An Axiomatic Distance Methodology for Aggregating Multimodal Evaluations

Adolfo R. Escobedo, Erick Moreno-Centeno, and Romena Yasmin

Abstract

This work introduces a multimodal data aggregation methodology featuring optimization models and algorithms for jointly aggregating heterogenous ordinal and cardinal evaluation inputs into a consensus evaluation. Mathematical modeling components are derived to enforce three types of logical couplings between the collective ordinal and cardinal evaluations: Rating and ranking preferences, numerical and ordinal estimates, and rating and approval preferences. The methodology is tailored for use with axiomatic distances rooted in social choice theory, and it is equipped to adequately deal with highly incomplete evaluations, tied values, and other challenging aspects of distributed decision-making contexts. The practicality of the proposed methodology for group decision-making is illustrated on a case study involving an academic student paper competition. These considerations and computational aspects are further explored via synthetic instances sampled from distributions parameterized by ground truths and varying noise levels. The results show that multimodal aggregation is effective at extracting a collective truth from noisy information sources and at capturing the distinctive evaluation qualities of ordinal and cardinal evaluations in group decision-making.

Index Terms

Multimodal data aggregation, group decision-making, axiomatic distances, incomplete rankings and ratings

I. Introduction

THE natural tendency to reconcile multiple sources of conflicting information into a representative whole continues to fuel the development of data aggregation techniques. The goal of many such techniques is to systematically eliminate error/noise—thereby enhancing the quality of extractable information—arising from individual information sources who collectively evaluate the same set of entities or system of interest (Mitchell, 2012). This motivation is especially prevalent within group decision-making, collective intelligence, and various other fields that seek to make sense of multiple subjective evaluations—judgments, preferences, estimates—which are inherently *heterogenous* or contradictory. While most existing data aggregation methods have been developed for a specific modality of data—continuous, ordinal, linguistic, etc.—there is a growing interest in developing new methodologies capable of integrating multiple modalities data (Lahat et al., 2015). Such efforts are driven by the notion that different modalities are equipped to capture distinctive qualities of interest and, hence, combining them could help extract more useful information than what is possible when each data modality is considered separately.

In group decision-making and various other applications of data aggregation, it is imperative for the aggregate evaluation to be a good representation of the underlying "collective truth" inherent in the inputs. This overriding concern has led to increased attention on methods founded on the socio-theoretical concept of a *consensus* (e.g., see Brandt et al. (2016); Cook (2006); Hassanzadeh and Milenkovic (2014)). Consensus aggregation methods find an aggregate evaluation of a set of *objects* that least disagrees (or, equivalently, most agrees) with the evaluations provided by a set of individuals or *judges*. Most of these methods are differentiated by whether the evaluations are elicited (or treated) in an ordinal or in a cardinal format and by the choice of measure of disagreement. An ordinal evaluation is one where the objects are ordered based on how they compare to one another relative to a specific criterion of interest. In group decision-making, *ranking* vectors are used to order the objects from "most preferred" to "least preferred" by assigning them a nondecreasing sequence of numerical values. A cardinal evaluation is one where individual objects are assigned a *score* (a scalar) signifying the degree by which each possesses one or more qualities of interest, quantified according to an explicit or implicit reference scale. In group decision-making, *rating* vectors are used to record the scores of multiple objects, where higher rating values signify the more preferred objects with respect to the evaluation criterion.

Adolfo R. Escobedo and Romena Yasmin are with the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287 (e-mail: adRes@asu.edu, ryasmin@asu.edu).

Erick Moreno-Centeno is with the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77840 (email: emc@tamu.edu).

The usage of cardinal versus ordinal evaluations for making fair and effective collective decisions is a longstanding point of contention (Ammar and Shah, 2011) dating back to the origins of voting theory (Young, 1988). Reasons that ordinal evaluations are advocated include their emphasis on pairwise comparisons, avoidance of subjective scales, and robustness against outliers offered by some of their associated aggregation methods. Reasons that cardinal evaluations are promoted include a lower cognitive load of elicitation (since each object can be ostensibly evaluated independently), reflection of different "intensities of preference" between the objects, and computational efficiency of their associated aggregation methods. The multimodal aggregation framework presented in this work can be regarded a unification of these contrasting theories. In particular, this work develops a general distance-based methodology that is capable of aggregating evaluations that are cardinal and/or ordinal. Although the proposed methodology is elaborated primarily within the context of group decision-making, the featured contributions could have wide-ranging applicability. This is because consensus aggregation methods have been widely employed to reconcile heterogeneous evaluations within a variety of fields and applications including artificial intelligence (Chopra et al., 2006), bioinformatics (Burkovski et al., 2014), information retrieval (Hassanzadeh and Milenkovic, 2014), and wireless sensor networks (Fishbain and Moreno-Centeno, 2016), to name a few.

This work makes six main contributions:

- Introduction of a general distance-based methodology for the joint aggregation of ordinal and cardinal evaluations into a consensus multimodal evaluation.
- Derivation of mathematical modeling components for enforcing three different logical couplings between cardinal and ordinal evaluations, which are applicable to any suitable cardinal-ordinal distance pair.
- Construction of a ranking and rating consensus aggregation model that combines two axiomatic distances for group decision-making; the distances allow the inputs to be incomplete and contain ties and are generalizations of the Kemeny and Snell (1962) ranking distance and the Cook and Kress (1985) rating distance.
- Derivation of exact and approximate optimization models for solving the joint rating and ranking aggregation problem and computational enhancements thereof based on polyhedral theory.
- Supplementary techniques for identifying sources of high inconsistency in the evaluations and cases that may warrant further inspection (e.g., judges whose evaluations drastically contradict those of most other judges).
- Description of a case study and computational experiments on synthetic instances for illustrating the practicality of the proposed methodology.

The paper is organized as follows: Section II provides an overview of principled methodologies for ordinal and cardinal aggregation within the context of group decision-making. Section III reviews the models and the axiomatic distances used to construct the functions of disagreement featured in this work, and it defines the notions of an aggregate cardinal evaluation and an aggregate ordinal evaluation used herein. Section IV derives mathematical modeling components for enforcing three different types of logical interrelationships between an ordinal and a cardinal evaluation. Section V focuses on a multimodal aggregation problem associated with one of these couplings: the rating and ranking aggregation problem. First, it proves the intractability of this optimization problem; second, it derives a mixed integer linear programming formulation and enhancements thereof based on polyhedral theory; third it derives a convexified formulation; fourth, it introduces mechanisms for identifying inconsistencies in the given evaluations. Section VI discusses a real-world case study involving the 2007 MSOM Student Paper Competition for illustrating the practicality of the proposed methodology; it also describes computational experiments on synthetic instances motivated by the case study and other practical considerations. Finally, Section VII concludes the work.

II. Literature Review

Previous group decision-making literature has addressed either the rankings-alone aggregation problem (e.g. Kemeny and Snell 1962; Arrow 1963; Bartholdi et al. 1989), or the ratings-alone aggregation problem (e.g. Keeney 1976; Saaty 1977; Hochbaum and Levin 2006). The ranking aggregation problem has been studied extensively, especially in the social choice literature. One of the most celebrated results is Arrow's impossibility theorem (Arrow, 1963), which states that there is no "satisfactory" method to aggregate a set of rankings, where a satisfactory method is one that satisfies all of the following properties: universal domain, no imposition, monotonicity, independence of irrelevant alternatives, and non-dictatorship. In spite of this landmark result, different ranking aggregation methods have been developed to guarantee the fulfillment of a selected number of these and other properties (Brandt et al., 2016). Kemeny and Snell (1962) proposed a set of axioms that a distance metric between two complete rankings should satisfy and proved that their distance uniquely satisfies all of these axioms. Their distance measures the number of *rank reversals* between two rankings, where one rank reversal is incurred whenever two objects have

a different relative order in the given rankings, and a half rank reversal is incurred whenever two objects are tied in one ranking but not in the other. Kemeny and Snell defined the *median ranking* as the ranking that minimizes the sum of the distances to each of the input rankings; their methodology has thenceforth become synonymous with robust ranking aggregation. Specifically, this principled aggregation framework is known to ensure fairness, thwart manipulation, and mitigate individual bias in the aggregate outcome, as has been evinced in group decision-making (Truchon et al., 1998), bioinformatics (Lin, 2010), relational databases (Farah and Vanderpooten, 2007), and many other applications. Moreno-Centeno and Escobedo (2016) devised a generalization of the Kemeny-Snell ranking distance for incomplete rankings, which reduces to the original distance when the rankings are complete. In Yoo et al. (2020), the intuitiveness of this distance function was bolstered by its connection to a corresponding generalization of the Kendall-τ ranking-correlation coefficient (Kendall, 1938). It must be noted that the optimization problem that needs to be solved to find the consensus ranking via any of the above-mentioned measures is NP-hard (Bartholdi et al., 1989).

The difficulties presented by Arrow's impossibility theorem and the intractability of finding the Kemeny-Snell's aggregate ranking can be overcome by replacing ordinal rankings by (cardinal) ratings. Following this direction, Keeney (1976) proved that the *averaging method* satisfies all of Arrow's desirable properties. In the averaging method, the aggregate rating of each object is the average of the scores it receives. An immediate drawback of this approach is that it implicitly requires that all judges use the same rating scale; that is, all individuals must be equally strict or equally lenient in their score assignments. Such a standard is nearly impossible to enforce, even when evaluation rubrics are provided, as is evinced by the case study featured later in this work. Moreover, the rating aggregation approach ignores the aspect of relative pairwise comparisons, which are fundamental towards avoiding certain undesirable outcomes—e.g., an object may be selected as the winner even though a majority may consider it inferior compared with another object. Pairwise comparison intensities are the input to Saaty's Analytic Hierarchy Process technique (Saaty, 1977). There, the optimal scores are found by the principal eigenvector technique. The readers are referred to Hochbaum (2010) for an analysis of the principal eigenvector method in the context of group decision-making.

The separation-deviation model of Hochbaum (2004, 2006); Hochbaum and Levin (2006) avoids the computational difficulties of the Kemeny-Snell model and the decision quality inadequacies of the principal eigenvector method. The model takes point-wise scores and potentially also pairwise comparisons as inputs, and it is one of the building blocks of the models proposed herein. The separation-deviation optimization problem is solvable in polynomial time when all of the penalty functions are convex (Hochbaum and Levin, 2006).

III. Preliminaries

This section introduces basic notation and definitions used throughout the paper, and it reviews the concepts of the separation-deviation (SD) model, axiomatic distance between incomplete ratings, and axiomatic distance between incomplete rankings.

A. Basic Notation and Definitions

Let V be the universal set of n objects to be evaluated; without loss of generality, assign a unique identifier to each element so that $V = \{1,2,...,n\}$. There are m judges indexed by $k \in \{1,2,...,m\}$, each of who provides a (possibly incomplete) vector of scores or ratings, $\boldsymbol{a}^k$, over V. Specifically, a_j^k is the score given by judge k to object j, and a_j^k is undefined or assigned the token "$\bullet$" if judge k did not rate object j; the subset of the objects rated by judge k is written as $V_{\boldsymbol{a}}^k \subseteq V$. It is also assumed that the input ratings may contain ties. Without loss of generality, the scores are contained in a pre-specified interval $[L,U]$; the *range* of the ratings is given by $R := U - L$. Given a judge, k, and two objects, i and j, the *implied (cardinal) separation gap* (or *intensity of preference*) of i to j is

$$p_{ij}^k = \begin{cases} a_i^k - a_j^k & \text{if } i \in V_{\boldsymbol{a}}^k \text{ and } j \in V_{\boldsymbol{a}}^k \\ \text{undefined} & \text{otherwise.} \end{cases}$$

In addition, each judge $k \in \{1,2,...,m\}$ provides a (possibly incomplete) ranking vector, $\boldsymbol{b}^k$, over V. Specifically, b_j^k is the rank position (an ordinal number) given by judge k to object j, and b_j^k is undefined or assigned the token "$\bullet$" if judge k did not rank object j; the subset of the objects ranked by judge k is written as $V_{\boldsymbol{b}}^k \subseteq V$. This work also assumes that the input rankings may contain ties using the following convention of expression. When $\boldsymbol{b}$ ties all the objects in a subset $V'_{\boldsymbol{b}} \subseteq V_{\boldsymbol{b}}$ and these objects are all ranked strictly worse than $(p\text{-}1)$ other objects in $V_{\boldsymbol{b}}$, where

$p \geq 1$, then $b_i = p$ for all $i \in V'_{\boldsymbol{b}}$. Likewise, an object $j \in V_{\boldsymbol{b}} \backslash V'_{\boldsymbol{b}}$ that holds the next (worse) ranking position relative to $i \in V'_{\boldsymbol{b}}$ receives the rank $b_j = p + |V'_{\boldsymbol{b}}|$. Stated otherwise, the expression $(|V^k_{\boldsymbol{b}}| - b^k_i)$ reflects how many objects from $V^k_{\boldsymbol{b}}$ are tied or ranked worse than i (excluding itself).

Given a judge, k, and two objects, i and j, the *implied (ordinal) separation gap* (or *preference*) of i to j is $\text{sign}(b^k_i - b^k_j)$ if judge k ranked both objects, and is undefined otherwise. The sign function is defined as

$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases} \tag{1}$$

We note that, even though in the previous discussion (and throughout the paper) we assume that every judge gives both a rating and a ranking, the herein proposed methodology also applies to situations where not all judges give both types of evaluations. Consequently, the methodology is also applicable to cases where one set of judges gives only ratings and another disjoint set of judges gives only rankings.

Since the axiomatic distances used in the proposed methodology rely on comparing the scores or ranks of pairs of objects, it is convenient to define the *pairwise comparison arc set* over the object set $V_{\boldsymbol{a}^k}$ (or $V_{\boldsymbol{b}^k}$) as follows:

$$\mathcal{A}^k := \{(i,j) : i \in V^k_{\boldsymbol{a}}, j \in V^k_{\boldsymbol{a}},\ i < j\}, \tag{2}$$

($\mathcal{B}^k$ is defined analogously for the ranking data) where, by convention, the pairwise comparison arc sets contain only arcs from lower indices to higher indices to eliminate duplicate comparison pairs.

Given a rating, $\boldsymbol{a}$, we denote by $\mathsf{rank}(\boldsymbol{a})$ the ranking obtained by first sorting the objects according to their nondecreasing scores in $\boldsymbol{a}$, and then assigning to each object the ordinal number corresponding to its position in the sorted list of objects. For example, the rating $(4.5, 5, 2, \bullet, 2, 1.7)$ induces the ranking $(2, 1, 3, \bullet, 3, 5)$. Throughout the paper, it is assumed that higher ratings are superior (preferred) to lower ratings—as is customary in product reviews, student grading, and various other group decision-making contexts.

