Reduction of search space for the mean partition problem

JYRKO CORREA-MORRIS
Mathemtics and Natural Sciences Department
Miami Dade College, Padron Campus
627 SW 27 Ave, Miami, Florida
USA

Abstract: The contributions of this paper are threefold. First, it conducts a formal comparison of the primary approaches to consensus clustering, using the concepts of *agreement* and *consent.* Secondly, it presents theoretical evidence justifying the preference for mean-based methods, which rely on consent, over other agreement-based procedural methods like the q-rule, which are now mostly used as quality evaluators in practice. Thirdly, the paper computes the exact reduction achieved by criteria available in existing literature to assess the quality of mean-based consensus solutions and reduce the search space's size. Finally, it compiles the regions where consensus functions associated with well-known dissimilarity measures, such as the Mirkin metric and Variation of Information, accumulate their consensus solutions.

Key-Words: Lattice of partitions, Consensus, Mean partition, Quota Rules, Cluster ensemble, Search Space Reduction

Received: May 25, 2023. Revised: August 26, 2023. Accepted: September 28, 2023. Published: October 20, 2023.

1 Introduction

Clustering problems are widespread and have led to the development of numerous algorithms; however, the lack of reliable performance information has created "The User's Dilemma," [1], making algorithm selection uncertain.

To address this, two approaches have emerged. First, formal theories aim to establish generic concepts and rules to classify and understand clustering algorithms, [2], [3], [4], [5]. Second, more advanced algorithms combine outcomes from multiple traditional methods to produce integrated solutions, known as "consensus methods," [6], [7], [8], [9]. This paper primarily focuses on consensus methods, particularly those based on the concept of a *mean partition*.

1.1 Consensus methods

Consensus algorithms take a set of partitions $\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_m$, usually called an *ensemble* or a *profile*, generated by standard clustering algorithms and merge them into a holistic solution called the "consensus partition." These input partitions represent partial solutions, akin to expert judgments from different perspectives. The consensus partition aims to condense these partial perspectives into a comprehensive representation. There are several interesting axiomatic studies, many of them published in journal of economics or social sciences, [10], [11], [12], [13], [14], which address partitions as equivalence relations and define the consensus mechanisms as a function that receives an m-tuple of equivalence relations (an ordered ensemble) and returns the equivalence relation resulting of taking the *meet* (see Section 2) of some subset of the inputs determined by a fixed subset of the indexes (e.g., the meet of the first, third, and fifth partitions in the tuple), [11]. Similar mechanisms are obtained if instead of the meet, the *join* operator is used. In the first case, the consensus mechanism is called a *meet aggregator*, while in the second case it is called *join aggregator*. What varies from one aggregator to another of the same kind is the set of indexes to be considered. Apart from the axiomatic approaches, there exist two closely related fundamental mechanisms of consensus in the ambit of partitions, which have been often applied in pattern recognition and machine learning tasks: quota rules [15] and mean-based consensus, [16], [6], [7], [8].

Quota rules constitute a slightly more flexible join aggregator. Given $0 \leq q \leq m$ (m denotes the total number of partitions in the ensemble), the q-quota rule is the *finest* partition that places in the same cluster all those pairs of objects that were placed in the same cluster of more than q partitions of the ensemble, [15], [17]. Given a subset A of X, we said that a partition of X *refines* A is every cluster of this partition either lies entirely within A or entirely outside A. The dual q-quota rule, $0 \leq q \leq m$, is the *coarsest* partition that refines all subsets A of X that are refined by more than q partitions of the ensemble.

In turn, the consensus solutions in the mean-based approach are defined as the minimizers of the consensus function $\varphi(\mathrm{p}) = \sum_{i=1}^{m} D(\mathrm{p}, \mathrm{p}_i)$, where D is a dissimilarity measure for quantifying the resemblance between two partitions and $\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_m$ are

the members of the ensemble.

When comparing different methodologies, quota rules align nicely with the idea of agreement, offering straightforwardly interpretable results. However, they are highly sensitive to factions within the ensemble, which is common in practice. This sensitivity leads to fragmented partitions and limited insights into data interrelations. Meet aggregators share similar sensitivity issues. In contrast, dual quota rules and join aggregators tend to produce compact partitions dominated by a few large clusters, known as the "chaining effect."

Mean-based methods, favored for their robustness in practical applications, handle diverse ensembles and provide more informative outcomes compared to quota rules, especially regarding cluster number and size. However, accurately computing mean-based solutions can be computationally expensive due to the large search space, unless additional information is available for effective reduction. Additionally, the accuracy of approximations by commonly used heuristic methods for averaging remains uncertain.

From a foundational perspective, there are uncertainties about how the consensus function φ that computes the mean using a dissimilarity measure aligns meet and join operations in the lattice of partitions with basic arithmetic operations on real numbers. These unique characteristics raise questions concerning the solutions quality and regarding the solutions location.

1.1.1 Quality of consensus solutions

As mentioned earlier, quota rules and dual quota rules align well with the concept of agreement, making their results easily interpretable and practical. While quota rules offer limited but consistently reliable information, dual quota rules might provide some noise by potentially including some irrelevant information. Quota rules may not be highly informative as consensus solutions, but they capture essential information. In contrast, dual quota rules encompass all the information (even if it includes noise) that a quality consensus solution should contain.

Consequently, these rules, although not serving as primary consensus solutions, serve as quality evaluators. They assess the quality of mean-based consensus solutions as follows: a consensus solution is considered acceptable if it contains at least the information provided by a quota rule and no more than what a dual quota rule offers. These rules play a crucial role in evaluating the effectiveness of mean-based consensus solutions.

This situation has been empirically addressed by several researchers, [18], [19], [20].

