \in \mathcal{P}^N$, then partition function $f : \mathcal{P}^N \to \mathbb{R}$ is sub-modular. Super-modularity obtains by reversing the inequality. As partitions are the transitive closure of graphs, partition lattice $\mathcal{P}^N$ is the *polygon matroid* (see again [1, pp. 54 and 274-5]), and matroids have sub-modular rank [1, Rank axioms 6.14 p. 265]. Conversely, the size is a super-modular partition function. To see this, firstly consider the *covering* relation $>^*$, with $P >^* Q$ meaning: $P > Q$ and there is no P' such that $P > P' > Q$.

[1]Recall that Q^A is the partition of $A \in 2^N$ induced by $Q \in \mathcal{P}^N$, see section 2 above.

Proposition 2: The size is a super-modular partition function:

$$s^{P\vee Q} + s^{P\wedge Q} \geq s^P + s^Q \text{ for all } P, Q \in \mathcal{P}^N.$$

Proof: If the two partitions are comparable, say $P \geqslant Q$, then $P = P \vee Q$ and $Q = P \wedge Q$, which makes the statement satisfied with equality. Otherwise, if $P \ngeqslant Q \ngeqslant P$, then there are two maximal chains of partitions (see [1], [32]), one meets $P\wedge Q$ and P as well as $P\vee Q$, while the other meets $P\wedge Q$ and Q as well as $P\vee Q$. Focusing on the relevant part or segment of the former chain, there are $\hat{P}_{\hat{r}} >^* \cdots >^* \hat{P}_1 >^* \hat{P}_0$, with $\hat{r} = r(P\vee Q) - r(P\wedge Q)$, such that $\hat{P}_0 = P\wedge Q$ and $\hat{P}_{\hat{r}} = P\vee Q$ as well as $\hat{P}_{k_P} = P$ for some $k_P, 0 < k_P < \hat{r}$. Similarly, focusing on the relevant segment of the latter chain[2], there are $\hat{Q}_{\hat{r}} >^* \cdots >^* \hat{Q}_1 >^* \hat{Q}_0$ such that $\hat{Q}_0 = P \wedge Q$ and $\hat{Q}_{\hat{r}} = P \vee Q$ as well as $\hat{Q}_{k_Q} = Q$ for some $k_Q, 0 < k_Q < \hat{r}$. Note that $r(P) = r(Q) \Leftrightarrow k_P = k_Q$.

Difference $s^{P\vee Q} + s^{P\wedge Q} - (s^P + s^Q)$ may be counted by summing across (Hasse diagram) levels for the two segments. The fact is that frequently many atoms are both *(a)* finer than $P \vee Q$, and *(b)* $\geqslant$-incomparable with respect to both P and Q. Atoms $[ij] \leqslant P \wedge Q$ may be disregarded because they are counted in the size of all the four involved partitions, namely $P, Q, P\wedge Q, P\vee Q$. As for the remaining ones, firstly observe that sets

$$S_P = \{[ij] \in \mathcal{P}_1^N : P \geqslant [ij] \nleqslant P \wedge Q\} \text{ and}$$

$$S_Q = \{[ij] \in \mathcal{P}_1^N : Q \geqslant [ij] \nleqslant P \wedge Q\}$$

are disjoint, i.e. have empty intersection $S_P \cap S_Q = \emptyset$. In fact, if an atom $[ij] \nleqslant P \wedge Q$ satisfied $P \geqslant [ij] \leqslant Q$, then $(P\wedge Q) \vee [ij]$ (and not $P \wedge Q$) would be the coarsest partition finer than both P, Q, i.e. a contradiction.

Now consider going upwards through the Hasse diagram from $P \wedge Q$ to $P \vee Q$ *twice*, starting with all atoms finer than $P \vee Q$ apart from those also finer than $P \wedge Q$. Specifically, a first route is through segment $\hat{P}_0, \ldots, \hat{P}_{\hat{r}}$ of the former maximal chain, and at each partition reached up to $\hat{P}_{k_P} = P$ all atoms finer than the current partition but not also finer than the preceding one are removed. A second, subsequent route starts with only the *residual* atoms and is through segment $\hat{Q}_0, \ldots, \hat{Q}_{\hat{r}}$ of the latter maximal chain. Again, up to $\hat{Q}_{k_Q} = Q$ at each reached level all atoms finer than the current partition but not also finer than the preceding are removed. The above intersection being empty, it is not possible that the same atom is found to be removed twice, and at the end of the second route there still remains a non-empty (generally large) collection of atoms, namely all those for reaching $P\vee Q$ from either P or Q. ■