A distance-based *consensus* is defined as the optimal solution to the cardinal aggregation (CA) problem or the ordinal aggregation (OA) problem, which can be written succinctly as:

$$\text{(CA)} \quad \min_{\boldsymbol{x}} \ \sum_{k=1}^{m} d(\boldsymbol{x}, \boldsymbol{a}^k) \qquad || \qquad \text{(OA)} \quad \min_{\boldsymbol{y}} \ \sum_{k=1}^{m} d(\boldsymbol{y}, \boldsymbol{b}^k) \tag{3}$$

where the respective solution vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are assumed to be complete. This work also assumes that the solution vectors may contain ties.

B. Review of the Separation-Deviation Model

The separation-deviation (SD) model can be applied to group-decision making problems where the input is given as pairwise comparisons and/or point-wise scores. In the model formulation, the variable x_i is the aggregate score of the ith object and, thus, $(x_i - x_j)$ represents the aggregate separation gap of the ith over the jth object. A set of separation gaps p_{ij} given as inputs must be *consistent*, that is, for all triplets (h,i,j), $p_{hi} + p_{ij} = p_{hj}$. The consistency of a set of separation gaps is equivalent to the existence of a set of scores ω_i for $i = 1, \ldots, n$ such that $p_{ij} = \omega_i - \omega_j$ (Hochbaum, 2010; Hochbaum and Levin, 2006).

The mathematical programming formulation of the SD model is as follows:

$$\text{(SD)} \qquad \min_{\boldsymbol{x}} \quad \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} f^k_{ij}\Big((x_i - x_j) - p^k_{ij}\Big) + \sum_{k=1}^{m} \sum_{i=1}^{n} g^k_i\Big(x_i - a^k_i\Big) \tag{4a}$$

$$\text{subject to} \quad L \ \leq \ x_i \leq U \quad i = 1, \ldots, n \tag{4b}$$

$$x_i \ \in \ \mathbb{Z} \qquad i = 1, \ldots, n. \tag{4c}$$

The function $f^k_{ij}(\cdot)$ penalizes the difference between the aggregate separation gap and the kth reviewer's separation gap for object-pair (i,j). The function $g^k_i(\cdot)$ penalizes the difference between the aggregate score of object i and the kth reviewer's score of object i. In order to ensure polynomial-time solvability, the functions $f^k_{ij}(\cdot)$ and $g^k_i(\cdot)$ must be convex. In the context of rating aggregation, the penalty functions assume the value 0 for the argument 0; meaning that if the output separation gap for object-pair (i,j) (given by $(x_i - x_j)$) agrees with p^k_{ij}, then $f^k_{ij}((x_i - x_j) - p^k_{ij}) = f^k_{ij}(0) = 0$. If $i \notin V_{\boldsymbol{a}^k}$, then $g^k_i(\cdot)$ is set to the constant function 0; similarly, if $i \notin V_{\boldsymbol{a}^k}$ or $j \notin V_{\boldsymbol{a}^k}$, then $f^k_{ij}(\cdot)$ is set to the constant function 0. Furthermore, for linear $f^k_{ij}(\cdot)$ and $g^k_i(\cdot)$ with $L, U \in \mathbb{Z}$, the resulting problem can be solved as a linear program and $\boldsymbol{x}$ is guaranteed to be integral due to the unimodularity of the constraint coefficient matrix.

The SD problem is a special case of the convex dual of the minimum cost network flow (CDMCNF) problem (Hochbaum, 2004, 2006; Hochbaum and Levin, 2006). The most efficient algorithm known for the CDMCNF has a running time of $O(mn\log\frac{n^2}{m}\log(U-L))$ (Ahuja et al., 2003), where m is the total number of given separation gaps, and n is the number of objects. Ahuja et al. (2004) presented an alternative algorithm that uses a minimum-cut algorithm as a subroutine.

C. Axiomatic Distance Between Incomplete Ratings (Possibly with Ties)

Defining a penalty function on separation gaps is equivalent to quantifying the distance between them. Cook and Kress (1985) proposed a distance between complete ratings. This distance function was adapted to incomplete ratings in Fishbain and Moreno-Centeno (2016). It was shown that this adaption, called the *normalized projected Cook-Kress distance* (NPCK), uniquely satisfies a set of desirable metric-like axioms stated therein. Given incomplete ratings $\boldsymbol{a}^1$ and $\boldsymbol{a}^2$, the NPCK distance between the implied separation gaps is:

$$d_{NPCK}(\boldsymbol{a}^1,\boldsymbol{a}^2)=\mathcal{C}^{1,2}\sum_{i\in V_{\boldsymbol{a}}^1\cap V_{\boldsymbol{a}}^2}\sum_{j\in V_{\boldsymbol{a}}^1\cap V_{\boldsymbol{a}}^2}\left|p_{ij}^1-p_{ij}^2\right| \tag{5}$$

where

$$\mathcal{C}^{1,2}=\left(4R\cdot\left\lceil\frac{|V_{\boldsymbol{a}}^1\cap V_{\boldsymbol{a}}^2|}{2}\right\rceil\cdot\left\lfloor\frac{|V_{\boldsymbol{a}}^1\cap V_{\boldsymbol{a}}^2|}{2}\right\rfloor\right)^{-1}, \tag{6}$$

and where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ represent the floor and ceiling functions, respectively. In the above expression, $\mathcal{C}^{1,2}$ is a normalization constant that guarantees that $0\le d_{NPCK}(\boldsymbol{a}^1,\boldsymbol{a}^2)\le 1$, and $R:=U-L$ is the range of the ratings. Equivalently, since $|p_{ij}^1-p_{ij}^2|=\left|p_{ji}^1-p_{ji}^2\right|$, equation 5 can be reexpressed as:

$$d_{NPCK}(\boldsymbol{a}^1,\boldsymbol{a}^2)=2\mathcal{C}^{1,2}\sum_{(i,j)\in\mathcal{A}^1\bigcap\mathcal{A}^2}|p_{ij}^1-p_{ij}^2|. \tag{7}$$

In Fishbain and Moreno-Centeno (2016) the *aggregate rating*, $\boldsymbol{x}^*$, is the optimal solution to the *Ratings Cardinal Aggregation* problem:

$$\text{(R-CA)}\qquad \min_{\boldsymbol{x}}\sum_{k=1}^{m}d_{NPCK}(\boldsymbol{a}^k,\boldsymbol{x}). \tag{8}$$

Problem R-CA is as a special case of the SD model, and therefore it is solvable in polynomial time.

We note that $d_{NPCK}(\boldsymbol{a}^1,\boldsymbol{a}^2)=0$ and $d_{NPCK}(\boldsymbol{a}^1,\boldsymbol{a}^2)=1$ indicate that there is total agreement and total disagreement, respectively, between the ratings $\boldsymbol{a}^1$ and $\boldsymbol{a}^2$. The normalization is necessary for the distances in problem (8) to be comparable to each other even when the individuals rate different numbers of objects. The normalization constant $\mathcal{C}^{1,2}$ was chosen to address the following difficulties: (i) The numbers of objects rated by each incomplete rating may be different; therefore the distances in problem (8) are over different dimensional spaces. (ii) Each of the distance calculations in problem (8) is between a complete rating, $\boldsymbol{x}^*$, and an incomplete rating, $\boldsymbol{a}^k$—meaning, it only considers the number of objects rated by the incomplete rating. (iii) Distances in higher dimensional spaces tend to be considerably larger than distances in lower dimensional spaces—from equation (5), observe that the number of summands of each distance term $d_{NPCK}(\boldsymbol{a}^k,\boldsymbol{x})$ in problem (8) is squarely proportional to the size of $V_{\boldsymbol{a}}^k$.

D. Axiomatic Distance Between Incomplete Rankings (Possibly with Ties)

Kemeny and Snell (1962) proposed a distance between complete rankings. This distance function was adapted to incomplete rankings in Moreno-Centeno and Escobedo (2016). The authors demonstrated that this adaptation, called the *normalized projected Kemeny-Snell distance* (NPKS), uniquely satisfies a set of desirable metric-like axioms stated therein. Given incomplete rankings $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$, the NPKS distance is:

$$d_{NPKS}(\boldsymbol{b}^1,\boldsymbol{b}^2)=\mathcal{D}^{1,2}\sum_{i\in V_{\boldsymbol{b}}^1\cap V_{\boldsymbol{b}}^2}\sum_{j\in V_{\boldsymbol{b}}^1\cap V_{\boldsymbol{b}}^2}\frac{1}{4}\left|\text{sign}(b_i^1-b_j^1)-\text{sign}(b_i^2-b_j^2)\right| \tag{9}$$

$$=\mathcal{D}^{1,2}\sum_{(i,j)\in\mathcal{B}^1\bigcap\mathcal{B}^2}\frac{1}{2}\left|\text{sign}(b_i^1-b_j^1)-\text{sign}(b_i^2-b_j^2)\right| \tag{10}$$

(since $|\text{sign}(b_i^1 - b_j^1) - \text{sign}(b_i^2 - b_j^2)| = |\text{sign}(b_j^1 - b_i^1) - \text{sign}(b_j^2 - b_i^2)|$ for all i, j), where

$$\mathcal{D}^{1,2} = \left[\frac{|V_{\boldsymbol{b}}^1 \cap V_{\boldsymbol{b}}^2| \cdot \left(|V_{\boldsymbol{b}}^1 \cap V_{\boldsymbol{b}}^2| - 1\right)}{2} \right]^{-1}. \tag{11}$$

The expression inside equation (10) sum corresponds to the Kemeny and Snell (1962) distance function term associated with the ranks given to objects i and j by $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$; $\mathcal{D}^{1,2}$ is a normalization constant that guarantees that $0 \leq d_{NPKS}(\boldsymbol{b}^1, \boldsymbol{b}^2) \leq 1$. The distance $d_{NPKS}(\boldsymbol{b}^1, \boldsymbol{b}^2)$ has the following natural interpretation: The distance between two incomplete rankings is proportional to the number of *rank reversals* between them. That is, a rank reversal is incurred whenever two objects have a different relative order in the rankings $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$. Similarly, a *half rank reversal* is incurred whenever two objects are tied in one ranking but not in the other ranking. In Moreno-Centeno and Escobedo (2016) the *aggregate ranking*, $\boldsymbol{y}^*$, is the optimal solution to the *Rankings Ordinal Aggregation* problem:

$$\text{(R-OA)} \qquad \min_{\boldsymbol{y}} \sum_{k=1}^{m} d_{NPKS}(\boldsymbol{b}^k, \boldsymbol{y}). \tag{12}$$

Problem R-OA is NP-hard whether the input rankings are complete (Bartholdi et al., 1989) or incomplete (Moreno-Centeno and Escobedo, 2016).

We note that when there is total agreement between $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$ with respect to the ordinal positions of the objects ranked in common, $d_{NPKS}(\boldsymbol{b}^1, \boldsymbol{b}^2)$ is equal to 0; when there is total disagreement, their distance is equal to 1; otherwise, the distance is strictly between 0 and 1 and proportional to the level of disagreement. The normalization is necessary for the distances in problem (12) to be comparable to each other even when the individuals rank different numbers of objects. The normalization constant $\mathcal{D}^{1,2}$ was chosen to address an analog set of difficulties as $\mathcal{C}^{1,2}$ for incomplete ranking aggregation.

IV. Logical Couplings for Multimodal Aggregation

This paper seek to develop mathematical models for the joint aggregation of a set of cardinal evaluations $\{\boldsymbol{a}^k\}_{k=1}^m$ and a set of ordinal evaluations $\{\boldsymbol{b}^k\}_{k=1}^m$. The proposed consensus aggregation models are designed to find a cardinal-ordinal evaluation that least disagrees with the multimodal inputs, quantified according to an appropriate pair of (axiomatic) distances. An abbreviated form of these models can be written as:

$$\text{(COA)} \qquad \min_{\boldsymbol{x}, \boldsymbol{y}} \quad \sum_{k=1}^{m} w_C^k \cdot d_C\left(\boldsymbol{a}^k, \boldsymbol{x}\right) + \sum_{k=1}^{m} w_O^k \cdot d_O\left(\boldsymbol{b}^k, \boldsymbol{y}\right), \tag{13}$$

where, respectively, $d_C(\cdot,\cdot), d_O(\cdot,\cdot)$ denote unspecified ordinal and cardinal distance functions; parameters w_C^k, w_O^k denote weights assigned to the cardinal and ordinal information from judge k; and variable vectors $\boldsymbol{x}$, $\boldsymbol{y}$ denote the aggregate cardinal and ordinal evaluations. The full contents of these models—objective function, constraints, additional auxiliary variables—depend on the choices of distance function and the aggregate-evaluation domains (e.g., complete, with ties, etc.). They also depend on special logic that may need to be enforced (e.g., bounds on the solution values, linearized expressions, etc.). This section introduces modeling components to *couple* or logically interrelate $\boldsymbol{x}$ and $\boldsymbol{y}$. In particular, it considers three general interrelationships between cardinal and ordinal evaluations: rating and ranking preferences, numerical and ordinal estimates, and rating and approval preferences.

A. Coupling Rating and Ranking Preferences

For the decision-making context, an aggregate rating (i.e., cardinal evaluation) and an aggregate ranking (i.e., ordinal evaluation) are coupled by requiring that $\boldsymbol{y} = \mathsf{rank}(\boldsymbol{x})$. This guarantees that objects that obtain higher rating values in the consensus solution also obtain better ranking positions. The following theorem demonstrates how to enforce this logic through the addition of $O(n^2)$ linear constraints and auxiliary binary variables.

Theorem 1 (Rating-Ranking Coupling): Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n$ be complete rating and ranking vectors that allow ties. The following mixed integer linear constraints interrelate $\boldsymbol{x}$ and $\boldsymbol{y}$ so that their preferences are logically interrelated, specifically, so that objects that receive higher cardinal values also receive lower ordinal values.

$$y_i + \sum_{j \neq i} z_{ij} = n \qquad i = 1, \ldots, n \tag{14a}$$

$$x_i - x_j + 1 \leq M_1 \, z_{ij} \qquad i, j = 1, \ldots, n; i \neq j \tag{14b}$$

$$-x_i + x_j \leq M_2 (1 - z_{ij}) \qquad i, j = 1, \ldots, n; i \neq j \tag{14c}$$

$$z_{ij} \in \{0, 1\} \qquad i, j = 1, \ldots, n; i \neq j, \tag{14d}$$

where $\boldsymbol{z} \in \{0,1\}^{n^2}$ are auxiliary variables and M_1, M_2 are constants large enough so that constraint (14a) is satisfied whenever $z_{ij} = 1$ and constraint (14b) is satisfied whenever $z_{ij} = 0$, for any feasible setting of $\boldsymbol{x}$.