1.1.2 Location of consensus solutions

To tackle computational challenges, various approaches have proposed reducing the search space, each with its own method. While these reductions may appear intuitive, they often lack substantial evidence or support from general principles governing consensus solution behavior.

The exponential growth of the Bell sequence (the nth Bell number represents the number of partitions of a set with n elements) underscores the impossibility of evaluating each individual partition of a given finite dataset to find consensus solutions. Hence, practical procedures frequently employ unverified search methods. Pruning the search space becomes necessary to handle computational constraints. This involves identifying a smaller region where consensus solutions are likely to be found, thereby improving computational efficiency and approximation quality.

However, certain questions arise that warrant consideration: Is there a common region where consensus solutions, regardless of the consensus function, should be sought? For many dissimilarity measures, the answer is yes. As mentioned earlier, quota rules and dual quota rules serve as lower and upper bounds of consensus solutions, significantly reducing the search space.

Another important question is whether all consensus functions accumulate their solutions in the same region of the search space. The answer is no. For example, consensus functions associated with the lattice metric and the symmetric difference metric accumulate their solutions in opposing regions of the partition space, [15].

1.2 Contributions

Motivated by these considerations, this paper delves into the distinction between "agreement" and "consent" within the context of clustering consensus methods, focusing on significant subtleties and nuances, and formalizes these concepts. Subsequently, the article translates the primary agreement-based consensus criteria into rules for evaluating the quality of mean-based consensus solutions.

In addition to these criteria, the paper incorporates other criteria derived from axiomatic approaches, though it is important to note that not all of them are equally applicable or desirable. These are necessary steps to facilitate access to the main contribution of the paper: the exact computation of the pruning capability of these criteria.

For each criterion, the paper investigates its potential as a means to prune the search space by precisely calculating the number of partitions it eliminates. This aspect of the methods has not been explored to the best of my knowledge, making it a novel and unexplored area of research.

By addressing these aspects, the paper aims to deepen our understanding of clustering consensus methods and provide valuable insights into their performance and effectiveness.

2 The size of the lattice of partitions

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a finite set. A *partition* p of X is a collection $\mathrm{p} = \{C_1, C_2, \ldots, C_s\}$ of subsets of X such that $C_i \cap C_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^{s} C_i = X$. The subsets $C_1, C_2, \ldots, C_s$ are called the clusters or blocks of p. Throughout this paper, m_X denotes the partition all whose clusters are singletons —$\mathrm{m}_X := \{\{x\} : x \in X\}$—, while g_X denotes the partition whose only cluster is the entire set X —$\mathrm{g}_X := \{X\}$. There is a natural partial order between partition: we say that the partition p *refines* the partition p′, in notation notation $\mathrm{p} \preceq \mathrm{p}'$, if the following condition holds: for every pair of objects $x, x' \in X$, if x and x' are in the same cluster of p, then they are also in the same cluster of p′. Naturally, $\mathrm{p} \prec \mathrm{p}'$ means that p strictly refines p′. We say that a partition p *covers* a partition p′ if (a) p and p′ are different partitions; (b) p′ refines p; and (c) if p′ refines p″ which in turn refines p, then either $\mathrm{p}'' = \mathrm{p}'$ or $\mathrm{p}'' = \mathrm{p}$ (i.e., there is no partition in between p′ and p). The notation $\mathrm{p}' \sqsubset \mathrm{p}$ means p covers p′. For any two partitions p and p′ there are partitions that refine both. The coarsest partition that satisfies this property is denoted by $\mathrm{p} \wedge \mathrm{p}'$ and is called the *meet* of p and p′. Two elements $x, y \in X$ are placed in the same cluster of $\mathrm{p} \wedge \mathrm{p}'$ if and only if they are simultaneously placed in the same cluster of p and in the same cluster of p′. Therefore the clusters of $\mathrm{p} \wedge \mathrm{p}'$ are all the possible non-empty intersections of one cluster of p with one cluster of p′. For any two partitions p and p′ there are partitions that are refined by both. The finest partition that satisfies this property is denoted by $\mathrm{p} \vee \mathrm{p}'$ and is called the *join* of p and p′. Two elements $x, y \in X$ are placed in the same cluster of $\mathrm{p} \vee \mathrm{p}'$ if and only if there is a sequence $x = x_{i_1}, x_{i_2}, \ldots, x_{i_k} = y$ such that, for all $j \in \{1, 2, \ldots, k-1\}$, x_{i_j} and $x_{i_{j+1}}$ are either placed in the same cluster of p or in the same cluster of p′. Thus, the cluster of $\mathrm{p} \vee \mathrm{p}'$ are the subsets of X that can be expressed both as unions of clusters of p and as unions of clusters of p′, and are minimal with respect to this property. Henceforth, $\mathbb{P}_X$ denotes the lattice of all partitions of X.

The partitions with exactly $n-1$ clusters, i.e., partitions all whose clusters are singletons except for one which includes exactly two elements, are called the *atoms* or *atomic partitions* of $\mathbb{P}_X$. From now on, $\mathrm{p}_{xx'}$ denotes the atomic partition whose only non-singleton cluster is $\{x, x'\}$. Moreover, $\mathcal{A}_X$, or simply $\mathcal{A}$, denotes the set of atoms of $\mathbb{P}_X$ while $\mathcal{A}(\mathrm{p})$ denotes the set of all the atoms of $\mathbb{P}_X$ that refine the given partition p. Analogously, the partitions with exactly 2 clusters are called the *co-atoms* or *co-atomic partitions* of $\mathbb{P}_X$. Henceforth, p_M denotes the co-atomic partition whose only clusters are M and $X - M$. $C\mathcal{A}_X$ denotes the set of co-atoms of $\mathbb{P}_X$ and $C\mathcal{A}(\mathrm{p})$ denotes the set of co-atoms of $\mathbb{P}_X$ that are refined the given partition p. It is well-known that every partition $\mathrm{p} \neq \mathrm{m}_X$ can be expressed as the join of atom while every partition $\mathrm{p} \neq \mathrm{g}_X$ can be represented as the meet of co-atoms.