IV. Hamming distance between partitions

In order to parallel the traditional Hamming distance between subsets given by (1) above, HD has to count the number of atoms finer than either one of any two partitions but not finer than both. Thus, in terms of cardinalities of subsets of atoms, distance $HD : \mathcal{P}^N \times \mathcal{P}^N \to \mathbb{Z}_+$ is given, for all $P, Q \in \mathcal{P}^N$, by

$$\begin{aligned} HD(P,Q) &= |\{[ij] \in \mathcal{P}_{(1)}^N : P \geqslant [ij] \nleqslant Q\}| + \\ &\quad + |\{[ij] \in \mathcal{P}_{(1)}^N : P \ngeqslant [ij] \leqslant Q\}| \qquad (7) \\ &= |\{[ij] \in \mathcal{P}_{(1)}^N : P \geqslant [ij] \text{ OR } Q \geqslant [ij]\}| + \\ &\quad - |\{[ij] \in \mathcal{P}_{(1)}^N : P \geqslant [ij] \leqslant Q\}|, \qquad (8) \end{aligned}$$

where OR in second expression (8) means precisely that the corresponding subset contains all atoms finer than *at least one* of the two partitions, while the other (second) subset clearly contains all atoms finer than both.

The size and the indicator functions allow to reproduce these expressions in a computationally efficient manner. Considering the size first, note that

$$P \wedge Q = \bigvee_{\substack{[ij] \in \mathcal{P}_{(1)}^N \\ P \geqslant [ij] \leqslant Q}} [ij], \qquad (9)$$

and this is the maximal representation of $P \wedge Q$ as a join of atoms, namely that utilizing $s^{P\wedge Q}$ atoms. Accordingly, a further expression of HD obtains immediately as follows

$$HD(P,Q) = s^P + s^Q - 2s^{P\wedge Q} \text{ for all } P, Q \in \mathcal{P}^N. \qquad (10)$$

In words, this subtracts twice the number $s^{P\wedge Q}$ of atoms finer than both P and Q from the sum of the number s^P of atoms finer than P and the number s^Q of atoms finer than Q.

Remark 3: In proposition 2 above, subtracting $2s^{P\wedge Q}$ from both sides yields

$$s^{P\vee Q} - s^{P\wedge Q} \geq s^P + s^Q - 2s^{P\wedge Q}.$$

If the two partitions are comparable (say $P \geqslant Q$), then equality holds, but the converse is not true: there exist incomparable partitions P, Q, that is $P \ngeqslant Q \ngeqslant P$, where equality holds as well. For example, $N = \{1, 2, 3, 4\}$ and $P = [12], Q = [34]$. Also, the left hand side is the *size-based* distance [28], classifiable as a *modular* partition distance measure.

Turning to the indicator function, there are two distinct ways of using it for the current computational needs. Let $\mathbf{1}$ denote the $\binom{n}{2}$-vector each of whose entries equals 1, as before. Then,

$$\begin{aligned} HD(P,Q) &= \sum_{[ij] \in \mathcal{P}_{(1)}^N} \left(I_P([ij]) - I_Q([ij])\right)^2 = \\ &= \langle (I_P - I_Q)^2, \mathbf{1} \rangle \qquad (11) \\ &= \langle I_P, \mathbf{1} \rangle + \langle I_Q, \mathbf{1} \rangle - 2\langle I_P, I_Q \rangle, \qquad (12) \end{aligned}$$

where $\langle I_P, I_Q \rangle = \langle I_{P\wedge Q}, \mathbf{1} \rangle$ (see expression (9) above), hence this is the analogue of expression (5) above.

From another perspective, for all $[ij] \in \mathcal{P}_{(1)}^N$ define

$$\begin{aligned} I_{P,Q}^{\max}([ij]) &= \max\{I_P([ij]), I_Q([ij])\}, \\ I_{P,Q}^{\min}([ij]) &= \min\{I_P([ij]), I_Q([ij])\}. \end{aligned}$$

Then, in terms of the indicator function, expression (8) yields

$$HD(P,Q) = \langle I_{P,Q}^{\max}, \mathbf{1} \rangle - \langle I_P, I_Q \rangle = \langle I_{P,Q}^{\max} - I_{P,Q}^{\min}, \mathbf{1} \rangle =$$

$$= \langle I_P + I_Q \text{ modulo } 2, \mathbf{1} \rangle.$$

[2]The length $\hat{r}$ is the same for both the two segments.