Proof. It is useful to give a preference interpretation to auxiliary variable z_{ij}:

$$z_{ij} = \begin{cases} 1 & \text{if object } i \text{ is preferred or tied with object } j, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Constraint (14a) provides a one-to-one relationship between variables z_{ij} and the ranking position of object $i \in V$, according to the convention for expressing rankings with ties described in Section III-A. That is, tallying the number of other objects over which i is preferred or tied, given by $\sum_{j \neq i} z_{ij}$, and subtracting this total from n provides the corresponding ranking position, y_i. Next, Constraints (14b) and (14c) enforce that $\boldsymbol{y} = \mathsf{rank}(\boldsymbol{x})$, by considering the implications of each preference modality onto the other.

First, the ordinal preferences implied by the relationships between aggregate cardinal variables x_i and x_j are encapsulated with two cases:

- **Cardinal to Ordinal Preferences, Case 1**: $x_i \geq x_j$.
  The left-hand side of Constraint (14b) is positive, which forces $z_{ij} = 1$, i.e., i is tied or ranked better than j. This setting makes the right-hand side of Constraint (14c) equal to 0; the constraint is automatically satisfied since $x_j - x_i \leq 0$.
- **Cardinal to Ordinal Preferences, Case 2**: $x_i < x_j$.
  The left-hand side of Constraint (14c) is positive, which forces $z_{ij} = 0$, i.e., i is ranked worse than j. This setting makes the right-hand side of Constraint (14b) equal to 0; the constraint is automatically satisfied since $x_i - x_j < 0$.

Second, the cardinal preferences implied by the aggregate ordinal preferences between i and j, represented by z_{ij}, are encapsulated with two cases:

- **Ordinal to Cardinal Preferences, Case 1**: $z_{ij} = 0$.
  The right-hand side of Constraint (14b) is 0, which implies that $x_i + 1 \leq x_j$ (note that (14b) becomes redundant). In other words, if i receives a worse rank than j, then it must also receive a lower rating.
- **Ordinal to Cardinal Preferences, Case 2**: $z_{ij} = 1$.
  The right-hand side of Constraint (14c) is 0, which implies that $x_j \leq x_j$. In other words, if i is tied or ranked better than j, then it must receive at least as high a rating as the latter. □

A couple of clarifications are in order. First, the assumption in Theorem 1 that the aggregate rating vector is integral is without loss of generality, since it is possible to specify any desired precision in the aggregating ratings through a special interpretation of $\boldsymbol{x}$. Expressly, bounding the elements of $\boldsymbol{x}$ as $L/\mu \leq x_i \leq U/\mu$, for $i = 1, \ldots, n$, where L and U are the lowest and highest values of the rating range and $\mu = 1/p$ is the desired score precision or minimum separation in rating values, with integer $p \geq 1$, x_i can be interpreted as the number of minimum separation gaps from L obtained by object $i \in V$ (see Section V for more details). Given this interpretation, the tightest possible "Big-M" constants for Constraints (14b) and (14c) are $M_1 = (U - L + 1)/\mu$ and $M_2 = (U - L)/\mu$, respectively. Second, notice that Constraints (14a)-(14d) are sufficient to prevent cycles in the aggregate ordinal preferences. This is because the consensus ranking positions of any three objects $h, i, j \in V$ are directly implied by the ordering of their consensus rating values (each of which cannot assume more than one cardinal value at a time).

B. Coupling Cardinal and Ordinal Estimates

Although axiomatic aggregation methods are traditionally associated with social (i.e., human) contexts, their use extends to a multitude of other situations requiring the aggregation of conflicting information from non-human sources (Mirkin and Fenner, 2019). For example, consensus aggregation has been widely used in information retrieval to derive representative lists of relevant documents in databases (Losada et al., 2018) and to perform metasearch (Dwork et al., 2001a,b) and in bioinformatics to build genetic maps (Jackson et al., 2008) and consolidate gene expression results (Li et al., 2019; Lin, 2010). Akin to the social context, consensus aggregation methods are used in these types of settings to consolidate heterogenous evaluations that may be inconsistent, unreliable, and/or biased (Brandt et al., 2016). The proposed multimodal consensus aggregation may be applicable to such contexts where cardinal and ordinal evaluations may be available. However, an important distinction is that there may be multiple choices for coupling $\boldsymbol{x}$ and $\boldsymbol{y}$, and the most appropriate of these options must be judged by the specific context at hand. The possibility that objects that receive higher cardinal values also receive lower ordinal values was covered in Section IV-A. This subsection covers another coupling that is relevant for estimating some quantifiable characteristic

exhibited by the objects, namely the requirement that objects that receive higher cardinal values also receive higher ordinal values.

Theorem 2 (Cardinal and Ordinal Estimate Coupling): Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n$ be complete cardinal and ordinal vectors that allow ties. The following mixed integer linear constraints interrelate $\boldsymbol{x}$ and $\boldsymbol{y}$ so that objects that receive higher cardinal values also receive higher ordinal values.

$$y_i + \sum_{j \neq i} z_{ij} = n \qquad i = 1, \ldots, n \tag{16a}$$

$$x_i - x_j + 1 \leq M_1(1 - z_{ij}) \qquad i, j = 1, \ldots, n; i \neq j \tag{16b}$$

$$-x_i + x_j \leq M_2\, z_{ij} \qquad i, j = 1, \ldots, n; i \neq j \tag{16c}$$

$$z_{ij} \in \{0, 1\} \qquad i, j = 1, \ldots, n; i \neq j, \tag{16d}$$

where $\boldsymbol{z} \in \{0,1\}^{n^2}$ are auxiliary variables and M_1, M_2 are constants large enough so that constraint (16b) is always satisfied when $z_{ij} = 0$ and constraint (16c) is always satisfied when $z_{ij} = 1$, for any feasible setting of $\boldsymbol{x}$.

Proof. It is useful to give a slightly different interpretation to auxiliary variable z_{ij}:

$$z_{ij} = \begin{cases} 1 & \text{if object } i \text{ exhibits a lower quantity of the observed characteristic than object } j, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

The remainder of the proof uses a similar logic as that of Theorem IV-A and is thus omitted for brevity. □

C. Coupling Rating and Approval Preferences

Approval voting is a relatively simple ordinal voting method that has received recent attention from various communities. The method seeks to divide the objects into an "approved" (i.e., winning) subset and a "disapproved" (i.e., losing) subset. An *approval ballot* is defined as a subset $V_+^k \subseteq V$ that indicates those objects judge k deems as approved; all other objects evaluated by judge k belong to a disapproved subset $V_-^k \subseteq V \backslash V_+^k$. Judge k's approval ballot can be equivalent expressed using a binary (ordinal) vector $\boldsymbol{b}^k$ defined as:

$$b_i^k = \begin{cases} 1 & \text{if object } i \text{ is approved,} \\ 0 & \text{if object } i \text{ is disapproved,} \\ \bullet & \text{if object } i \text{ is not evaluated by judge } k. \end{cases} \tag{18}$$

An aggregate approval voting vector $\boldsymbol{y} \in \{0,1\}^n$ is defined similarly, with the only difference that the last case from Equation (18) is unnecessary since the solution vector is assumed to be complete. Approval voting has been criticized for its oversimplification of the collective preferences. Here, we propose logical expressions that can be used to construct ratings and approval voting joint aggregation models. To that end, the following proposition introduces mixed integer linear constraints for coupling the collective approval voting vector $\boldsymbol{y}$ with the collective rating vector $\boldsymbol{x}$.

Theorem 3 (Rating and Approval Aggregation): Let $\boldsymbol{x} \in \mathbb{Z}^n$ be a complete rating that allows ties and $\boldsymbol{y} \in \{0,1\}^n$ be a complete approval voting vector. The addition of the following mixed integer linear constraints couples $\boldsymbol{x}$ and $\boldsymbol{y}$ so that approved objects in the aggregate evaluation receive higher cardinal values than disapproved objects.

$$x_i - x_j + 1 \leq M_1 z_{ij} \qquad i, j = 1, \ldots, n; i \neq j \tag{19a}$$

$$y_i - y_j \leq z_{ij} \qquad i, j = 1, \ldots, n; i \neq j, \tag{19b}$$

$$-y_i + y_j \leq (1 - z_{ij}) \qquad i, j = 1, \ldots, n; i \neq j, \tag{19c}$$

$$z_{ij} \in \{0, 1\} \qquad i, j = 1, \ldots, n; i \neq j, \tag{19d}$$

Proof. Constraints (19a)-(19d) reflect the required coupling logic through the use of auxiliary variables z_{ij}, which can be interpreted as in Equation (15). To demonstrate this, we evaluate the implications of $\boldsymbol{x}$ on $\boldsymbol{y}$ and then the implications of $\boldsymbol{y}$ on $\boldsymbol{x}$. First, the approval preferences implied by the relationships between aggregate cardinal variables x_i and x_j is encapsulated with two cases:

- **Cardinal to Approval, Case 1**: $x_i \geq x_j$.
  The left-hand side of Constraint (19a) is positive, which forces $z_{ij} = 1$. In turn, this setting makes the right-hand side of Constraint (19c) equal to 0, which is equivalent to requiring that $y_j \leq y_i$ (note that (19b) becomes redundant). In other words, when object i receives a rating value that is at least as good as the rating object j

receives, it is not possible simultaneously for i to be disapproved and j to be approved, i.e., we cannot have that $y_j > y_i$.

- **Cardinal to Approval, Case 2**: $x_i < x_j$.
  The left-hand size of (19a) is non-positive and, therefore, z_{ij} is allowed to assume any value, i.e., no coupling with y_i and y_j is enforced.

Second, the cardinal preferences implied by the relationships between aggregate approval voting variables y_i and y_j is encapsulated with three cases:

- **Approval to Cardinal, Case 1**: $y_i = y_j$.
  The left-hand sides of Constraints (19b) and (19c) are 0 and, therefore, z_{ij} is allowed to assume any value, i.e., no coupling with x_i and x_j is enforced.
- **Approval to Cardinal, Case 2**: $y_i = 0, y_j = 1$.
  Constraint (19c) implies that $z_{ij} = 0$. This in turn implies that $x_i + 1 \leq x_j$ in Constraint (19a). In other words, if j is approved and i is disapproved in the aggregate approval preferences, it must also be the case that j receives a higher cardinal value than i.
- **Approval to Cardinal Case 3**: $y_i = 1, y_j = 0$.
  Constraint (19b) implies that $z_{ij} = 1$, which makes constraint (19a) redundant. □

It is important to remark that because constraints are generated for $i, j = 1, \ldots, n$, Case 3 forces $z_{ji} = 0$ in the respective constraints where the labels i and j are exchanged (i.e., according to Case 2 after the exchange).

By adding Constraints (19a)-(19d) to Problem (13), it is possible to find a rating-approval consensus using a suitable pair of distances. Example distances that can be used to aggregate ratings include d_{NPCK} (discussed in Section III-C). Example distances that can be used to aggregate approval ballots include the Hamming distance (Brams et al., 2007).

V. An Axiomatic Distance-based Rating and Ranking Aggregation Model

The previous section derived linear expressions for logically interrelating different types of cardinal and ordinal evaluations. The respective constraint sets are only one part of the mathematical modeling components needed to obtain an explicit representation of the consensus aggregation problem for a specific cardinal-ordinal distance pair (see Problem (13)). To complete the associated optimization model, it is necessary to include the corresponding objective function expressions and other specialized constraints and auxiliary variables. The remainder of this paper focuses on a multimodal consensus aggregation model that combines the rating aggregation problem via distance d_{NPCK} (Problem R-CA, given by (8)) and the ordinal aggregation problem via distance d_{NPKS} (Problem R-OA, given by (12)). We denote this as the *Ratings and Rankings Cardinal and Ordinal Aggregation* problem (RR-COA).

This section is organized as follows. Section V-A introduces an abbreviated version of RR-COA and demonstrates that the problem is NP-hard. Section V-B derives an exact mixed integer linear program (MILP) reformulation to solve this problem exactly. It also proposes a strengthened version of the formulation that incorporates structural valid inequalities. Section V-C derives a convex relaxation, which serves as an efficient and effective heuristic capable of solving instances with a very large number of objects. Lastly, Section V-D describes supplementary techniques for analyzing the RR-COA solution.

A. The Ratings and Rankings Cardinal and Ordinal Aggregation Problem (RR-COA)

Problem RR-COA can be written in abbreviated form as:

$$\text{(RR-COA)} \qquad \min_{\boldsymbol{x},\boldsymbol{y}} \quad \sum_{k=1}^{m} d_{NPCK}\left(\boldsymbol{a}^k, \boldsymbol{x}\right) + \sum_{k=1}^{m} d_{NPKS}\left(\boldsymbol{b}^k, \boldsymbol{y}\right) \tag{20a}$$

$$\text{subject to} \qquad \boldsymbol{y} = \mathsf{rank}\left(\boldsymbol{x}\right) \tag{20b}$$

$$0 \leq x_i \leq \frac{U-L}{\mu} \qquad i = 1, \ldots, n \tag{20c}$$

$$x_i, y_i \in \mathbb{Z} \qquad i = 1, \ldots, n. \tag{20d}$$

It is important to elaborate on the key assumptions inherent in this formulation. The goal of this model is to give fair representation/weight to each of the evaluation inputs and to each data modality; this is enforced by giving equal weight to the cumulative d_{NPCK} distance term and to the cumulative d_{NPKS} distance term. If justified by a particular context, different weight parameters can be assigned to the two cumulative distance terms and/or to their

individual summands, as is illustrated in (13). Additionally, the parameter $\mu = 1/p$ in Problem (20) specifies the *score precision* or *minimum separation gap* in rating values of non-tied objects in the solution, where $p \in \mathbb{Z}_+$. Accordingly, x_i gives the number of minimum separation gaps from the lowest rating value, L, of object $i \in V$. That is, the consensus rating value according to the original scale is obtained via the expression $L + \mu x_i$. Higher solution precision than is specified in the rating inputs may be needed to incorporate enough separation gaps in the aggregate rating. At a minimum, μ must be sufficiently small to allow any ranking of n objects to be eligible as a solution for RR-COA (since the ranking solution is induced by ordering the $\boldsymbol{x}$ values in non-increasing order). As an example, using a scoring range of 0.0 to 10.0 and minimum precision of $\mu = .5$, it is not possible to obtain a strict ranking of more than 21 objects. Furthermore, μ should be small enough to capture large intensities of preference transitively implied through multiple pairwise comparisons. As a continuation of the previous example, if a judge evaluates only objects h and i and gives them scores of 10.0 and 0.0 (a difference of 20 minimum-separation gaps), respectively, and another judge evaluates only objects i and j and gives them scores of 10.0 and 0.0, respectively, the combination of their scores would suggest a stronger intensity of preference between objects i and h than between the two compared pairs (a difference of 40 minimum-separation gaps). However, to enhance solution interpretability the precision should not be too small.