The number of partitions in $\mathbb{P}_X$ is given by the nth Bell number, B_n, where n is the number of elements in X. Bell numbers can be computed by using any of the following equivalent formulas:

$$B_n = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!} \quad \text{(Dobinski's formula)}; \quad (1)$$

$$B_n = \sum_{k=0}^{n-1}\binom{n-1}{k}B_k, \; B_0 = 1; \quad (2)$$

$$B_n = \sum_{k=1}^{n} S(n,k). \quad (3)$$

Equation (3) defines the Bell numbers as the sum of the Stirling numbers of the second kind, $S(n,k)$, where, for specific values of n and k, $S(n,k)$ gives the number of partitions of n elements into k clusters. In turn, Equations (1) and (2) together show that the difference between two consecutive Bell numbers is an exponential, which implies that B_n increases extremely quickly as n increases. Table 1 shows the first ten Bell numbers, which are additionally compared to 2^{n-1}, as well as the difference between any two consecutive of them, in order to numerically illustrate the claim above.

Table 1: The first ten Bell numbers

n	7	8	9	10
B_n	877	4,140	21,147	115,975
$B_n - B_{n-1}$	674	3,273	17,007	94,828
2^{n-1}	64	128	256	512

3 Consensus by agreement vs. consensus by consent: Quota rules and mean-based consensus

This section commences by exploring the distinction between two closely related concepts: "agreement" and "consent." The primary objective of this discussion is to offer formal evidence regarding the dissimilarity between the two main approaches in consensus

methods: quota rules and mean-based methods. Furthermore, it seeks to clarify why the latter approach has gained more popularity in applied fields such as pattern recognition and machine learning.

In essence, agreement can be viewed as what is desired, whereas consent represents what is accepted. Consequently, in an ensemble, a member may choose to consent to a proposal that is not their top choice, aiming to achieve a solution that might not be optimal but is satisfactory to all members. As a result, obtaining consent appears to be easier than achieving agreement. This flexibility in consent-based methods leads to more informative solutions with a better balance between the size and number of clusters. On the contrary, agreement-based methods tend to reveal only a limited amount of the underlying data structure, primarily due to either a high level of fragmentation or a high level of concentration. Organizing the lattice of partitions, we observe that agreement methods often produce partitions that lie either in the lower or higher levels of the partition space.

Despite the limited information provided by agreement-based methods, it is undeniable and can be valuable as a source of information when seeking to prune the search space in mean-based methods. In this context, it plays a crucial role, even though it may not suffice as a standalone consensus solution.

To formalize these ideas, I shall introduce the following definition:

Definition 1 *In the process of creating the consensus solution* p^**, the* j*th member of the ensemble,* p_j*, is said to* agree *that the pair of objects* x *and* x' *to be placed in the same cluster of* p^* *if and only if* x *and* x' *are in the same cluster of* p_j*. Thus, the ensemble* $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \dots \mathrm{p}_m\}$ *is said to contain a quota* $q \in [0,1]$ *of agreement with the fact that the pair of objects* x *and* x' *in* X *to be placed in the same cluster of* p^* *if and only if more than* mq *members of the ensemble* $\mathcal{E}$ *agree that* x *and* x' *to be placed in the same cluster of* p^**. In other words, if*

$$|\{\mathrm{p}_j \in \mathcal{E} \,:\, \mathrm{p}_{xx'} \preceq \mathrm{p}_j\}| > mq.$$

In the process of creating the consensus solution p^**, the* j*th member of the ensemble,* p_j*, is said to* consent, according to the dissimilarity D*, that the pair of objects* x *and* x' *in* X *to be placed in the same cluster of* p^* *if and only if the partition obtained from* p^* *by merging those clusters containing* x *and* x'*, respectively, either maintains or lowers the dissimilarity with respect to* p_j*. In other words,* $D(\mathrm{p}_j, \mathrm{p} \vee \mathrm{p}_{xx'}) \leq D(\mathrm{p}_j, \mathrm{p})$*. The entire ensemble* $\mathcal{E}$ *is said to consent, according to* D*, that the pair of objects* x *and* x' *in* X *to be placed in the same cluster of* p^* *if and only if* $\varphi(\mathrm{p}^* \vee \mathrm{p}_{xx'}) \leq \varphi(\mathrm{p}^*)$*, which amounts to saying that*

$$\sum_{i=1}^{m} D(\mathrm{p}_i, \mathrm{p}^* \vee \mathrm{p}_{xx'}) \leq \sum_{i=1}^{m} D(\mathrm{p}_i, \mathrm{p}^*).$$

The dissimilarity $D(\mathrm{p}_j, \mathrm{p}^*)$ *can be thought as the degree of consent that the* j*th member of the ensemble,* p_j*, assigns to the proposal "the partition* p^* *is the consensus solution".*

This definition establishes that a consensus reached by agreement is the solution of certain q-quota rule, while a consensus obtained by consent is the solution of a certain mean-based method. Now, it is worthy, for the sake of completeness, to introduce a more formal definition of quota rules.

Quota rules*:* Given an ensemble $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \dots \mathrm{p}_m\}$ of partial partitionings of the data set X, and a real number $q \in [0,1]$, the q-quota rule, in notation p_q, is the finer partition from among those that are refined by every atom that, in turn, refines more than mq members of the ensemble.