Remark 4: While $I^{\min}_{P,Q} = I_{P\wedge Q}$, in general both $\langle I^{\max}_{P,Q}, \mathbf{1}\rangle$ and $\langle I_P + I_Q$ modulo 2 $, \mathbf{1}\rangle$ may well fail to be available sizes (see section 3 above), as

$$s^{P\vee Q} = \langle I_{P\vee Q}, \mathbf{1}\rangle \geq \langle I^{\max}_{P,Q}, \mathbf{1}\rangle =$$

$$= |\{[ij] \in \mathcal{P}^N_{(1)} : P \geqslant [ij] \text{ OR } Q \geqslant [ij]\}|.$$

Attention is now placed on comparing HD with both MMD and VI described above. While VI displays a precise axiomatic characterization [22], its applicative possibilities and computational complexity are not yet extensively investigated [12]. Conversely, MMD is widely used in bioinformatics applications and tightly bounded in terms of required computations [15], [19], [18].

V. HD AND VI: AXIOMS

PARTITION distance measures HD and VI display very similar axiomatic behaviors [22], as detailed hereafter.

Proposition 5: HD is a *metric*: for all $P, P', Q \in \mathcal{P}^N$,

1) $HD(P,Q) \geq 0$, with equality if and only if $P = Q$,
2) $HD(P,Q) = HD(Q,P)$,
3) $HD(P,P') + HD(P',Q) \geq HD(P,Q)$.

Proof: The second condition is obvious. The first one is also immediate as $\min\{s^P, s^Q\} \geq s^{P\wedge Q}$. Concerning the third one, known as triangle inequality,

$$HD(P,P') + HD(P',Q) - HD(P,Q) =$$

$= 2(s^{P'} - s^{P\wedge P'} - s^{P'\wedge Q} + s^{P\wedge Q})$. It must be shown that this is positive for all $P, P', Q \in \mathcal{P}^N$. For any such a triplet, the quantity is minimized when $s^{P\wedge P'} + s^{P'\wedge Q}$ is maximized, which in turn occurs for $P' \leqslant P, Q$ or $P \wedge P' = P' = P' \wedge Q$. This entails $s^{P'} = s^{P\wedge P'} = s^{P'\wedge Q} \leq s^{P\wedge Q}$. ■

Triangle inequality is satisfied with equality by both HD and VI when $P' = P \wedge Q$.

Proposition 6: HD satisfies *horizontal collinearity*:

$$HD(P, P\wedge Q) + HD(P\wedge Q, Q) = HD(P,Q)$$

for all $P, Q \in \mathcal{P}^N$.

Proof: Observe that summing $HD(P, P\wedge Q) = s^P - s^{P\wedge Q}$ and $HD(P\wedge Q, Q) = s^Q - s^{P\wedge Q}$ yields $HD(P,Q)$. ■

A further axiom characterizing both HD and VI is expressed in terms of those two partitions $P_\perp = \{\{1\}, \ldots, \{n\}\}$ and $P^\top = \{\{N\}\}$ occupying the bottom and top levels of the Hasse diagram, with $r(P^\top) = n - 1$.

Proposition 7: HD satisfies *vertical collinearity*:

$HD(P_\perp, P) + HD(P, P^\top) = HD(P_\perp, P^\top)$ for all $P \in \mathcal{P}^N$.

Proof: Sum $HD(P_\perp, P) + HD(P, P^\top) = s^P + s^{P^\top} - s^P$ yields $s^{P^\top} = \binom{n}{2} = HD(P_\perp, P^\top)$. ■

Two partitions P, Q are each a complement of the other if $P \wedge Q = P_\perp$ and $P \vee Q = P^\top$. While every subset $A \in 2^N$ admits a unique complement $A^c = N \backslash A$, a partition generally admits several complements. In particular, the Hamming distance between subsets given by expression (1) above yields $d(A, A^c) = n = d(\emptyset, N)$. That is, the distance between any two complements is the same as the distance between the bottom and top elements. Analogously, a partition distance measure $\delta : \mathcal{P}^N \times \mathcal{P}^N \to \mathbb{R}_+$ satisfies such a *complement maximality* condition [28] if for any two complements P', Q', it holds $\delta(P', Q') = \max_{(P,Q)\in\mathcal{P}^N\times\mathcal{P}^N} \delta(P,Q) = \delta(P_\perp, P^\top)$.