It is also worth highlighting that, while the formulation enforces the rating and ranking coupling featured in Section IV-A via Constraint (20b), a different logical relationship between the aggregate cardinal and ordinal evaluations can be utilized if warranted in certain situations. For instance, Kemmer et al. (2020) recently applied a modified version of RR-COA (using one of the MILPs developed herein) in the context of crowdsourced computation, which is a field that studies how to combine the abilities of multiple humans to complete difficult tasks (Ipeirotis and Paritosh, 2011). The authors enforced the coupling for cardinal and ordinal estimates introduced in IV-B to perform two related but distinct crowdsourced computation tasks: ordering a set of images based on the number of dots they contain (fewest to most) and estimating the number of dots each image contains. The results therein attest that eliciting and aggregating multimodal information can improve the quality of crowdsourced estimates.

Since d_{NPCK} and d_{NPKS} are generalized versions of the Cook and Kress (1985) complete rating distance and the Kemeny and Snell (1962) complete ranking distance, respectively, problem (20) can be used to solve the complete ranking and rating aggregation problem, also previously undefined in the literature. Next, we establish that RR-COA is NP-hard by reducing it from problem (12), which is NP-hard (Bartholdi et al., 1989).

Lemma 4: Problem RR-COA is NP-hard.

Proof. Given an instance of problem (12), that is a set of incomplete rankings $\{\boldsymbol{b}^k\}_{k=1}^m$, one can transform it (in polynomial time) to an instance of RR-COA as follows. Keep $\{\boldsymbol{b}^k\}_{k=1}^m$ unchanged and create a set of ratings $\{\boldsymbol{a}^k\}_{k=1}^m$ such that each rating evaluates exactly one object (the choice of object is irrelevant; in fact, all of the ratings can evaluate the same object). From the definition of d_{NPCK} (equation (5)), it follows that, for every $\boldsymbol{x}$, the first summand in RR-COA will be equal to 0. Therefore, with this choice of ratings the optimal solution to RR-COA will be $\boldsymbol{y}^*$, that is, the optimal solution to problem (12). □

B. Deriving an Exact MILP Formulation of Problem RR-COA

The objective functions of Problems R-CA and R-OA problems are nonlinear. This subsection linearizes and combines both objectives to construct an exact mixed integer linear programming (MILP) formulation of RR-COA. It is useful to begin with Problem R-OA and to define parameters $\hat{b}_{ij}^k$ as:

$$\hat{b}_{ij}^k = \begin{cases} 1 & \text{if } b_i^k \leq b_j^k, \\ -1 & \text{if } b_i^k > b_j^k, \\ 0 & \text{if } i = j, \end{cases} \tag{21}$$

for $(i,j) \in \mathcal{B}^k$ and $k = 1, \ldots, m$. The solution to Problem R-OA can be reexpressed as:

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{k=1}^{m} d_{NPKS}\left(\boldsymbol{b}^k, \boldsymbol{y}\right) = \underset{\boldsymbol{y}}{\operatorname{argmax}} \; -2\left[\sum_{k=1}^{m} d_{NPKS}\left(\boldsymbol{b}^k, \boldsymbol{y}\right)\right] + m \tag{22a}$$

$$= \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{k=1}^{m} 1 - 2 d_{NPKS}\left(\boldsymbol{b}^k, \boldsymbol{y}\right) \tag{22b}$$

$$= \underset{\boldsymbol{z}}{\operatorname{argmax}} \sum_{k=1}^{m} \mathcal{D}^k \sum_{(i,j)\in\mathcal{B}^k} \hat{b}_{ij}^k \, z_{ij}, \tag{22c}$$

where the latter equation applies an equivalent representation of distance d_{NPKS} derived in Yoo et al. (2020). Expressly, the resulting maximization problem linearizes (10) through the introduction of parameters $\hat{b}^k_{ij}$ defined above and binary variables $z_{ij} \in \{0,1\}$, where $i,j = 1,\ldots,n$ and $k = 1,\ldots,m$ (these substitute variables can interpreted as in (15)). To yield the corresponding aggregate ranking, the equation $y_i + \sum_{j \neq i} z_{ij} = n$ is solved, for all i (see Section IV-A for more details). Next, the solution to Problem R-CA can be reexpressed as:

$$\begin{aligned}\underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{k=1}^{m} 2C^k \sum_{(i,j)\in\mathcal{A}^k} \left|\mu(x_i - x_j) - p^k_{ij}\right| &= \underset{\boldsymbol{x}}{\operatorname{argmax}} \; -2\left[\sum_{k=1}^{m} 2C^k \sum_{(i,j)\in\mathcal{A}^k} \left|\mu(x_i - x_j) - p^k_{ij}\right|\right] \\ &= \underset{\boldsymbol{x}}{\operatorname{argmax}} \sum_{k=1}^{m} -4C^k \sum_{(i,j)\in\mathcal{A}^k} t^k_{ij},\end{aligned}$$

where auxiliary variables $t^k_{ij} \geq 0$ are used to substitute the respective absolute value terms by requiring equivalently that $t^k_{ij} \geq \mu(x_i - x_j) - p^k_{ij}$ and $t^k_{ij} \geq -\mu(x_i - x_j) + p^k_{ij}$, for $(i,j) \in \mathcal{A}^k$ and $k = 1,\ldots,m$. We remark that the d_{NPCK} and d_{NPKS} normalization constants are indexed above only by a single index in contrast to their two-index definitions, (6) and (11). This simplification can be made because in the consensus aggregation problem each input rating/ranking is always compared to a complete rating/ranking (the aggregate evaluation), meaning each constant fluctuates based only on the number of objects ranked by the k^{th} judge.

From the above derivations and the results of Section IV-A, the *Base RR-COA MILP* is given by:

$$\underset{\boldsymbol{t},\boldsymbol{x},\boldsymbol{z}}{\operatorname{argmax}} \sum_{k=1}^{m} -4C^k \sum_{(i,j)\in\mathcal{A}^k} t^k_{ij} \; + \; \sum_{k=1}^{m} \mathcal{D}^k \sum_{(i,j)\in\mathcal{B}^k} \hat{b}^k_{ij}\, z_{ij} \tag{23a}$$

$$\begin{aligned}
\text{subject to} \quad t^k_{ij} - \mu(x_i - x_j) &\geq -p^k_{ij} && (i,j) \in \mathcal{A}^k, k = 1,\ldots,m && \text{(23b)}\\
t^k_{ij} + \mu(x_i - x_j) &\geq p^k_{ij} && (i,j) \in \mathcal{A}^k, k = 1,\ldots,m && \text{(23c)}\\
x_i - x_j &\leq M_1 z_{ij} - 1 && i,j = 1,\ldots,n; i \neq j, && \text{(23d)}\\
-x_i + x_j &\leq M_2(1 - z_{ij}) && i,j = 1,\ldots,n; i \neq j, && \text{(23e)}\\
x_i &\leq \tfrac{U-L}{\mu} && i = 1,\ldots,n && \text{(23f)}\\
z_{ij} &\in \{0,1\} && i,j = 1,\ldots,n; i \neq j, && \text{(23g)}\\
x_i &\in \mathbb{Z}^+_{\cup\{0\}} && i = 1,\ldots,n. && \text{(23h)}
\end{aligned}$$

Additionally, we seek to enhance the computational performance of RR-COA MILP through the incorporation of structural valid inequalities (VIs). The insight behind the VIs is linked with the preference relations that are guaranteed by pairs and triplets of variables z_{ij}. More specifically, the following linear expressions are satisfied by any setting of variables z_{ij} that induces a complete non-strict ranking of n objects (Yoo and Escobedo, 2020):

$$\begin{aligned}
z_{ij} + z_{ji} &\geq 1 \quad i,j = 1,\ldots,n; \quad i \neq j && \text{(24a)}\\
z_{ij} - z_{kj} - z_{ik} &\geq -1 \quad i,j,k = 1,\ldots,n; \quad i \neq j \neq k \neq i. && \text{(24b)}
\end{aligned}$$

Briefly stated, these expressions enforce the properties of a weak ordering, that is a binary relation that is reflexive, transitive, and total. We denote the resulting formulation as the *Enhanced RR-COA MILP* and evaluate its comparative performance with the Base RR-COA MILP in Section VI. It is worth adding that (24a) and (24b) are in fact logically equivalent expressions of two of the three members of the *basic family* of facet defining inequalities of the weak order polytope (the convex hull of the characteristic vectors induced by all weak orders on n objects); the third member of this family are the upper bound constraints $z_{ij} \leq 1$. Note that these VIs represent only a subset of all facet defining inequalities of the polytope known to date (e.g., see Doignon et al. (2014); Fiorini (2003); Fiorini and Fishburn (2004)).

C. Convex Relaxation of Problem RR-COA

It is useful to return to the original (nonlinear, nonconvex) formulation of RR-COA, which can be written as:

$$\min_{\boldsymbol{x}} \sum_{k=1}^{m} \left[2C^k \sum_{(i,j)\in\mathcal{A}^k} \left|\mu(x_i - x_j) - p^k_{ij}\right|\right] + \sum_{k=1}^{m} \left[\frac{1}{2}\mathcal{D}^k \sum_{(i,j)\in\mathcal{B}^k} \left|\operatorname{sign}(x_i - x_j) - \operatorname{sign}(b^k_j - b^k_i)\right|\right] \tag{25a}$$

$$\text{subject to} \quad 0 \le x_i \le \frac{U-L}{\mu} \qquad i=1,\ldots,n \tag{25b}$$

$$x_i \in \mathbb{Z} \qquad i=1,\ldots,n. \tag{25c}$$

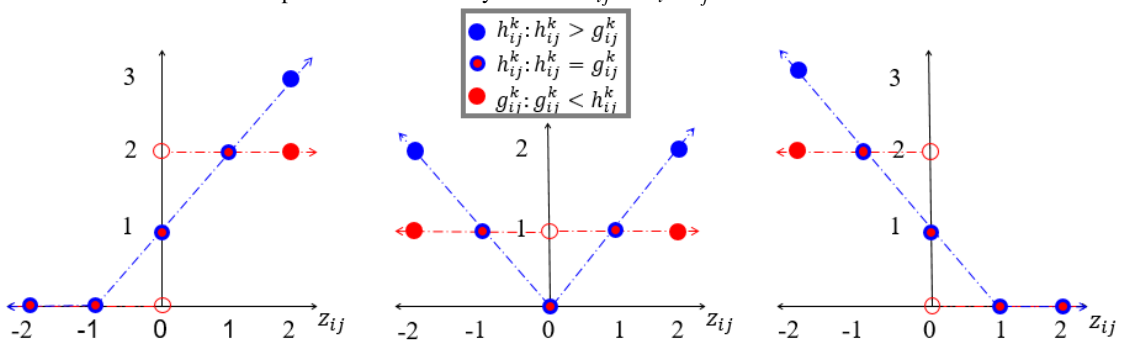
Notice that the above formulation replaces the ordinal vector $\boldsymbol{y}$ and cardinal-ordinal coupling expressions in Formulation (23) with terms $\text{sign}(x_i - x_j)$ in the objective function. Additionally, the argument of the other sign function in the objective function is $(b_j^k - b_i^k)$ and not $(b_i^k - b_j^k)$, as in equation (10). This modified argument matches the rating-ranking coupling discussed in Section IV-A, which requires that higher cardinal numbers are assigned to objects judged as more preferable, while higher ordinal numbers are assigned to objects judged as less preferable.

As the preceding subsection demonstrates, each cardinal-aggregation term in (25a) is a convex piecewise linear term that is easily linearized. By contrast, each ordinal-aggregation term in (25a) is highly nonconvex and nonlinear (see (10)). This subsection proposes to approximate the function $g^k(x_i,x_j) := |\text{sign}(x_i - x_j) - \text{sign}(b_j^k - b_i^k)|$, the k^{th} summand of the second sum in (25a), with a tight upper-convex (piecewise-linear) envelope, $h^k(x_i,x_j)$:

$$h^k(x_i,x_j) = \begin{cases} \max\{0, x_i - x_j + 1\} & \text{if } \text{sign}(b_j^k - b_i^k) = -1 \\ \max\{-x_i + x_j, x_i - x_j\} & \text{if } \text{sign}(b_j^k - b_i^k) = 0 \\ \max\{-x_i + x_j + 1, 0\} & \text{if } \text{sign}(b_j^k - b_i^k) = 1. \end{cases} \tag{26}$$

Figure 1 shows that $h^k(x_i,x_j)$ approximates $g^k(x_i,x_j)$ and provides a tight convex envelope.

Fig. 1: Relationships between $g^k(x_i,x_j)$ (shorthand g_{ij}^k) and its upper convex envelope $h^k(x_i,x_j)$ (shorthand h_{ij}^k), for each of the three possible values of $\text{sign}(b_j^k - b_i^k)$. For ease of illustration, the domain $(x_i,x_j) \in \mathbb{Z}^2$ is projected into a 1-dimensional space via the auxiliary variable $z_{ij} := x_i - x_j \in \mathbb{Z}^1$



(a) If $\text{sign}(b_j^k - b_i^k) = -1$ (b) If $\text{sign}(b_j^k - b_i^k) = 0$ (c) If $\text{sign}(b_j^k - b_i^k) = 1$

Replacing $g^k(x_i,x_j)$ with $h^k(x_i,x_j)$ and linearizing the cardinal-aggregation terms in the objective function yields the following convex relaxation of the RR-COA problem, denoted as c-RR-COA:

$$\min_{\boldsymbol{x}} \sum_{k=1}^{m} 2C^k \sum_{(i,j)\in\mathcal{A}^k} t_{ij}^k \;+\; \sum_{k=1}^{m} \frac{1}{2} D^k \sum_{(i,j)\in\mathcal{B}^k} h_{ij}^k \tag{27a}$$

$$\text{subject to} \quad t_{ij}^k - \mu(x_i - x_j) \ge -p_{ij}^k \qquad (i,j)\in\mathcal{A}^k, k=1,\ldots,m \tag{27b}$$

$$t_{ij}^k + \mu(x_i - x_j) \ge p_{ij}^k \qquad (i,j)\in\mathcal{A}^k, k=1,\ldots,m \tag{27c}$$

$$h_{ij}^k - x_i + x_j \ge 1 \qquad (i,j)\in\mathcal{B}^k, k=1,\ldots,m;\ \text{s.t. } \text{sign}(b_j^k - b_i^k) = -1 \tag{27d}$$

$$h_{ij}^k - x_i + x_j \ge 0 \qquad (i,j)\in\mathcal{B}^k, k=1,\ldots,m;\ \text{s.t. } \text{sign}(b_j^k - b_i^k) = 0 \tag{27e}$$

$$h_{ij}^k + x_i - x_j \ge 0 \qquad (i,j)\in\mathcal{B}^k, k=1,\ldots,m;\ \text{s.t. } \text{sign}(b_j^k - b_i^k) = 0 \tag{27f}$$

$$h_{ij}^k + x_i - x_j \ge 1 \qquad (i,j)\in\mathcal{B}^k, k=1,\ldots,m;\ \text{s.t. } \text{sign}(b_j^k - b_i^k) = 1 \tag{27g}$$

$$h_{ij}^k \ge 0 \qquad (i,j)\in\mathcal{B}^k, k=1,\ldots,m \tag{27h}$$

$$t_{ij}^k \ge 0 \qquad (i,j)\in\mathcal{A}^k, k=1,\ldots,m \tag{27i}$$

$$0 \le x_i \le \tfrac{U-L}{\mu},\ \text{integer} \qquad i=1,\ldots,n. \tag{27j}$$

Problem (27) is a special case of the convex SD model and, therefore, it is solvable in polynomial time.