By definition, the finer partition refined by some given partitions is their join. It can be therefore concluded that:

$$\mathrm{p}_q = \bigvee \{\mathrm{p}_{xx'} \in \mathcal{A} : |\{\mathrm{p}_i \in \mathcal{E} : \mathrm{p}_{xx'} \preceq \mathrm{p}_i\}| > mq\}.$$

In the case that the set of such atoms is empty, then $\mathrm{p}_q = \mathrm{m}_X$.

Among the q-quota rules, unanimity rule and majority rule stand out by being used the most in practice (see, for instance, [15]).

Unanimity rule*:* The unanimity rule is the q-quota rule that corresponds to $q = \dfrac{m-1}{m}$. (In this case, $mq = m - 1$, which means that the atoms to be considered are those that refine all the members of the ensemble.) Accordingly, the unanimity consensus is given by the meet of all the members of the ensemble: $\bigwedge_{i=1}^{m} \mathrm{p}_i$.

Majority rule*:* The majority rule is the q-quota rule that corresponds to $q = 1/2$.

In an effort to address the issue of quota rules tending to generate small clusters, the concept of Dual Quota Rules has been introduced [15].

Dual quota rules*:* Given an ensemble $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \dots \mathrm{p}_m\}$ of partial partitionings of the data set X and a real number $q \in [0,1]$, the dual q-quota rule, in notation p_q^d, is the coarser partition from among those that refine every co-atoms that, in turn, is refined by more than mq members of the ensemble.

Since the coarser partition that refines some given partitions is their meet, it can be concluded that:

$$\mathrm{p}_q^d = \bigwedge \{\mathrm{p}_M \in C\mathcal{A}, : |\{\mathrm{p}_i \in \mathcal{E} : \mathrm{p}_i \preceq \mathrm{p}_M\}| > mq\}.$$

In the event that the set of such co-atoms is empty, then $\mathrm{p}_q^d = \mathrm{g}_X$.

The Dual unanimity rule and the Dual majority rule correspond to $q = \frac{m-1}{m}$ and $q = \frac{1}{2}$, respectively. Similar to quota rules, these dual rules are frequently employed in practical applications.

Restated in terms of consent, quota rules and dual quota serve as criteria designed to evaluate the inner workings of other consensus methods that may not be equally transparent. This assessment provides valuable information about the quality and location of their consensus solutions, which can then be utilized to effectively prune the search space.

***Quota rules for consent**:* Given an ensemble $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \ldots \mathrm{p}_m\}$ of partial partitionings of the data set X and a real number $q \in [0,1]$, the q-quota rule for consent establishes that any consensus partition, particularly the minimizers of the mean-based consensus function $\varphi(\mathrm{p}) = \sum_{i=1}^{m} D(\mathrm{p}, \mathrm{p}_i)$, must be refined by p_q.

In essence, q-quota rules for consent stipulate that pairs of objects appearing together in the same cluster for more than qm partitions of the ensemble should be grouped together in the same cluster of any consensus partition. Consequently, this rule ensures that any consensus solution must be at least as inclusive as p_q, rather than declaring p_q itself as the consensus solution. To illustrate, in terms of intervals in $\mathbb{P}_X$, these rules propose replacing a larger interval, namely the entire set $\mathbb{P}_X$ of partitions X, with a smaller interval defined as $[\mathrm{p}_q, \mathrm{g}_X]$.

Similarly, the dual q-quota rules for consent assert that any consensus solution lies within the interval $[\mathrm{m}_X, \mathrm{p}_q]$.

***Dual quota rules for consent**:* Given an ensemble $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \ldots \mathrm{p}_m\}$ of partial partitionings of the data set X and a real number $q \in [0,1]$, the dual q-quota rule for consent establishes that any consensus partition, particularly the minimizers of the mean-based consensus function $\varphi(\mathrm{p}) = \sum_{i=1}^{m} D(\mathrm{p}, \mathrm{p}_i)$, must refine p_q^d.

It is time to introduce the Undesired Atoms criterion. Instead of listing the pairs that should appear together in the same cluster of a consensus partition, the following criterion has a prohibitive approach.

***Undesired atoms**:* Given an ensemble $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \ldots \mathrm{p}_m\}$ of partial partitionings of the dataset X, this rule establishes that any atomic partition $\mathrm{p}_{xx'}$ for which there is no sequence $x = x_{i_1}, x_{i_2}, \ldots, x_{i_k} = x'$ such that every atom $\mathrm{p}_{x_{i_j}, x_{i_{j+1}}}$, $j = 1, 2, \ldots, j-1$, refines at least one partition of the ensemble $\mathcal{E}$, must refine no consensus solution.

This rule finds its foundation in the notion that there are essentially two fundamental reasons for placing a pair of objects in the same cluster of a consensus partition: (i) when the ensemble reaches an agreement on this pairing, and (ii) when the ensemble decides to do so through consent or cooperation. For (i), this pair must appear together in a sufficient number of ensemble members, ensuring the required quota of agreement. Conversely, in (ii), this pair is placed together in the same cluster of the consensus partition due to the chaining effect resulting from cooperation among the ensemble members. An undesirable atom is one that lacks justification for its elements to be grouped together in the same cluster, neither through agreement nor cooperation.

To elaborate, if a partition p refined by an undesirable atom $\mathrm{p}xx'$ is prohibited from being a consensus solution, it implies that the union of the intervals $[\mathrm{p}_{xx'}, \mathrm{g}_X]$, where $\mathrm{p}_{xx'}$ ranges over the set of undesired atoms, does not contain any consensus partitions.