For generic $P = \{A_1, \ldots, A_{|P|}\}$, consider P_* with size $s^{P_*} = |P| - 1$ and maximal representation as a join of atoms (see above) $P_* = [ij]_1 \vee \cdots \vee [ij]_{|P|-1}$. Also, for $1 \leq k < |P|$, let $P \vee [ij]_k = \{A_1, \ldots, A_{k-1}, A_k \cup A_{k+1}, A_{k+2}, \ldots, A_{|P|}\}$ (assume $|A| \geq 2$ for all $A \in P$). In words, P_* is the join of $|P| - 1$ pair wise disjoint atoms each merging two distinct blocks of P. Such a P_* is a complement of P. A further complement of P is $P^* = B \cup P_\perp^{B^c}$ with $|B \cap A| = 1$ for every block $A \in P$. This partition P^* has $n - |P| + 1$ blocks: one is B containing $|P|$ elements, while all remaining ones are singletons. A sensible partition distance measure should distinguish between such complements. That is, it should *not* satisfy complement maximality. Indeed, this is the behavior displayed by all distances MMD, VI and HD, and is easily checked by means of an example: the partitioned set is $N = \{1, \ldots, 7\}$ while partitions are $P = 123|456|7$ and $P_* = 1|2|34|5|67$ as well as $P^* = 147|2|3|5|6$, with $|$ separating blocks. Then, $VI(P, P_*) = \frac{6}{7}\log 6 - \frac{2}{7} < \frac{4}{7}\log 9 + \frac{2}{7}\log 3 - \frac{1}{7} = VI(P, P^*)$ as well as $HD(P, P_*) = 8 < 9 = HD(P, P^*)$. On the other hand, $MMD(P, P_*) = 4 > 2 = MMD(P, P^*)$.

VI. HD AND MMD: COMPUTATIONS AND EXAMPLES

A Computational procedure determining HD accepts as inputs two $\binom{n}{2}$-dimensional and $\{0,1\}$-valued arrays as inputs. These are the indicator functions I_P, I_Q of the two partitions P, Q between which the distance is to be computed. In fact, the indicator function alone constitutes a novel, neat and mostly natural way of coding partitions (while how to code partitions as inputs when computing MMD is not straightforward [15], [19], [18]). Most importantly, HD does not rise any algorithmic issue. In fact, consider expression (12) above. The number of elementary operations (or running time [20]) needed to compute HD is easily counted. Each of the three scalar products requires $\binom{n}{2}$ multiplications (between simple numbers, namely 0s or 1s) and $\binom{n}{2} - 1$ additions (between positive integers never exceeding $\binom{n}{2}$ itself). Next, another multiplication and two further additions provide total $3\left(2\binom{n}{2} - 1\right) + 3 = 3n(n-1)$. This is evidently better than the best running time $O(n^3)$ when computing MMD in bioinformatics [18].

Beside computations, HD and MMD can also be compared through examples showing the different sensitivity of their normalized versions (see below). This seems achievable even for the small value $n = 7$ applying to Table 2 below. Consider that the partition lattice has n levels. Maintaining the notation used thus far for level 1 (above level 0), populated by $\binom{n}{2}$ atoms, let $\mathcal{P}^N_{(n-k)}$ denote level $n - k$, for $k = 1, \ldots, n$. Each level has a number of elements given by the *Stirling numbers of the second kind* $\mathcal{S}_{n,k}$ [1], [14]. For $k = 1, \ldots, n$,

$$\mathcal{S}_{n,k} = \frac{1}{k!}\sum_{0\leq h\leq n}(-1)^{k-h}\binom{k}{h}h^n = |\mathcal{P}^N_{(n-k)}|,$$

TABLE II
Comparing HD and MMD, normalized values in parentheses.