D. Supplementary Analyses from the RR-COA Solution

Next, we propose a mechanism to identify inconsistencies in the given evaluations (e.g. outliers, judges that are too lenient or too strict, etc.). This information may be helpful, for instance, for the lead decision maker to initiate an investigation of the nature of unusual discrepancies and to justify further deliberations (e.g., discussing these inconsistencies with the judges and promoting a discussion with the objective of alleviating them).

The mechanism uses the ratings solution to RR-COA, denoted as $\boldsymbol{x}^{(RR)}$, to identify (i) judges whose evaluations differ the most with the rest of the evaluations and (ii) objects about which judges had particularly divergent evaluations. These judges (objects) are those that assigned (received) scores that disagree the most with $\boldsymbol{x}^{(RR)}$. Specifically, we use the individual contributions to the separation penalty to identify the judges whose evaluations are farthest from $\boldsymbol{x}^{(RR)}$. The contribution of judge $\underline{k} \in \{1,\ldots,m\}$ to the separation penalty is calculated as:

$$2C^{\bar{k}} \sum_{(i,j)\in\mathcal{A}^{\underline{k}}} \left|(x_i^{(RR)} - x_j^{(RR)}) - (a_i^{\underline{k}} - a_j^{\underline{k}})\right|. \tag{28}$$

Similarly, we use the separation penalty to identify the objects that engendered particularly divergent evaluations. These objects are those with the highest contribution to the separation penalty. The contribution of object $\bar{i}$ to the separation penalty is calculated as:

$$\sum_{k=1}^{m} \left[\sum_{(\underline{i},j)\in\mathcal{A}^k} 2C^k \left|(x_{\underline{i}}^{(RR)} - x_j^{(RR)}) - (a_{\underline{i}}^k - a_j^k)\right| + \sum_{(j,\underline{i})\in\mathcal{A}^k} 2C^k \left|(x_j^{(RR)} - x_{\underline{i}}^{(RR)}) - (a_j^k - a_{\underline{i}}^k)\right| \right]. \tag{29}$$

VI. Computational Tests and Analysis

This section assesses various practical dimensions of the featured aggregation methodology. To this end, it first considers a real-world case study involving the 2007 MSOM Student Paper Competition, written succinctly as 2007 MSOM SPC. Afterwards, it introduces a procedure for generating synthetic instances motivated by the response styles and other practical considerations seen in the 2007 MSOM SPC dataset. This then allows for more comprehensive computational analyses of the featured methodology. The experiments were performed on machines equipped with 36GB of RAM memory shared by two Intel Xeon E5-2680 processors running at 2.40 GHz. The code was written in Python, and the optimization models were solved with CPLEX version 12.9.

A. Analysis of 2007 MSOM SPC

This subsection considers the case study of the 2007 MSOM SPC, which consists of 58 submitted papers and 63 participating judges. In the competition, each judge evaluated only three to five papers, and each paper was reviewed by only three to five judges. Although the input evaluations are highly incomplete, the paper-judge allocation of this instance is considered to be robust (see next subsection for more details). It is worthwhile to note that some of the input ratings tended to use the entire (i.e., an expanded) rating scale, others tended to use only the middle scores (i.e., a condensed scale), others tended to use only the top scores (i.e., an optimistic scale), and yet others tended to use only the bottom scores (i.e., a pessimistic scale). Such diverse response styles align with a large number of real-world studies (see, for example, Baumgartner and Steenkamp 2001; Smith 2004; Harzing 2006). Moreover, there were self-contradictions between the input ranking evaluations and the rankings induced by the input rating evaluations of individual judges.

The papers were rated based on six different attributes (see Table I), gauged according to a precise numerical rubric (see Table II). The rubric was set according to the respective qualities of papers published in three top-tier domain journals (heretofore referred to as *the journals*): *Manufacturing & Service Operations Management* (MSOM), *Operations Research* (OR), and *Management Science* (MS). Each judge also provided an ordinal evaluation (a ranking) of the papers he/she reviewed (1 = best, 2 = second best, etc.), allowing for ties.

TABLE I: Description of evaluation attributes with respective numerical scales

Attribute	Description	Scale
(A)	Problem importance/interest	1–10
(B)	Problem modeling	0–10
(C)	Analytical results	0–10
(D)	Computational results	0–10
(E)	Paper writing	1–10
(F)	Overall contribution to the field (field contribution, for short)	1–10

TABLE II: Numerical score rubric (the journals: MSOM, OR, and MS)

Score	Definition / Interpretation
10	Attribute considered is comparable to that of the best papers published in the journals.
8,9	Attribute considered is comparable to that of the average papers published in the journals.
7	Attribute considered is at the minimum level for publication in the journals.
5,6	Attribute considered independently would require a minor revision before publication in the journals.
3,4	Attribute considered independently would require a major revision before publication in the journals.
1,2	Attribute considered would warrant by itself a rejection if the paper were submitted to the journals.
0	Attribute considered is not relevant or applicable to the paper being evaluated.

Although this precise rubric was provided to the judges, they nevertheless differed significantly in their evaluations and presumably interpreted the scores differently. Examples of this phenomenon are illustrated for papers 18 and 26 in Tables III and IV, respectively. We remark that labels identifying judges and papers have been randomly permuted from their original assignments in order to preserve the participants' anonymity.

TABLE III: Evaluations of paper 18

	Attribute Ratings						Paper
Judge	(A)	(B)	(C)	(D)	(E)	(F)	Ranking
3	3	3	4	0	2	3	4
4	5	3	4	0	5	3	4
11	6	4	4	0	5	4	4
12	7	8	8	0	7	8	2
42	2	2	2	0	3	2	4

TABLE IV: Evaluations of paper 26

	Attribute Ratings						Paper
Judge	(A)	(B)	(C)	(D)	(E)	(F)	Ranking
21	8	10	8	8	5	8	3
24	8	9	8	10	7	8	1
14	7	2	3	2	2	2	5
26	8	8	7	8	8	7	3
49	10	7	6	9	9	8	1

As Table III illustrates, in Attributes (B), (C), and (F), paper 18 was given a score of 8 by one judge—i.e., the paper is comparable in problem modeling, analytical results, and field contribution to an average paper in the journals—and scores no greater than 4 by the other four judges—i.e., the paper requires at best a major revision. Such scoring discrepancies are not insignificant and are especially pronounced between judges 12 and 42. For the most part, the former considers the paper to be at the average level for publication, while the latter holds it should be rejected by the journals. A similar discrepancy in subjective judgments involving paper 26 can be seen in Table IV. Therein, judge 14's evaluations do not appear to be on the same scale as the evaluations of the other four judges. In particular, in every attribute except for (A), judge 14's evaluation indicates that the paper would be rejected; on the other hand, in all attributes the other judges deemed the paper worthy of publication—some of them even indicated it would be among the best papers published in the journals! Such glaring discrepancies in the judges' evaluations over the same papers are commonplace throughout this data set. Henceforth, we use the average scores over all six attributes, excepting scores of 0 (which connote lack of relevance rather than poor quality), as the input ratings of each judge.

Table V compares the optimal solutions obtained by the three aggregation models: i) $\boldsymbol{x}^*$, obtained by aggregating only the ratings via problem (8); ii) $\boldsymbol{y}^*$, obtained by aggregating only the rankings via problem (12); and iii) $\boldsymbol{x}^{(RR)}$ and $\mathsf{rank}\big(\boldsymbol{x}^{(RR)}\big)$, obtained by jointly aggregating the ratings and rankings via problem (20).

TABLE V: Aggregate evalutions for 2007 MSOM SPC returned by different aggregation models

Paper	$\boldsymbol{x}^*$	$\boldsymbol{y}^*$	$\boldsymbol{x}^{(\mathbf{RR})}$	$\mathsf{rank}\big(\boldsymbol{x}^{(\mathbf{RR})}\big)$	Paper	$\boldsymbol{x}^*$	$\boldsymbol{y}^*$	$\boldsymbol{x}^{(\mathbf{RR})}$	$\mathsf{rank}\big(\boldsymbol{x}^{(\mathbf{RR})}\big)$	Paper	$\boldsymbol{x}^*$	$\boldsymbol{y}^*$	$\boldsymbol{x}^{(\mathbf{RR})}$	$\mathsf{rank}\big(\boldsymbol{x}^{(\mathbf{RR})}\big)$
1	6.1	51	7.62	51	21	7.4	45	8.18	46	41	8.4	10	9.40	15
2	8.0	19	9.13	17	22	6.6	39	8.21	40	42	8.2	33	8.69	33
3	7.5	47	8.07	48	23	7.6	50	7.63	49	43	8.3	28	8.87	26
4	7.0	29	8.22	38	24	7.9	22	8.88	25	44	7.1	14	8.91	23
5	6.5	52	7.33	55	25	8.1	21	9.11	19	45	7.3	41	8.21	40
6	9.0	10	9.51	13	26	8.6	24	9.00	22	46	8.6	9	9.60	10
7	7.8	43	8.20	44	27	10.0	14	9.50	14	47	8.8	8	9.58	11
8	8.6	14	9.01	21	28	6.9	52	7.61	52	48	8.7	10	9.40	15
9	8.2	26	8.87	26	29	7.8	34	8.70	30	49	9.1	5	9.72	4
10	8.6	7	9.61	9	30	7.1	30	8.71	29	50	6.1	58	7.11	57
11	8.3	6	9.62	8	31	8.0	43	8.20	44	51	8.4	24	8.89	24
12	7.5	38	8.29	37	32	8.8	20	9.12	18	52	7.3	47	8.60	36
13	9.5	22	9.10	20	33	7.5	34	8.21	40	53	7.5	34	8.70	30
14	8.8	14	9.71	7	34	8.9	3	9.72	4	54	9.1	4	9.99	2
15	7.3	41	8.21	40	35	7.6	34	8.61	35	55	7.2	52	7.61	52
16	6.8	52	7.60	54	36	8.1	26	8.87	26	56	8.7	2	9.72	4
17	8.8	18	9.58	11	37	6.8	39	8.22	38	57	8.7	1	10.00	1
18	5.3	49	7.63	49	38	5.8	57	7.12	56	58	7.9	31	8.70	30
19	7.9	32	8.69	33	39	8.9	10	9.98	3					
20	6.0	52	6.71	58	40	7.3	46	8.08	47					

As table V demonstrates, there are many conflicts between the ratings-only ($\boldsymbol{x}^*$) and the rankings-only ($\boldsymbol{y}^*$) solutions; for example, while paper 27 attains the top position in the R-CA solution, it is only judged as 14th best in the R-OA solution. Such outcomes can be explained in part by the different qualities that are encapsulated through each input evaluation format in the SPC. Cardinal evaluations measure average performance across all attributes and implicitly capture judges' intensities of preference between papers with respect to these averages. Conversely, ordinal evaluations capture the net preferences between papers, effectively allowing each judge to weigh and condense performance from the individual attributes differently based on what he/she regards as most relevant. At the same time, differences in assigned ranks generally do not capture intensities of preference (e.g., the preference for the first-ranked over the second-ranked paper may be marginal, but the preference for either over the third-ranked paper may be substantial). Due to the different qualities encapsulated through the two input evaluation formats, it was not uncommon in this dataset for a judge to rank a paper that performs well over all six rating attributes, but not exceptionally on any single one, lower than a paper that performs exceptionally on certain key attributes, but comparatively worse on average over all six attributes. The featured multimodal aggregation approach yields a rating-ranking pair that minimizes cumulative disagreement with the two types of input evaluations but is devoid of such conflicts.

It is important to remark that $\boldsymbol{x}^*$ and $\boldsymbol{x}^{(RR)}$ in Table V have a higher precision than the individual attribute ratings (0.5 was the highest precision given in the attribute scores). As Section V-A explains, added precision is necessary to incorporate enough separation gaps in the aggregate rating. From a modeling point of view, this does not represent a problem since R-CA and RR-COA can be solved to any rating precision, which is specified a priori via $\mu > 0$; herein, this parameter was set to $\mu = .01$.

Next we give a specific example of objects/papers whose aggregate score in $\boldsymbol{x}^*$ and aggregate rank in $\boldsymbol{y}^*$ are in conflict. For instance, paper 54 has a relatively high aggregate score of 9.1, but it conflicts with others (e.g., paper 57) that have a lower aggregate score but a higher aggregate rank. Table VI gives the evaluations received by papers 54 and 57, table VII gives the number of papers reviewed by their respective judges and the average rating these judges gave over all their assigned papers, and table VIII gives each paper's adjusted rating, obtained by dividing the paper's rating by the respective judge's average rating. From these tables we observe the following:

1) The ranking evaluations assigned to paper 57 seem slightly better than those assigned to paper 54.
2) The rating evaluations assigned to paper 54 were lower in magnitude than those assigned to paper 57. However, juxtaposing the paper ratings with the average rating from each respective judge suggests there was a stronger intensity of preference for paper 54 over the papers against which it was compared than there was for paper 57. Indeed, note that the top-3 adjusted ratings of the former are greater than those of the latter.
3) The lowest ranking of paper 57 was 2 while that of paper 54 was 4. Moreover, paper 54 received a below-average paper rating while paper 57 did not.