To conclude this section, I will introduce two novel criteria that I have recently uncovered while investigating the influence of submodularity on the localization of consensus partitions, [21], [22]. Submodular functions can be considered akin to discrete convex functions, and as it is widely recognized, convexity is one of the most advantageous scenarios in continuous optimization. The exploration of the impact of this seemingly natural property has garnered increasing interest within several areas of mathematics and computer science, including optimization and machine learning.

***Unanimity rule for consent bounded from above by the ensemble**:* Given an ensemble $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \ldots \mathrm{p}_m\}$ of partial partitionings of the dataset X, this rule establishes that any consensus solution must be refined by the meet of all the members of $\mathcal{E}$, $\bigwedge_{j=1}^{m} \mathrm{p}_j$, and refines at least one member of $\mathcal{E}$.

Essentially, this rule asserts that any consensus partition can be obtained by refining a certain partition of the ensemble to be discovered, ensuring that no pairs placed together in the same cluster for any partition in the ensemble are split apart. Formally, this statement implies that all consensus partitions lie within the union of the intervals $\left[\bigwedge_{j=1}^{m} \mathrm{p}_j, \mathrm{p}_i\right]$, $\mathrm{p}_i \in \mathcal{E}$. While this rule may impose a restrictive condition, it is not intended to be satisfied by every consensus function; rather, it has emerged naturally when investigating the behavior of consensus solutions produced by submodular consensus functions.

In a similar vein and stemming from the same source, the following and final rule posits that any consensus partition can be identified by merging clus-

ters from a particular member of the ensemble to be revealed, with the assurance that all new pairs are the result of cooperation among the members of the ensemble.

Dual unanimity rule for consent bounded from below by the ensemble: Given an ensemble $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \ldots \mathrm{p}_m\}$ of partial partitionings of the dataset X, this rule establishes that any consensus solution must be refined by at least one the p_i's in $\mathcal{E}$ and refines $\bigvee_{j=1}^{m} \mathrm{p}_j$, the join of all the members of $\mathcal{E}$. Accordingly, any consensus partition lies in the union of the intervals $\left[\mathrm{p}_i, \bigvee_{j=1}^{m} \mathrm{p}_j\right]$, $\mathrm{p}_i \in \mathcal{E}$.

4 On the exact prune of the search space

In this section, we calculate the exact reduction offered by each criterion examined in the previous section. The phrase "reduction provided by the rule $\mathcal{R}$" implies that if an average consensus function satisfies the rule $\mathcal{R}$, the minimizers of φ are among the partitions of $\mathbb{P}_X$ accepted by $\mathcal{R}$.

As each criterion provides a region that is either an interval or a collection of intervals, it is convenient to determine the number of partitions contained within an interval $[\mathrm{p}, \mathrm{p}']$.

Lemma 1 *Let* p *and* $\mathrm{p}' = \{C_1, C_2, \ldots, C_k\}$ *be partitions such that* $\mathrm{p} \prec \mathrm{p}'$. *If the cluster* C_i, $1 \leq i \leq k$, *of* p' *is the union of* n_i *clusters of* p, *then the number of partitions in the interval* $[\mathrm{p}, \mathrm{p}']$ *is* $\prod_{i=1}^{k} B_{n_i}$.

Table 2 shows some illustrative numbers.

Table 2: Computing the number of partitions in $[\mathrm{p}, \mathrm{p}']$ by applying Lemma 1

n	k	#p	$n_1; \ldots; n_k$	$\prod_{i=1}^{k} B_{n_i}$	B_n
10	2	7	3; 4	75	115,975
10	2	8	4; 4	225	115,975
10	2	8	2; 6	406	115,975
12	3	10	2; 3; 5	520	4,213, 597
12	4	11	2; 2; 3; 4	300	4,213,597
15	3	12	3; 4; 5	3 900	1,382,958,545
15	4	10	3; 3; 4	375	1,382,958,545

PROOF 1 *Any partition that refines the partition* p' *can be thought of as an union of partitions of the clusters of* p'. *Thus, as a first step toward our desired conclusion, we could think that if we compute the total number* a_i *of ways in which every cluster* C_i *of* p' *can be partitioned, then the product of the product of the* a_i*'s would provide us with an estimate of the number that we are looking for. Notice that* C_i *is a set itself, hence the number of partitions of* C_i *is the Bell number corresponding to the number of elements in* X *that form* C_i. *However, we do not want to count every partition that refines* p', *but only those that are refined by* p *as well. To take into consideration the latter condition, it suffices to count only those partitions of* C_i *in which the elements lying in the same cluster of* p *remain together. This means that the finest partition of* C_i *that we can consider is that consisting of those clusters of* p *that are contained in* C_i. *This suggests that, instead of considering the elements of* X *that form* C_i *to determine the Bell number that matches* a_i, *the clusters of* p *contained in* C_i *should be considered as individual elements to determine such Bell number. Since there are* n_i *of such clusters of* p, *the number of ways of partitioning* C_i *under the required condition is* B_{n_i}, *the* n_i*th Bell number. Therefore, the number of partitions in the interval* $[\mathrm{p}, \mathrm{p}']$ *is* $\prod_{i=1}^{k} B_{n_i}$.

Now, we are poised to calculate the precise reduction offered by each criterion presented in the previous section. I will present them in the same order in which they were introduced earlier.

Proposition 1 *If* p_q *has* k *clusters, then the reduction provided by* q*-quota rule for consent is of* $B_n - B_k$ *partitions.*

PROOF 2 *Remember that the* q*-quota rule for consent basically establishes that any consensus partition must lie in the interval* $\left[\mathrm{p}_q, g_X\right]$. *Since the only cluster of* g_X *is the entire set* X, *which is the union of* k *clusters of* p_q, *it can be concluded, in view of Lemma 1, that the number of partitions in the interval* $\left[\mathrm{p}_q, g_X\right]$ *is* B_k. *Consequently, the reduction provided by* q*-quota rule is of* $B_n - B_k$ *partitions.*

Let us analyze now the reduction provided by the dual q-quota rule.