#	P	Q	$HD(P,Q)$	$MMD(P,Q)$
1	12|3|4|5|6|7	1|2|3|4|5|67	2 (0.095)	2 (0.333)
2	12|3|4|5|67	1|2|345|6|7	5 (0.238)	4 (0.667)
3	123|4|5|6|7	1|2|3|4|567	6 (0.286)	4 (0.667)
4	1234|5|6|7	1|2|3|4|567	9 (0.429)	5 (0.833)
5	1234|5|6|7	1235|4|6|7	6 (0.286)	2 (0.333)
6	123|45|67	17|25|36|4	8 (0.381)	4 (0.667)
7	123|4567	1456|237	10 (0.472)	5 (0.833)
8	123|45|67	1|2|34567	11 (0.524)	4 (0.667)
9	123|45673	1234567	8 (0.381)	5 (0.833)
10	123|45|67	123|4567	4 (0.190)	3 (0.5)
11	123|4567	167|2345	12 (0.571)	4 (0.667)
12	12|34|56|7	1|23|45|67	6 (0.286)	3 (0.5)
13	123|45|6|7	167|2|34|5	8 (0.381)	5 (0.833)
14	1|2|3|4|5|6|7	1234567	21 (1)	6 (1)
15	12|34|56|7	17|23|45|6	6 (0.286)	4 (0.667)
16	123|4|56|7	1|72|345|6	8 (0.381)	3 (0.5)
17	1|234567	12|3|4|5|6|7	16 (0.762)	5 (0.833)
18	1234567	12|3|4|5|67	19 (0.905)	5 (0.833)
19	1234567	12|3|4|5|6|7	20 (0.952)	5 (0.833)
20	1234567	1|234567	6 (0.286)	1 (0.167)

where $h > k \Rightarrow \binom{k}{h} = 0$. These numbers are known to be such that most populated levels are central ones [1, Proposition 3.30 p. 91]. Also, upper levels are comparatively more populated than lower ones. For $n = 7$,

$$\begin{aligned}
\mathcal{S}_{7,7} &= 1 = \mathcal{S}_{7,1}, \\
\mathcal{S}_{7,6} &= 21 = \binom{7}{2}, \\
\mathcal{S}_{7,5} &= 140, \\
\mathcal{S}_{7,4} &= 350, \\
\mathcal{S}_{7,3} &= 301, \\
\mathcal{S}_{7,2} &= 63.
\end{aligned}$$

Examples 1-20 in Table II above compare measurement sensitivity between MMD and HD with partitions picked initially from less populated levels, and subsequently from more populated ones.

It seems worth noting again that while MMD takes all n integer values $0, 1, \ldots, n-1$ in its interval $[0, n-1]$, neither VI nor HD have a range displaying the same even distribution over associated intervals $[0, \log n]$ and $[0, \binom{n}{2}]$, respectively. Concerning HD, the size of partitions does not take all values $0, 1, \ldots, \binom{n}{2}$, as explained in section 1.3. In particular, some parts of the interval are more densely populated by range elements than others. This is another consequence of linear dependence, characterizing all (non-corrected) partition distance measures based on counting pairs; see [22, section 2.1]. Most importantly, the number of range elements is greater with HD (and VI) than with MMD, i.e. $|rg_n(HD)| > |rg_n(MMD)|$ as soon as $n > 3$, where $rg_n(MMD) = \{0, 1, \ldots, n-1\}$ and $rg_n(HD) \subseteq \{0, 1, \ldots, \binom{n}{2}\}$ denote the range of MMD and HD, respectively, for any given n. These are sets of integer numbers, and a larger range provides greater measurement sensitivity. Finally, any distance measure may be normalized so to range in the unit interval $[0, 1]$. That is, $\frac{MMD(P,Q)}{n-1}, \frac{VI(P,Q)}{\log n}, \frac{HD(P,Q)}{\binom{n}{2}} \in [0, 1]$ for all $P, Q \in \mathcal{P}^N$. These normalized values are within parentheses in Table II.

VII. Conclusion

In statistical classification, measuring the distance between partitions is an important combinatorial problem, attracting attention since the '60s [26], [21], [11]. Solution HD provided here parallels the (traditional) Hamming distance between subsets by counting atoms in the corresponding lattices. From an applicative viewpoint, HD is added to MMD and VI as tools for clustering comparison in bioinformatics. Easily computed, HD has fine measurement sensitivity and displays suitable axiomatic features.

References

[1] M. Aigner. *Combinatorial Theory*. Springer, 1997. Reprint of the 1979 edition.

[2] A. Almudevar. A commentary on some recent methods in sibling group reconstruction based on set coverings. *Optimization Methods and Software*, 26(6):993–1003, 2011.

[3] A. Almudevar and E. C. Anderson. A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. *Molecular Ecology Resources*, 12(1):164–178, 2012.

[4] A. Almudevar and C. Field. Estimation of single-generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological and Environmental Statistics*, 4(2):136–165, 1999.

[5] M. V. Ashley, I. C. Caballero, W. A. Chaovalitwongse, B. DasGupta, P. Govindan, S. I. Sheikh, and T. Y. Berger-Wolf. KINALYZER, a computer program for reconstructing sibling groups. *Molecular Ecology Resources*, 9(4):1127–1131, 2009.