All of this suggests that paper 57 slightly edges out paper 54 when the ranking and rating evaluations are considered jointly. Indeed, in the combined aggregate rating-ranking pair, $\boldsymbol{x}^{(RR)}$ and $\mathsf{rank}(\boldsymbol{x}^{(RR)})$ (the solution to RR-COA), paper 57 is rated and ranked slightly higher than paper 54; this, as discussed previously, seems appropriate. In contrast, the aggregate rating $\boldsymbol{x}^*$ rates paper 54 higher than 57. This provides evidence that the combined rating-ranking solution (which jointly aggregates the multimodal evaluations) more effectively represents the judges' multimodal evaluations than the aggregate rating (which takes into consideration only the ratings).

TABLE VI: Evaluations of papers 54 and 57

Paper	Judge	Paper Rating	Paper Ranking
54	22	7.3	1
54	25	7.0	1
54	30	6.2	1
54	32	4.6	4
57	16	7.0	1
57	17	7.4	1
57	32	7.4	1
57	57	6.3	2
57	62	6.0	1

TABLE VII: Statistics of judges who evaluated papers 54 and 57

Judge	# of Papers Evaluated	AVG Rating
22	4	4.95
25	5	5.30
30	5	4.96
32	4	6.13
16	3	6.73
17	4	5.00
32	4	6.13
57	4	5.93
62	4	5.15

TABLE VIII: Adjusted rating received by papers 54 and 57

Paper	Judge	Adjusted Rating
54	22	1.47
54	25	1.32
54	30	1.25
54	32	0.75
57	16	1.04
57	17	1.46
57	32	1.21
57	57	1.06
57	62	1.17

Papers 14, 18, and 50 had the top-three (object-wise) contributions to the separation penalty. As noted previously

and illustrated in Table III, paper 18 elicited polarized responses: four judges gave it a very low evaluation and one judge a very high evaluation. In such a situation, it may be prudent to further deliberate on the assigned scores/ranks. Judges 44, 18, and 24 had the top-three (judge-wise) contributions to the separation penalty. This information suggests that these judges expressed relatively unpopular opinions. For instance, table IX shows that judge 44 assigned a near-perfect rating of 9.7 to paper 42 and a relatively low rating of 5.3 to paper 14 (second-worst on the judge's list), even though the solutions to R-CA, R-OA, and RR-COA all rated and/or ranked paper 42 significantly worse than paper 14. Additionally, judge 44's ratings and rankings of papers 45 and 56, whose respective evaluations are shown in tables X and XI, appear to be at odds with the evaluations of all other judges who reviewed them, and, consequently, are also at odds with the aggregate evaluations.

TABLE IX: Evaluations of judge 44

Paper	Paper Rating	Paper Ranking
14	5.3	4
38	4.2	5
42	9.7	1
45	8.3	2
56	8.3	3

TABLE X: Evaluations of paper 45

Judge	Paper Rating	Paper Ranking
23	5.2	2
33	4.4	3
40	6.2	2
43	5.0	2
44	8.3	2

TABLE XI: Evaluations of paper 56

Judge	Paper Rating	Paper Ranking
5	6.8	1
37	8.0	1
44	8.3	3
51	7.8	1

A potential promising line of inquiry is to examine how insights like those in the preceding two paragraphs could be used to determine the appropriateness of the initial paper-to-judge evaluation assignment and/or the existence of conflicts of interest or manipulative judges. It would also be interesting to determine when the outputs from these analyses should lead to further deliberations on divergent evaluations and what specific processes can be employed. However, while these questions are relevant, they are outside the scope of this paper and are left for future work.

B. Description of Synthetic Instances

Synthetic instances consist of joint ranking and rating evaluations with varying degrees of collective similarity. Individual input rankings are sampled from an adaptation of the Mallows ϕ-distribution of ranking data (Mallows, 1957). The standard ϕ-distribution is parameterized by a reference (i.e., ground-truth) complete strict ranking $\underline{\boldsymbol{b}}$ and dispersion $\phi \in (0,1]$, which quantify the probability of observing a complete strict ranking $\boldsymbol{b}$ as:

$$P(\boldsymbol{b}) = P(\boldsymbol{b}|\underline{\boldsymbol{b}},\phi) = \frac{1}{Z}\phi^{d_\tau(\boldsymbol{b},\underline{\boldsymbol{b}})},$$

where $d_\tau(\cdot,\cdot)$ signifies the Kendall (1938) distance (equivalent to d_{KS} when $\boldsymbol{b},\underline{\boldsymbol{b}}$ are complete strict rankings) and $Z = \Sigma_{\boldsymbol{b}}\ \phi^{d_\tau(\boldsymbol{b},\underline{\boldsymbol{b}})} = (1)\times(1+\phi)\times(1+\phi+\phi^2)\times\cdots\times(1+\cdots+\phi^{n-1})$ is a normalization constant. It is important to point out that setting $\phi = 1$ yields the (discrete) uniform distribution over the space of complete strict rankings and setting it nearer to 0 centers the distribution mass closer to $\underline{\boldsymbol{b}}$ (Lu and Boutilier, 2014). In other words, ϕ can be said to control the proximity of $\boldsymbol{b}$ to $\underline{\boldsymbol{b}}$ and the collective similarity of multiple rankings within a sample. It can also be said to control the difficulty of the generated instances, specifically, computation times tend to increase with ϕ. The featured experiments use these synthetic instances to assess the aggregation methodology's ability to recover an aggregate ranking that is close to the underlying ground truth $\underline{\boldsymbol{b}}$ as collective similarity weakens.

We sample instances of complete and incomplete strict rankings based on the repeated insertion model introduced by Doignon et al. (2004). Since this sampling approach is not readily applicable for incomplete rankings, we utilize an extension developed in Yoo et al. (2020) to sample from smaller projected spaces. Specifically, assuming the object set to be ranked by the k^{th} judge ($V_{\boldsymbol{b}}^k$) has been predetermined, $\boldsymbol{b}^k$ is generated according to the ϕ-distribution parameterized by $(\underline{\boldsymbol{b}}|_{V_{\boldsymbol{b}}^k}, \phi_{\boldsymbol{b}^k})$, with $b_i^k = \bullet$ for all $i \in V \backslash V_{\boldsymbol{b}}^k$—that is, $\underline{\boldsymbol{b}}|_{V_{\boldsymbol{b}}^k}$ and $\boldsymbol{b}^k|_{V_{\boldsymbol{b}}^k}$ are complete strict rankings in the projected space, and the latter of these rankings is extended to the full set of objects by assigning null values to the unranked objects ($V \backslash V_{\boldsymbol{b}}^k$). The ground truth ranking vector $\underline{\boldsymbol{b}}$ is fixed to $(1,2,\ldots,n)$ in all the generated instances. Accordingly, the projected ground truth used to generate incomplete ranking $\boldsymbol{b}^k$ is given by $\underline{\boldsymbol{b}}|_{V_{\boldsymbol{b}}^k} = (1,2,\ldots,|V_{\boldsymbol{b}}^k|)$.

Individual input ratings are generated using reference rating vectors and a rating error parameter. The rating scale $[L,U]$ of all inputs is $[1.0,10.0]$. Assuming the object set to be rated by the k^{th} judge ($V_{\boldsymbol{a}}^k$) has been predetermined, the reference rating vector $\underline{\boldsymbol{a}}^k|_{V_{\boldsymbol{a}}^k}$ is set proportional to the ground truth ranking vector $\underline{\boldsymbol{b}}$ based on the number of objects rated ($|V_{\boldsymbol{a}}^k|$) and on an assigned response style. Motivated by the characteristics of the 2007 MSOM SPC dataset, four response styles are defined: expanded, condensed, optimistic and pessimistic. The first two styles

differ in the expansiveness of their ranges, but each contains a balanced number of high and low *rating markers* (i.e., reference rating values); the last two styles share the same range magnitude, but each contains an unbalanced number of high or low rating markers. Table XII lists the reference rating vectors defined for sizes $|V_{\boldsymbol{a}}^k| = 4,5,6,7,8$ for each of the four response styles. Reference rating vectors of size $|V_{\boldsymbol{a}}^k| \geq 8$ are set by assigning rating markers from $|V_a^k| = 7$ to multiple objects. Objects 1 to $\lfloor \frac{|V_a^k|}{7} \rceil$ are set to the first rating marker from the respective column under $|V_a^k| = 7$, objects $\lfloor \frac{|V_a^k|}{7} \rceil$ to $\lfloor \frac{2|V_a^k|}{7} \rceil$ are set to the second marker, etc. The error parameter ε is used to introduce random deviations from the reference rating markers and is defined as:

$$\varepsilon = 1.5 * \text{rand}(\{1.0, 1.5\}),$$

where $\text{rand}(\{1.0, 1.5\})$ selects one of the two scaling factors with equal probability. Given the generated ranking $\boldsymbol{b}^k$, the rating $\boldsymbol{a}^k$ is generated as follows. The k^{th} judge is first assigned one of the four responses styles. Next, the objects in $V_{\boldsymbol{a}}^k = V_{\boldsymbol{b}}^k$ are sorted based on their ascending order in $\boldsymbol{b}^k$. For each ranking position $i = 1, \ldots, |V_{\boldsymbol{b}}^k|$, an error ε is sampled and the object that the k^{th} judge ranks in position i receives the rating value $\underline{a}_i^k + U(-\varepsilon, \varepsilon)$ (i.e., the deviation term follows a continuous uniform distribution based on the sampled error parameter). The ensuing subsection describes additional details for generating the rating and ranking aggregation instances.

	$\|V_a^k\| = 4$				$\|V_a^k\| = 5$				$\|V_a^k\| = 6$				$\|V_a^k\| = 7$				$\|V_a^k\| = 8$			
Rank	**E**	**C**	**O**	**P**	**E**	**C**	**O**	**P**	**E**	**C**	**O**	**P**	**E**	**C**	**O**	**P**	**E**	**C**	**O**	**P**
1	9.0	7.5	9.5	7.5	9.0	7.5	9.5	7.5	9.0	7.5	9.5	7.5	9.0	7.5	9.5	7.5	9.0	7.5	9.5	7.5
2	7.0	6.5	7.5	5.5	7.0	6.5	7.5	5.5	8.0	7.0	8.5	6.5	8.0	7.0	8.5	6.5	8.0	7.0	8.5	6.5
3	4.0	4.5	5.5	3.5	5.5	5.5	6.5	4.5	7.0	6.5	7.5	5.5	7.0	6.5	7.5	5.5	7.0	6.5	7.5	5.5
4	2.0	3.5	3.5	1.5	4.0	4.5	5.5	3.5	4.0	4.5	5.5	3.5	5.5	5.5	6.5	4.5	5.5	5.5	6.5	4.5
5					2.0	3.5	3.5	1.5	3.0	4.0	4.5	2.5	4.0	4.5	5.5	3.5	5.5	5.5	6.5	4.5
6									2.0	3.5	3.5	1.5	3.0	4.0	4.5	2.5	4.0	4.5	5.5	3.5
7													2.0	3.5	3.5	1.5	3.0	4.0	4.5	2.5
8																	2.0	3.5	3.5	1.5

TABLE XII: Setting of the (projected) ground truth rating $\underline{\boldsymbol{a}}^k|_{V_{\boldsymbol{a}}^k}$ for sizes $|V_{\boldsymbol{a}}^k| = 4,5,6,7,8$ and four response styles: Expanded (**E**), Contracted (**C**), Optimistic (**O**), Pessimistic (**P**)

Before proceeding, we discuss two practical considerations of ranking and/or rating aggregation instances and related metrics herein implemented to assess them. First, it is necessary for the evaluation assignments to be allocated to judges in such a way that a direct or indirect comparison between every pair of objects in V is possible. The robustness of the object-to-judge allocation can be measured according to the number of *hops* or the length in the sequence pairwise comparisons needed to obtain an implied comparison between two objects (Hochbaum and Levin, 2010). As an example of this concept, for $h, i, j \in V$, if h and i are compared by one judge, i and j are compared by a different judge, and no single judge compares h and j, then there is one hop between h and i, one hop between i and j and two hops between h and j. The robustness and reliability of the consensus aggregation solution decreases as the maximum number of hops between the object pairs increases; an allocation with a maximum hop of one over all $i, j \in V$ is ideal, and an allocation with two maximum hops is also robust. Second, it is possible for a judge's rating and ranking inputs to be in conflict with another, meaning that an object i is simultaneously ranked better and rated worse than object j (or vice versa). To quantify the degree of individual contradiction, we define the *inner distance* of judge k as the d_{KS} distance between $\boldsymbol{b}^k$ and $\text{rank}(\boldsymbol{a}^k)$ (the ranking obtained by sorting the values of $\boldsymbol{a}^k$ in non-increasing order). Note that no such contradictions can occur between the rating and ranking solution returned by RR-COA based on the enforced coupling discussed in Section IV-A. It is worth mentioning that for the 2007 MSOM SPC data set the maximum number of hops is 2 and the average inner distance is 0.06.

C. Experiments on Synthetic Instances

The first experiment seeks to carry out a basic computational comparison of the Base RR-COA MILP (given by (23a)-(23h)) and the Enhanced RR-COA MILP (given by (23a)-(23h),(24a),(24b)). The formulations are tested on instances of incomplete non-strict rankings/ratings and on instances of complete non-strict rankings/ratings. The experiment primarily evaluates the impact of different number of objects ($n = |V|$) and dispersion levels (ϕ) on the solution times. For each defined combination of these two parameter values, 32 different instances are generated and solved by each of the two formulations within a two-hour time limit. The generated rankings do not contain ties but the output rankings are allowed to contain ties. For simplicity, the number ratings/rankings of each instance is set to $m = \lfloor 1.75n \rfloor$ and the expanded response style is assigned to all of the generated rating vectors.

For the set of incomplete non-strict rankings/ratings instances, the size of the individual evaluation subset $V_{\boldsymbol{a}}^k = V_{\boldsymbol{b}}^k$ is drawn from the discrete uniform distribution $U(4,8)$ for every k; the specific objects in the evaluation subset are selected at random from universal set V. The tested numbers of objects and dispersion values are $n \in \{40,45,50,55\}$ and $\phi \in \{.1,.2,\ldots,1.0\}$, respectively. Table XIII reports the average instance and solution statistics obtained over 32 repetitions of each tested n and ϕ-value. The instance statistics are maximum number of hops and inner distance (label d_{KS}^{INNER}). The solution statistics are wall-clock time in seconds (label Times (s)), d_{KS} distance between the aggregate ranking and the ground truth ranking (label d_{KS}^{GT}), and the relative optimality gap percentage (label Gap%).