Proposition 2 *If the cluster* C_i *of* p_q^d *has* n_i *elements, then the reduction provided by the dual* q*-quota rule is of* $B_n - \prod_{i=1}^{k} B_{n_i}$ *partitions.*

PROOF 3 *The dual* q*-quota rules basically states that any consensus partition lies in the interval* $\left[m_X, \mathrm{p}_q^d\right]$. *Since all the clusters of* m_X *are singletons, each cluster of* p_q^d *is the union of* n_i *clusters of* m_X *if and only if it contains* n_i *elements of* X. *The result is now a direct consequence of Lemma 1.*

Now, let us direct our attention to the Undesired Atoms rule. Before calculating the reduction provided by this rule, we must address a more fundamental question: How many undesirable atoms are there? An atom $\mathrm{p}_{xx'}$ is deemed undesirable if and only if the elements x and x' belong to different clusters of $\bigvee_{j=1}^{m} \mathrm{p}_j$—the join of all members in the ensemble $\mathcal{E}$. Thus, the number of undesirable atoms is equivalent to the number of possibilities for selecting a pair of elements from X that reside in distinct clusters of $\bigvee_{j=1}^{m} \mathrm{p}_j$.

Given this, each element in an arbitrary cluster C_u of the join can be paired with each element in another arbitrary cluster C_v of this partition, resulting in $\#C_u \cdot \#C_v$ undesirable pairs. This proves the following lemma.

Lemma 2 *If $\bigvee_{j=1}^{m} \mathrm{p}_j$ has k distinct clusters $C_1, C_2, \ldots, C_k$ with $n_1, n_2, \ldots, n_k$, respectively, then, according to $\mathcal{E}$, there are*

$$\sum_{1 \leq u < v \leq k} n_u n_v$$

undesired atoms.

Lemma 3 *The join of $r \geq 1$ atoms consists of at least $n - r$ clusters (and at most $n - 1$ clusters).*

PROOF 4 *Given an arbitrary partition* p *and an arbitrary atom $\mathrm{p}_{xx'}$, either x and x' are together in the same cluster of* p *or not. In the case that x and x' are placed in the same cluster of* p*, then $\mathrm{p} \vee \mathrm{p}_{xx'} = \mathrm{p}$. Otherwise, when the join of* p *and $\mathrm{p}_{xx'}$ is taken, the clusters of* p *that contain x and x', respectively, merge into a single cluster. Thus, $\mathrm{p} \vee \mathrm{p}_{xx'} \sqsubset \mathrm{p}$, which means that $\mathrm{p} \vee \mathrm{p}_{xx'}$ has one cluster less than* p*. Therefore, each time the join of any partition and an atom is taken, either the partition states the same or its number of clusters decreases by one.*

Suppose the join of r atoms is now considered. We can think of the initial partition p *as any of these atoms, and then, we will take the join with the $r - 1$ remaining atoms, one at a time. By virtue of the reasoning that we just made, after having performing the join with all the $r - 1$ remaining atoms, the initial* p *lost no more than $r - 1$ clusters.*

Proposition 3 *Let $\mathcal{UA} = \{\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_q\}$ be the set of all the undesired atoms according to the ensemble $\mathcal{E}$. The reduction provided by Undesired Atoms rule is of*

$$qB_{n-1} + \sum_{j=2}^{n-1} \left(\sum_{r=j}^{q} (-1)^{r+1} p_{rj} \right) B_{n-j} \qquad (4)$$

partitions, where p_{rj} is the number of sets consisting of exactly r undesired atoms whose join has $n - j$ clusters ($1 \leq j \leq r$).

PROOF 5 *According to this rule, if a partition lies in some of the intervals $[\mathrm{A}_j, g_X]$, then it is prohibited from being a consensus solution. As a result, the union of these intervals must be removed from the search space and the reduction provided by this rule is the cardinality of such a union.*

Applying the inclusion-exclusion principle, we get that $\#\bigcup_{j=1}^{q} [\mathrm{A}_j, g_X]$ is equal to

$$\sum_{r=1}^{q} \sum_{\boldsymbol{u} \in \mathcal{O}_q^r} (-1)^{r+1} \# \left[\bigvee_{\boldsymbol{u}} \mathrm{p}_{\boldsymbol{u}}, \mathrm{g}_X \right], \qquad (5)$$

where $\mathcal{O}_q^r := \{\boldsymbol{u} = (u_1, u_2, \ldots, u_r) \in \mathbb{Z}^r : 0 < u_1 < u_2 < \ldots < u_r < q\}$ and $\bigwedge_{\boldsymbol{u}} \mathrm{p}_{\boldsymbol{u}}$ should be understood as $\mathrm{p}_{u_1} \wedge \mathrm{p}_{u_2} \wedge \ldots \wedge \mathrm{p}_{u_r}$, provided $\boldsymbol{u} = (u_1, u_2, \ldots, u_r) \in \mathcal{O}_q^r$.