[6] T. Y. Berger-Wolf, S. I. Sheikh, B. DasGupta, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, and S. L. Putrevu. Reconstructing sibling relationship in wild populations. *Bioinformatics*, 23(13):i49–i56, 2007.

[7] B. Bollobas. *Combinatorics. Set Systems, Hypergraphs, Families of Vectors, and Combinatorial Probability*. Cambridge University Press, 1986.

[8] D. G. Brown and T. Y. Berger-Wolf. Discovering kinship through small subsets. *Algorithms in Bioinformatics*, LNCS 6293:111–123, 2010.

[9] D. G. Brown and D. Dexter. Sibjoin: a fast heuristic for half-sibling reconstruction. *Algorithms in Bioinformatics*, LNCS 7534:44–56, 2012.

[10] W. A. Chaovalitwongse, T. Y. Berger-Wolf, B. DasGupta, and M. V. Ashley. Set covering approach for reconstruction of sibling relationships. *Optimization Methods and Software*, 22(1):11–24, 2007.

[11] W. H. E. Day. The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, 1(3):269–287, 1981.

[12] D. Dexter. *Reconstruction of half-sibling population structures*. Master thesis, Computer Science, Waterloo, 2012.

[13] M. M. Deza and E. Deza. *Encyclopedia of Distances - Second Edition*. Springer, 2013.

[14] R. Graham, D. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, 1994.

[15] D. Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82:159–164, 2002.

[16] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[17] D. A. Konovalov. Accuracy of four heuristics for the full sibship reconstruction problem in the presence of genotype errors. *Series on Advances in Bioinformatics and Computational Biology*, 3:7–16, 2006.

[18] D. A. Konovalov, N. Bajema, and B. Litow. Modified Simpson $O(n^3)$ algorithm for the full sibship reconstruction problem. *Bioinformatics*, 21(20):3912–3917, 2005.

[19] D. A. Konovalov, B. Litow, and N. Bajema. Partition-distance via the assignment problem. *Bioinformatics*, 21(10):2463–2468, 2005.

[20] B. Korte and J. Vygen. *Combinatorial Optimization: Theory and Algorithms (2nd edition)*. Springer, 2002.

[21] I. C. Lerman. *Classification et Analyse Ordinale des Données*. Dunod, 1981.

[22] M. Meila. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[23] B. G. Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic Press, 1996.
[24] B. G. Mirkin and L. B. Cherny. Measurement of the distance between distinct partitions of a finite set of objects. *Automation and Remote Control*, 31(5):786–792, 1970.
[25] B. G. Mirkin and I. Muchnik. Some topics of current interest in clustering: Russian approaches 1960-1985. *Electronic Journal for History of Probability and Statistics*, 4(2):1–12, 2008.
[26] S. Rénier. Sur quelques aspects mathématiques des problémes de classification automatique. *ICC Bulletin*, 4:175–191, 1965. Reprinted in Mathématiques et Sciences Humaines 82:13-29, 1983.
[27] G. Rossi. Information functions and expectation. In *RUD Proceedings*, 2004. www.kellogg.northwestern.edu/research/risk/rud/risk_papers.htm.
[28] G. Rossi. *Partition distances*. arXiv:1106.4579v1, 2011.
[29] G.-C. Rota. The number of partitions of a set. *American Mathematical Monthly*, 71:499–504, 1964.
[30] S. I. Sheikh, T. Y. Berger-Wolf, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, and B. DasGupta. Error-tolerant sibship reconstruction in wild populations. *Computational Systems Bioinformatics*, 7:273–284, 2008.
[31] S. I. Sheikh, T. Y. Berger-Wolf, A. A. Khokhar, I. C. Caballero, M. V. Ashley, W. Chaovalitwongse, C.-A. Chou, and B. DasGupta. Combinatorial reconstruction of half-sibling groups from microsatellite data. *Journal of Bioinformatics and Computational Biology*, 8(2):337–356, 2010.
[32] M. Stern. *Semimodular Lattices. Theory and Applications. Encyclopedia of Mathematics and its Applications 73*. Cambridge University Press, 1999.
[33] M. J. Warrens. On the equivalence of Chen's Kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification*, 25(1):177–183, 2008.
[34] H. Whitney. On the abstract properties of linear dependence. *American Journal of Mathematics*, 57:509–533, 1935.