The same experiment is repeated on instances consisting of all complete non-strict rankings and complete ratings as the generated inputs; table XIV reports these results (the maximum number of hops is exactly 1.0 for all instances and is thereby omitted from the table). The tested numbers of objects and dispersion values for these instances are $n \in \{60,70,80,90\}$ and $\phi \in \{.2,.3,\ldots,.9\}$, respectively. The narrower selection of dispersion values was selected to exclude the easiest and hardest instances; in particular, most complete RR-COA instances with $\phi = 1.0$ did not finish solving due to memory errors. For the tested parameter settings, the number of instances unsolved due to memory errors are reported in the last column of Table XIV (label Instances Exited); no such errors occurred for the incomplete RR-COA instances.

TABLE XIII: Average statistics for incomplete rating and ranking instances

n	ϕ	Max Hops	d_{KS}^{INNER}	Times (s)		Gap%		d_{KS}^{GT}	
				Base	**Enhanced**	**Base**	**Enhanced**	**Base**	**Enhanced**
40	0.1	2.00	0.075	77.07	49.52	0.006	0.009	0.02	0.02
	0.2	2.00	0.074	68.83	62.93	0.008	0.008	0.04	0.04
	0.3	2.00	0.077	66.15	94.10	0.008	0.009	0.06	0.06
	0.4	2.00	0.075	163.84	154.16	0.010	0.009	0.10	0.10
	0.5	2.00	0.075	445.54	195.76	0.009	0.008	0.13	0.13
	0.6	2.00	0.075	2343.38	560.54	2.263	0.006	0.18	0.18
	0.7	2.00	0.073	4406.53	871.19	3.637	0.007	0.25	0.25
	0.8	2.00	0.075	5317.86	1439.75	2.567	0.008	0.32	0.32
	0.9	2.00	0.077	6092.94	1738.68	3.955	0.009	0.41	0.42
	1.0	2.00	0.073	6517.68	1984.64	4.894	0.009	0.51	0.51
45	0.1	2.00	0.075	52.80	73.08	0.010	0.009	0.03	0.03
	0.2	2.00	0.076	715.36	96.16	0.010	0.009	0.04	0.04
	0.3	2.00	0.075	110.39	254.10	0.009	0.009	0.07	0.07
	0.4	2.00	0.076	503.33	580.48	0.010	0.009	0.10	0.10
	0.5	2.00	0.076	2445.59	776.05	0.079	0.009	0.13	0.13
	0.6	2.00	0.075	5001.70	2105.43	31.246	0.007	0.17	0.17
	0.7	2.00	0.077	6962.57	4302.36	49.958	1.042	0.24	0.24
	0.8	2.00	0.076	7203.52	5636.27	14.895	1.205	0.32	0.33
	0.9	2.00	0.075	7204.67	5076.80	15.045	0.740	0.39	0.41
	1.0	2.00	0.076	7203.55	5572.64	15.126	0.637	0.47	0.50
50	0.1	2.00	0.075	58.67	162.87	0.008	0.009	0.03	0.03
	0.2	2.00	0.077	139.61	330.80	0.009	0.009	0.05	0.05
	0.3	2.00	0.077	351.85	711.22	0.010	0.009	0.06	0.06
	0.4	2.00	0.078	2229.86	2052.67	0.034	0.009	0.09	0.09
	0.5	2.00	0.078	5054.74	3893.04	1.433	0.236	0.13	0.13
	0.6	2.00	0.076	7148.99	5401.88	92.161	6.741	0.19	0.19
	0.7	2.00	0.077	7203.93	6347.51	36.472	3.546	0.24	0.25
	0.8	2.00	0.076	7203.43	7029.63	29.933	6.045	0.31	0.32
	0.9	2.00	0.076	7204.47	7018.61	22.687	3.318	0.43	0.42
	1.0	2.00	0.078	7204.74	7147.53	28.312	5.700	0.51	0.49
55	0.1	2.00	0.076	89.02	293.27	0.01	0.009	0.03	0.03
	0.2	2.00	0.078	234.97	574.27	0.01	0.01	0.04	0.04
	0.3	2.00	0.075	1484.05	1268.65	0.045	0.009	0.07	0.07
	0.4	2.00	0.076	5128.91	3553.84	0.452	0.076	0.09	0.09
	0.5	2.00	0.077	7047.71	6348.02	7.371	2.005	0.13	0.13
	0.6	2.00	0.077	7204.67	7202.82	37.666	30.129	0.19	0.19
	0.7	2.00	0.077	7203.76	7203.47	57.053	18.282	0.25	0.25
	0.8	2.00	0.076	7203.22	7201.87	37.287	12.634	0.31	0.33
	0.9	2.00	0.075	7203.74	7200.38	31.481	32.287	0.41	0.41
	1.0	2.00	0.075	7201.96	7200.25	31.788	24.976	0.52	0.42

A few notable observations can be drawn from Tables XIII and XIV. First, the ability of the models to recover the ground truth ranking diminishes as ϕ increases. We remark that the d_{KS}^{GT} values attained by the two formulations coincide when both achieve optimality, but they differ when either or both formulations return a suboptimal solution upon reaching the two-hour time limit. Second, the inner distances of incomplete RR-COA instances are roughly half the value of the complete RR-COA instances; a simple explanation for this difference is that there are more

TABLE XIV: Average statistics for complete rating and ranking instances

n	ϕ	d_{KS}^{INNER}	Times (s)		Gap%		d_{KS}^{GT}		Instances Exited	
			Base	**Enhanced**	**Base**	**Enhanced**	**Base**	**Enhanced**	**Base**	**Enhanced**
60	0.2	0.139	178.02	517.69	0.009	0.008	0.000	0.000	0	0
	0.3	0.140	208.44	386.03	0.009	0.009	0.000	0.000	0	0
	0.4	0.139	169.93	321.14	0.009	0.009	0.000	0.000	0	0
	0.5	0.139	146.59	327.42	0.010	0.008	0.000	0.000	0	0
	0.6	0.140	195.1	307.12	0.010	0.008	0.000	0.000	0	0
	0.7	0.140	260.1	427.19	0.009	0.009	0.001	0.001	0	0
	0.8	0.139	1116.11	1229.68	0.010	0.009	0.005	0.005	0	0
	0.9	0.140	6845.16	6288.98	0.064	0.044	0.020	0.020	0	0
70	0.2	0.140	710.66	1651.74	0.009	0.008	0.000	0.000	0	0
	0.3	0.141	637.05	1416.27	0.009	0.009	0.000	0.000	0	0
	0.4	0.140	615.92	1366.3	0.008	0.008	0.000	0.000	0	0
	0.5	0.140	466.15	1015.6	0.009	0.008	0.000	0.000	0	0
	0.6	0.140	504.17	915.82	0.009	0.009	0.000	0.000	0	0
	0.7	0.141	932.61	1258.47	0.010	0.009	0.001	0.001	0	0
	0.8	0.140	2112.26	2568.42	0.010	0.010	0.004	0.004	0	0
	0.9	0.141	7130.7	6510.05	0.037	0.031	0.016	0.016	7	1
80	0.2	0.141	4454.02	4200.32	0.019	0.007	0.000	0.000	0	0
	0.3	0.140	3672.85	3920.3	0.010	0.006	0.000	0.000	0	0
	0.4	0.140	2870.4	3358.8	0.009	0.007	0.000	0.000	0	0
	0.5	0.140	1807.71	2963.99	0.009	0.009	0.000	0.000	0	0
	0.6	0.140	2058.19	2187.84	0.010	0.009	0.000	0.000	1	0
	0.7	0.141	2561.92	3284.15	0.010	0.010	0.001	0.001	2	0
	0.8	0.141	4036.47	5517.98	0.012	0.012	0.003	0.003	4	0
	0.9	0.141	6990.61	7206.75	0.028	0.045	0.013	0.013	4	0
90	0.2	0.141	6589.1	7130.26	0.038	0.101	0.000	0.000	8	6
	0.3	0.140	7003.75	7142.12	0.040	0.084	0.000	0.000	6	5
	0.4	0.141	7122.33	7024.69	0.034	0.041	0.000	0.000	4	4
	0.5	0.140	6760.81	6777.71	0.024	0.020	0.000	0.000	7	5
	0.6	0.140	5382.17	5956.99	0.013	0.012	0.000	0.000	4	3
	0.7	0.141	4524.02	6147.29	0.011	0.012	0.000	0.000	3	3
	0.8	0.141	5936.78	6765.65	0.013	0.017	0.002	0.002	2	1
	0.9	0.140	7216.56	7216.34	0.053	0.124	0.011	0.011	10	8

possibilities for an individual's evaluations to be contradictory when more objects are evaluated. Third, computation times tend to increase with n and ϕ, and incomplete RR-COA instances tend to be more difficult to solve than complete RR-COA instances even though the latter had higher inner distance values. Fourth, the base formulation outperforms the enhanced formulation over most of the complete RR-COA instances, while the latter outperforms the former over most of the incomplete RR-COA instances. This is partly justified by the higher difficulty of incomplete RR-COA instances, for which incorporating the valid inequalities seems to be worthwhile. In fact, the enhanced formulation has a comparable performance to the base formulation over complete RR-COA instances with higher ϕ values, but the base formulation has a markedly inferior performance for a sizable portion of the incomplete RR-COA instances.

The second experiment aims to assess the value of multimodal aggregation and to compare the computational performance of the featured optimization models. To that end, it tests the ability of RR-COA, c-RR-COA, and R-CA to recover the ground truth ranking from the generated set of noisy rating and/or ranking inputs; the R-CA consensus ranking is obtained from the non-increasing ordering of the consensus rating. The R-OA model is not tested since it possesses an inherent advantage for this task. Moreover, the ability to recover the ground truth rating is not assessed since it is not well defined—it depends on the response styles assigned to each judge.

The three models are tested on instances of incomplete non-strict rankings/ratings similar to those of the first experiment. The main difference of these instances is how the rating response styles are assigned. Two distinct rating response profiles are considered. The first profile apportions the pessimistic response style to 55% of the judges and the expanded, condensed, and the optimistic response styles equally to the remaining 45% of the judges. The second profile is similar to the first, with the difference that it apportions the condensed response style to 55% of the judges and the expanded, optimistic, and pessimistic response styles equally to the remaining 45% of the judges. Another key difference from the previous experiment is that the numbers of judges (i.e., input evaluations) varies, specifically $m \in \{30, 40, 70, 90\}$, primarily to test the effects of different object-to-judge allocations on the ability to recover the ground truth. Additional minor differences are that the number of objects is fixed to $n = 40$ and the tested dispersion values are slightly narrowed to $\phi \in \{.1, .2, \ldots, 0.8\}$, both to allow for all models to solve to optimality within two hours.

TABLE XV: Average statistics for incomplete rating and ranking instances with 55% pessimistic judges

m	ϕ	d_{KS}^{INNER}	Max Hops	Times (s)			d_{KS}^{GT}		
				RR-COA	**c-RR**	**R-CA**	**RR-COA**	**c-RR**	**R-CA**
30	0.1	0.062	2.78	64.83	0.11	0.05	0.063	0.065	0.116
	0.2	0.062	2.94	66.08	0.08	0.05	0.089	0.101	0.135
	0.3	0.064	2.94	77.09	0.09	0.06	0.125	0.140	0.168
	0.4	0.063	2.75	94.12	0.09	0.05	0.161	0.170	0.195
	0.5	0.061	2.88	153.37	0.09	0.05	0.218	0.220	0.234
	0.6	0.062	2.88	245.55	0.10	0.06	0.254	0.257	0.266
	0.7	0.063	2.94	473.16	0.10	0.05	0.325	0.325	0.328
	0.8	0.063	2.88	701.02	0.10	0.05	0.381	0.379	0.378
40	0.1	0.064	2.25	88.32	0.13	0.10	0.048	0.057	0.108
	0.2	0.063	2.31	70.40	0.12	0.11	0.064	0.083	0.115
	0.3	0.063	2.28	83.28	0.11	0.13	0.102	0.126	0.139
	0.4	0.063	2.38	123.87	0.12	0.13	0.132	0.159	0.168
	0.5	0.064	2.28	236.88	0.12	0.12	0.182	0.204	0.200
	0.6	0.064	2.28	687.10	0.13	0.12	0.242	0.256	0.253
	0.7	0.062	2.22	1542.45	0.13	0.12	0.284	0.305	0.299
	0.8	0.065	2.45	2872.76	0.15	0.14	0.373	0.375	0.376
70	0.1	0.064	2.00	45.66	0.16	0.19	0.025	0.043	0.077
	0.2	0.063	2.00	58.16	0.16	0.18	0.044	0.077	0.091
	0.3	0.063	2.00	68.18	0.19	0.19	0.062	0.105	0.105
	0.4	0.063	2.00	125.18	0.18	0.21	0.097	0.140	0.131
	0.5	0.066	2.00	600.00	0.19	0.19	0.134	0.173	0.153
	0.6	0.063	2.00	1587.61	0.19	0.17	0.188	0.219	0.200
	0.7	0.066	2.00	4728.64	0.22	0.19	0.236	0.266	0.247
	0.8	0.063	2.00	6079.27	0.23	0.20	0.316	0.339	0.331
90	0.1	0.063	2.00	38.58	0.24	0.23	0.017	0.040	0.066
	0.2	0.063	2.00	50.20	0.24	0.24	0.033	0.069	0.078
	0.3	0.062	2.00	63.96	0.28	0.26	0.050	0.098	0.090
	0.4	0.063	2.00	152.89	0.27	0.23	0.075	0.131	0.114
	0.5	0.063	2.00	430.42	0.31	0.26	0.117	0.166	0.141
	0.6	0.063	2.00	1740.19	0.31	0.29	0.159	0.206	0.174
	0.7	0.063	2.00	4737.39	0.25	0.28	0.214	0.255	0.227
	0.8	0.063	2.00	6524.59	0.25	0.29	0.298	0.317	0.296

Tables XV and Table XV report the results of instances generated with the first response profile (with 55% pessimistic judges) and with the second response profile (with 55% condensed judges), respectively (we remark that two other profiles with 55% expanded judges and 55% optimistic judges were tested, but they yielded similar results and are thus ommitted). The first general observation is that the d_{KS}^{GT} values decrease as m increases, owing both to a larger amount of information and to the better object-to-judge assignments resulting from the added evaluations. In fact, the values for instances with $m = 30, 40$ and $\phi = 0.1$ are approximately equal to those with $m = 70, 90$ and $\phi = 0.4$. All instances with the two highest m values achieve a maximum number of 2.0 hops, matching the object-to-judge assignment robustness of the 2007 MSOM SPC data set—their respective inner distances are also similar. The maximum number of hops for instances with lower m values are between 2 and 3 on average. It is also worth remarking that computation times also decreased going from the two lowest to the two highest m values.