Now, by virtue of Lemma 3, the join of $r \geq 2$ atoms at least $n - r$ clusters. Let us group all the possible joins of r undesired atoms according to their number of clusters. Using the fact that if a partition p *has $n - j$ clusters, the interval $[\mathrm{p}, g_X]$ has B_{n-j} partitions, we can conclude that $\mathrm{A}_{j_1} \vee \mathrm{A}_{j_2} \vee \ldots \mathrm{A}_{j_r}$ contributes to (5) with $(-1)^{r+1} B_{n-j}$ provided this join has $n - j$ clusters. Denoting now by p_{rj} the number of joins of r atoms that have $n - j$ clusters, we get that the total contribution such undesired atoms is $(-1)^{r+1} p_{rj} B_{n-j}$. Adding up these contributions, the result follows.*

Next, we examine the Unanimity Rule for Consent Bounded from Above by the Ensemble. As demonstrated earlier, this rule stipulates that consensus solutions should lie in the union of the intervals $\left[\bigwedge_{j=1}^{m} \mathrm{p}_j, \mathrm{p}_i \right]$, $i = 1, 2, \ldots, m$. Applying again the inclusion-exclusion principle combined with the property that the intersection of two intervals of the form $[\mathrm{p}, \mathrm{p}']$ and $[\mathrm{p}, \mathrm{p}'']$, respectively, is the interval $[\mathrm{p}, \mathrm{p}' \wedge \mathrm{p}'']$ because $\mathrm{p}' \wedge \mathrm{p}''$ is the coarser partition that simultaneously refines the partitions p' and p'', it follows that $\bigcup_{i=1}^{m} \left[\bigwedge_{j=1}^{m} \mathrm{p}_j, \mathrm{p}_i \right]$ is equal to:

$$\sum_{r=1}^{m-1}(-1)^{r-1}\sum_{\mathbf{u}\in\mathcal{O}_m^r}\#\left[\bigwedge_{j=1}^{m}\mathrm{p}_j,\bigwedge_{\mathbf{u}}\mathrm{p}_{\mathbf{u}}\right].$$

It suffices now to apply Lemma 1 to prove the following proposition.

Proposition 4 *Given $\boldsymbol{u}\in\mathcal{O}_m^r$, $r=1,2,\ldots,m-1$, let $n_{s_{\boldsymbol{u}}}^{\boldsymbol{u}}$ stand for the number of clusters of $\bigwedge_{j=1}^{m}\mathrm{p}_j$ whose union is the $s_{\boldsymbol{u}}$th cluster of $\bigwedge_{\boldsymbol{u}}\mathrm{p}_{\boldsymbol{u}}$, where $\mathrm{p}_{u_1},\mathrm{p}_{u_2},\ldots,\mathrm{p}_{u_r}$ belong to ensemble $\mathcal{E}$. The reduction provided by Unanimity Rule for Consent Bounded From Above by the Ensemble is of:*

$$B_n-\sum_{r=1}^{m-1}(-1)^{r-1}\sum_{\boldsymbol{u}\in\mathcal{O}_m^r}\prod_{s_{\boldsymbol{u}}}B_{n_{s_{\boldsymbol{u}}}^{\boldsymbol{u}}}.$$

Analogously to the previous rule, we can calculate the reduction provided by the Dual Unanimity Rule for Consent Bounded from Below by the Ensemble. This rule essentially states that any consensus solution lies in the union of the intervals $\left[\mathrm{p}_i,\bigvee_{j=1}^{m}\mathrm{p}_j\right]$, $i=1,2,\ldots,m$. The cardinality of these intervals can be determined using the inclusion-exclusion principle coupled with the fact that the intersection of two intervals of the form $[\mathrm{p}',\mathrm{p}]$ and $[\mathrm{p}'',\mathrm{p}]$ is the interval $[\mathrm{p}'\vee\mathrm{p}'',\mathrm{p}]$ because $\mathrm{p}'\vee\mathrm{p}''$ is the finer partition that simultaneously refines the partitions p' and p''.

Thus, the cardinality $\#\bigcup_{i=1}^{m}\left[\mathrm{p}i,\bigvee_{j=1}^{m}\mathrm{p}_j\right]$ is equal to:

$$\sum_{r=1}^{m-1}(-1)^{r-1}\sum_{\mathbf{u}\in\mathcal{O}_m^r}\#\left[\bigvee_{\mathbf{u}}\mathrm{p}_{\mathbf{u}},\bigvee_{j=1}^{m}\mathrm{p}_j\right].$$

On account of Lemma 1 we get the following proposition.

Proposition 5 *Given $\boldsymbol{u}\in\mathcal{O}_m^r$, $r=1,2,\ldots,m-1$, let $m_s^{\boldsymbol{u}}$ stand for the number of clusters of $\bigvee_{\boldsymbol{u}}\mathrm{p}_{\boldsymbol{u}}$, $\mathrm{p}_{u_1},\mathrm{p}_{u_2},\ldots,\mathrm{p}_{u_r}\in\mathcal{E}$, whose union is the sth cluster of $\bigvee_{j=1}^{m}\mathrm{p}_j$, $1\le s\le k$. The reduction provided by Dual Unanimity Rule for Consent Bounded From Below by the Ensemble is of:*

$$B_n-\sum_{r=1}^{m-1}(-1)^{r-1}\sum_{\boldsymbol{u}\in\mathcal{O}_m^r}\prod_{s=1}^{k}B_{m_s^{\boldsymbol{u}}}.$$

5 State-of-the-art dissimilarity measures

The Table 3 compiles the most significant results reported in literature for consensus functions associated to well-known dissimilarity functions.

Table 3: State-of-the-art measures' compliance with reduction criteria (Unanimity (U); Dual Unanimity (DU); Majority (M); Dual Majority (DM)).

Measure	U	DU	M	DM
Var. of Inform.	✓	✓		
Mirkin metric	✓	✓		
Dual Symm. Diff. Metric	✓	✓		✓
Lattice Metric	✓	✓		✓
0-1 Disagreement meas.	✓			
Split-merge meas.	✓		✓	

As previously mentioned, the works of Barthélemy and Leclerc ensure that the consensus functions linked to any metric constructed from a lower valuation exhibit a behavior akin to the consensus functions associated with the Variation of Information and Mirkin metric, respectively. Likewise, the consensus functions related to any metric constructed from an upper valuation display a behavior similar to the consensus function associated with the lattice metric [15].

6 Conclusions and future work

This paper embarked on an exploration of lattice theory to delve into the fundamental distinctions between the two primary approaches to consensus within the domain of partitional clustering. By presenting formal evidence, we clarified the preference for average-based methods over quota rules, despite the latter's intuitive and easily interpretable nature. We underscored that quota rules and similar criteria have not diminished in importance; rather, they have assumed a new role.

These methods are now instrumental in appraising the quality of average-consensus solutions and furnishing valuable insights into the whereabouts of average solutions. They assist in the efficient pruning of the search space. The principal contribution of this article is the compilation of key criteria for evaluating average-consensus solutions and the quantification of the exact reduction in the search space achieved by each criterion. Moreover, these findings offer insights into the pruning capabilities of each criterion, enhancing our understanding of mean-based consensus methods, which carry significant practical implications.

Looking ahead, further investigation is warranted to analyze the various average consensus functions corresponding to well-known dissimilarity measures. While the Symmetric Difference measure has received extensive scrutiny, other widely-used metrics such as Variation of Information and Van Dongen's metric remain less explored. The results presented here expand the horizons for in-depth exploration of mean-based consensus methods in future research.

References:

[1] Anil K. Jain, Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* 31(8), 2010, 651-666.

[2] Jon M. Kleinberg, An Impossibility Theorem for Clustering, in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, MIT Press, (2002) 446–453.

[3] Margareta Ackerman, Shai Ben-David, Simina Brânzei, David Loker: Weighted clustering: Towards solving the user's dilemma. *Pattern Recognit.* 120, 2021, 108152.

[4] Carlsson, Gunnar and Mamoli, Facundo, Classifying Clustering Schemes, *Foundations of Computational Mathematics*, 13 (2), (2013) 221–252.

[5] Jyrko Correa-Morris, An indication of unification for different clustering approaches, *Pattern Recognit.*, 46 (9), (2013) 2548–2561.

[6] Tossapon Boongoen, Natthakan Iam-On, Cluster ensembles: A survey of approaches with recent extensions and applications, *Computer Science Review*, 28, 2018, 1-25.

[7] Mustafa R. Kadhim, Guangyao Zhou, Wenhong Tian, A novel self-directed learning framework for cluster ensemble, *Journal of King Saud University - Computer and Information Sciences*, 34(10), Part A, 7841-7855.

[8] Hanan G. Ayad, Mohamed S. Kamel, On voting-based consensus of cluster ensembles, *Pattern Recognition*, 43(5), 2010, 1943-1953.

[9] Caroline X. Gao, Dominic Dwyer, Ye Zhu, Catherine L. Smith, Lan Du, Kate M. Filia, Johanna Bayer, Jana M. Menssink, Teresa Wang, Christoph Bergmeir, Stephen Wood, Sue M. Cotton, An overview of clustering methods with guidelines for application in mental health research, *Psychiatry Research*, 327, 2023, 115265.

[10] Fishburn, Peter C., and Ariel Rubinstein, Aggregation of equivalence relations, *Journal of classification* 3, 1986, 61-65.

[11] Dinko Dimitrov, Thierry Marchant and Debasis Mishra, Separability and aggregation of equivalence relations, *Economic Theory*, 51 (1), 2012, 212.

[12] Livieratos, John, Phokion G. Kolaitis, and Lefteris Kirousis, On the computational complexity of non-dictatorial aggregation, *Journal of Artificial Intelligence Research*, 72, 2021, 137-183.

[13] Baronchelli Andrea, The emergence of consensus: a primer, *R. Soc. Open Sci.*, 2018, 5172189172189.

[14] Jeff E. Biddle; Daniel S. Hamermesh, Theory and Measurement: Emergence, Consolidation, and Erosion of a Consensus, *History of Political Economy*, 49 (Supplement), 2017, 34–57.

[15] Jean-Pierre Barthélemy and Bruno Leclerc, The Median Procedure for Partitions, *Partitioning Data Sets, American Mathematics Society, Series in Discrete Math*, 1995, 3-34.

[16] Sandro Vega-Pons, Jyrko Correa-Morris, José Ruiz-Shulcloper: Weighted partition consensus via kernels. Pattern Recognit. 43(8), 2010, 2712-2724.

[17] Zoi Terzopoulou, Quota rules for incomplete judgments, *Mathematical Social Sciences,* 107, 2020, 23-36.

[18] J. He, L. Cai and X. Guan, Differential Private Noise Adding Mechanism and Its Application on Consensus Algorithm, *IEEE Transactions on Signal Processing*, 68, 2020, 4069-4082, doi: 10.1109/TSP.2020.3006760.

[19] Teixeira, P., Marques, M. M., Hagger, M. S., Silva, M., Brunet, J., Duda, J., … Hagger, M., Classification of techniques used in self-determination theory-based interventions in health contexts: an expert consensus study, 2019, https://doi.org/10.31234/osf.io/z9wqu

[20] Adrianna Kozierkiewicz-Hetmańska, The analysis of expert opinions' consensus quality, *Information Fusion*, 34, 2017, 80-86.

[21] Jyrko Correa-Morris, Comparing Partitions: Shortest Path Length Metrics and Submodularity, *Intern. J. of Math. Models and Meth. in Appl. Sci.*, 13, 2019, 45-51.

[22] Jyrko Correa-Morris: The median partition and submodularity, *Appl. Math. Comput.* 410, 2021, 126450.

Contribution of Individual Authors to the Creation of a Scientific Article

Jyrko Correa-Morris was responsible for the sections of the paper.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

Partial support for this research has been provided by the U.S. Department of Education under grant P120A200005.

Conflicts of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.