The RR-COA model (solved with the enhanced formulation) is the best of the three models at recovering the ground truth ranking, particularly for instances with $\phi \leq 0.5$. Its performance over instances with higher dispersion values is less pronounced, which can be explained by the fact that the abilities of various voting methods becomes less distinguishable when there is little to no consensus in the data (Ali and Meilă, 2012; Mao et al., 2013; Young, 1988). The c-RR-COA model (with abbreviated label c-RR in the tables) significantly outperforms the R-CA model in this respect, which is particularly impressive since the computational times of the two models are virtually identical. These two models solved all problems in under a second while the RR-COA model almost reached the two-hour time limit as m and ϕ increased. This suggests c-RR-COA is an efficient and effective heuristic for solving large-scale RR-COA instances.

TABLE XVI: Average statistics for incomplete rating and ranking instances with 55% condensed judges

m	ϕ	d_{KS}^{INNER}	Max Hops	Times (s)			d_{KS}^{GT}		
				RR-COA	**c-RR**	**R-CA**	**RR-COA**	**c-RR**	**R-CA**
30	0.1	0.083	2.91	81.33	0.12	0.06	0.071	0.073	0.130
	0.2	0.082	2.94	75.87	0.10	0.05	0.098	0.099	0.146
	0.3	0.084	2.81	80.39	0.10	0.06	0.128	0.139	0.171
	0.4	0.080	2.88	116.41	0.11	0.06	0.166	0.176	0.205
	0.5	0.081	2.78	176.23	0.11	0.05	0.216	0.220	0.238
	0.6	0.081	2.66	331.21	0.11	0.06	0.254	0.268	0.280
	0.7	0.083	2.88	334.89	0.11	0.04	0.329	0.326	0.340
	0.8	0.082	2.91	708.91	0.12	0.07	0.362	0.363	0.373
40	0.1	0.084	2.34	71.87	0.11	0.11	0.046	0.058	0.107
	0.2	0.082	2.28	72.47	0.10	0.10	0.071	0.092	0.124
	0.3	0.084	2.44	88.82	0.11	0.12	0.102	0.120	0.146
	0.4	0.083	2.28	122.27	0.11	0.11	0.141	0.168	0.184
	0.5	0.082	2.31	237.32	0.13	0.12	0.181	0.203	0.213
	0.6	0.083	2.28	449.05	0.13	0.13	0.235	0.249	0.250
	0.7	0.081	2.31	956.32	0.14	0.13	0.295	0.311	0.307
	0.8	0.083	2.28	1206.03	0.18	0.19	0.353	0.360	0.361
70	0.1	0.084	2.00	48.73	0.17	0.20	0.026	0.047	0.086
	0.2	0.083	2.00	52.47	0.18	0.20	0.043	0.074	0.101
	0.3	0.082	2.00	68.30	0.19	0.21	0.062	0.102	0.109
	0.4	0.081	2.00	132.83	0.19	0.19	0.094	0.141	0.134
	0.5	0.081	2.00	486.87	0.18	0.18	0.133	0.170	0.159
	0.6	0.082	2.00	1097.31	0.23	0.21	0.184	0.218	0.202
	0.7	0.083	2.00	3512.05	0.22	0.22	0.231	0.263	0.251
	0.8	0.082	2.00	4702.84	0.21	0.18	0.322	0.335	0.324
90	0.1	0.083	2.00	47.83	0.28	0.27	0.015	0.039	0.076
	0.2	0.081	2.00	55.39	0.31	0.28	0.031	0.071	0.086
	0.3	0.085	2.00	72.10	0.32	0.29	0.049	0.095	0.096
	0.4	0.083	2.00	134.16	0.33	0.30	0.076	0.130	0.121
	0.5	0.084	2.00	331.46	0.34	0.34	0.112	0.159	0.142
	0.6	0.083	2.00	1196.42	0.39	0.37	0.159	0.202	0.180
	0.7	0.083	2.00	4052.48	0.24	0.27	0.208	0.255	0.232
	0.8	0.083	2.00	5551.90	0.27	0.29	0.302	0.319	0.304

VII. Concluding Remarks

We propose a distance-based methodology for jointly aggregating cardinal and ordinal evaluations. The methodology is designed to find a multimodal consensus evaluation, that is a logically coupled cardinal and ordinal evaluation pair that minimizes the sum of the distances to the multimodal inputs. Linearized expressions are introduced to enforce three types of logical couplings, and different optimization models are derived to solve the rating and ranking aggregation variant of the methodology, which is demonstrated to be an NP-hard problem. The effectiveness of the new methodology for distributed decision-making is illustrated through a case study involving the 2007 MSOM Student Paper Competition and through synthetic instances with controllable degrees of collective similarity motivated by this case study and by other practical considerations. We provide evidence that obtaining a combined aggregate cardinal and ordinal evaluation better represents the judges' opinions as compared to a consensus rating that aggregates only the judges' cardinal evaluations or a consensus ranking that aggregates only the judges' ordinal evaluations.

The proposed methodology is founded on axiomatic distances based on social choice theory, and it is designed to adequately deal with highly incomplete evaluations and other challenging aspects of group decision-making. Aggregating incomplete evaluations is challenging because the aggregate evaluation is especially prone to be biased by the judges' subjective scales; for example, objects assigned to a particularly strict (lenient) judge have an advantage (disadvantage) compared to those objects not assigned to this specific judge. Our methodology can identify such inconsistencies in the given evaluations. This information is helpful in that the lead decision maker can initiate an investigation of the nature of the conflicts and act accordingly (for example, by having the specific judges discuss and possibly resolve these inconsistencies).

The problem of aggregating complete evaluations (in which all judges evaluate all objects) is a special case of the problem of aggregating incomplete evaluations (in which the judges are allowed to evaluate only some of the objects). Therefore the methodology is also applicable to aggregating complete multimodal evaluations. The proposed methodology may also applicable to various other contexts outside of group decision-making where cardinal and ordinal evaluations over a set of entities can be obtained. For instance, Kemmer et al. (2020) recently applied a modified version of the exact multimodal aggregation model developed herein in the context of crowdsourced computation. Their results demonstrated that eliciting and aggregating incomplete multimodal estimates can improve the quality of collective estimates and the efficiency by which they are obtained.

The code used to generate the synthetic aggregation instances is available upon request.

ACKNOWLEDGEMENTS

A preliminary version of some of the results are contained in an unpublished version of this manuscript entitled, "Joint aggregation of cardinal and ordinal evaluations with an application to a student paper competition", which can be retrieved on arXiv.com.

The authors acknowledge Research Computing at Arizona State University for providing computing resources that have contributed to the research results reported within this paper. The first and third authors gratefully acknowledge funding from the Army Research Office under grant W911NF1910260 and from NSF under grant 1850355.

REFERENCES

H. B. Mitchell, *Data fusion: concepts and ideas*. Springer Science & Business Media, 2012.

D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, *Handbook of computational social choice*. Cambridge University Press, 2016.

W. D. Cook, "Distance-based and ad hoc consensus models in ordinal preference ranking," *European Journal of Operational Research*, vol. 172, no. 2, pp. 369–385, 2006.

F. F. Hassanzadeh and O. Milenkovic, "An axiomatic approach to constructing distances for rank comparison and aggregation," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6417–6439, 2014.

A. Ammar and D. Shah, "Ranking: Compare, don't score," in *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE, 2011, pp. 776–783.

H. P. Young, "Condorcet's theory of voting," *American Political science review*, vol. 82, no. 04, pp. 1231–1244, 1988.

S. Chopra, A. Ghose, and T. Meyer, "Social choice theory, belief merging, and strategy-proofness," *Information Fusion*, vol. 7, no. 1, pp. 61–79, 2006.

A. Burkovski, L. Lausser, J. M. Kraus, and H. A. Kestler, "Rank aggregation for candidate gene identification," in *Data Analysis, Machine Learning and Knowledge Discovery*. Springer, 2014, pp. 285–293.

B. Fishbain and E. Moreno-Centeno, "Self calibrated wireless distributed environmental sensory networks," *Scientific reports*, vol. 6, p. 24382, 2016.

J. G. Kemeny and L. J. Snell, *Preference ranking: An axiomatic approach*, ser. Mathematical Models in Social Science. Boston, MA: Ginn, 1962, pp. 9–23.

W. D. Cook and M. Kress, "Ordinal ranking with intensity of preference," *Management Science*, vol. 31, pp. 26–32, 1985.

K. J. Arrow, *Social Choice and Individual Values*. New York: Wiley, 1963.

J. Bartholdi, C. A. Tovey, and M. A. Trick, "Voting schemes for which it can be difficult to tell who won the election," *Social Choice and Welfare*, vol. 6, pp. 157–165, Apr. 1989. [Online]. Available: http://dx.doi.org/10.1007/BF00303169

R. L. Keeney, "A group preference axiomatization with cardinal utility," *Management Science*, vol. 23, pp. 140–145, Oct. 1976.

T. Saaty, "A scaling method for priorities in hierarchical structures," *Journal of Mathematical Psychology*, vol. 15, no. 3, pp. 234–281, 1977.

D. S. Hochbaum and A. Levin, "Methodologies and algorithms for group-rankings decision," *Management Science*, vol. 52, pp. 1394–1408, 2006. [Online]. Available: http://portal.acm.org/citation.cfm?id=1246175.1246184\&coll=GUIDE\&dl=G

M. Truchon *et al.*, "An extension of the condorcet criterion and kemeny orders," *Cahier*, vol. 9813, 1998.

S. Lin, "Rank aggregation methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 555–570, 2010.

M. Farah and D. Vanderpooten, "An outranking approach for rank aggregation in information retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 591–598.

E. Moreno-Centeno and A. R. Escobedo, "Axiomatic aggregation of incomplete rankings," *IIE Transactions*, vol. 48, no. 6, pp. 475–488, 2016.

Y. Yoo, A. Escobedo, and K. Skolfield, "A new correlation coefficient for comparing and aggregating non-strict and incomplete rankings," *European Journal of Operational Research*, vol. 285, no. 3, pp. 1025–1041, 2020.

M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

D. S. Hochbaum, "The separation, and separation-deviation methodology for group decision making and aggregate ranking," in *TutORials in Operations Research*, J. J. Hasenbein, Ed. Hanover, MD: INFORMS, 2010, vol. 7, pp. 116–141.

——, "50th anniversary article: Selection, provisioning, shared fixed costs, maximum closure, and implications on algorithmic methods today," *Management Science*, vol. 50, pp. 709–723, 2004.

——, "Ranking sports teams and the inverse equal paths problem," in *Internet and Network Economics, Second International Workshop, WINE 2006*. Greece: Springer, 2006, pp. 307–318.

R. K. Ahuja, D. S. Hochbaum, and J. B. Orlin, "Solving the convex cost integer dual network flow problem," *Management Science*, vol. 49, pp. 950–964, 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=970414.970469

——, "A cut-based algorithm for the nonlinear dual of the minimum cost network flow problem," *Algorithmica*, vol. 39, pp. 189–208, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=996555.996557\&coll=GUIDE\&dl=GUIDE

B. Mirkin and T. I. Fenner, "Distance and consensus for preference relations corresponding to ordered partitions," *Journal of Classification*, vol. 36, no. 2, pp. 350–367, 2019.

D. E. Losada, J. Parapar, and A. Barreiro, "A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation," *Information Fusion*, vol. 39, pp. 56–71, 2018.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on the World Wide Web*. New York, NY, USA: ACM, 2001, pp. 613–622.

——, "Rank aggregation revisited," 2001.

B. N. Jackson, P. S. Schnable, and S. Aluru, "Consensus genetic maps as median orders from inconsistent sources," *IEEE/ACM Transactions on computational biology and bioinformatics*, vol. 5, no. 2, pp. 161–171, 2008.

X. Li, X. Wang, and G. Xiao, "A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications," *Briefings in bioinformatics*, vol. 20, no. 1, pp. 178–189, 2019.

S. J. Brams, D. M. Kilgour, and M. R. Sanver, "A minimax procedure for electing committees," *Public Choice*, vol. 132, no. 3-4, pp. 401–420, 2007.

R. Kemmer, Y. Yoo, A. R. Escobedo, and R. Maciewjewski, "Enhancing collective estimates by aggregating cardinal and ordinal inputs," in *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2020.

P. G. Ipeirotis and P. K. Paritosh, "Managing crowdsourced human computation: a tutorial," in *Proceedings of the 20th international conference companion on World wide web*, 2011, pp. 287–288.

Y. Yoo and A. R. Escobedo, "A new binary programming formulation and social choice property for Kemeny rank aggregation," *Under second round of review*, 2020, [Available at `optimization-online.org/DB_HTML/2020/08/7958.html`].

J.-P. Doignon, S. Fiorini, and S. Rexhep, "Facets of order polytopes," in *2014 International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2014, pp. 093–097.

S. Fiorini, "A combinatorial study of partial order polytopes," *European Journal of Combinatorics*, vol. 24, no. 2, pp. 149–159, 2003.

S. Fiorini and P. C. Fishburn, "Weak order polytopes," *Discrete mathematics*, vol. 275, no. 1-3, pp. 111–127, 2004.

H. Baumgartner and J. E. M. Steenkamp, "Response styles in marketing research: A cross-national investigation." *Journal of Marketing Research*, vol. 38, pp. 143–156, 2001.

P. B. Smith, "Acquiescent response bias as an aspect of cultural communication style," *Journal of Cross-Cultural Psychology*, vol. 35, pp. 50–61, 2004.

A. W. K. Harzing, "Response styles in cross-national survey research: A 26-country study," *International Journal*

of Cross Cultural Management, vol. 6, pp. 243–266, 2006.

C. L. Mallows, "Non-null ranking models. i," *Biometrika*, vol. 44, no. 1/2, pp. 114–130, 1957.

T. Lu and C. Boutilier, "Effective sampling and learning for mallows models with pairwise-preference data." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3783–3829, 2014.

J.-P. Doignon, A. Pekeč, and M. Regenwetter, "The repeated insertion model for rankings: Missing link between two subset choice models," *Psychometrika*, vol. 69, no. 1, pp. 33–54, 2004.

D. S. Hochbaum and A. Levin, "How to allocate review tasks for robust ranking," *Acta informatica*, vol. 47, no. 5-6, pp. 325–345, 2010.

A. Ali and M. Meilă, "Experiments with kemeny ranking: What works when?" *Mathematical Social Sciences*, vol. 64, no. 1, pp. 28–40, 2012.

A. Mao, A. D. Procaccia, and Y. Chen, "Better human computation through principled voting." in *AAAI*, 2013.