Top-k List Aggregation: Mathematical Formulations and Polyhedral Comparisons

Sina Akbari, Adolfo R. Escobedo

School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281 USA
{Sina.Akbari, adres}@asu.edu

Abstract

Top-k lists are being increasingly utilized in various fields and applications including information retrieval, machine learning, and recommendation systems. Since multiple top-k lists may be generated by different algorithms to evaluate the same set of entities or system of interest, there is often a need to consolidate this collection of heterogeneous top-k lists to obtain a more robust and coherent list. This work introduces various exact mathematical formulations of the top-k list aggregation problem under the generalized Kendall tau distance. Furthermore, the strength of the proposed formulations is analyzed from a polyhedral point of view.

Keywords— Top-k list aggregation, rank aggregation, Kendall tau distance, mixed integer programming, polyhedral analysis

1 Introduction

Top-k lists are a special form of item orderings (i.e., rankings) wherein out of n total items only a small number of them, k, are explicitly ordered. Top-k lists have many advantages that can overcome some of the practical drawbacks of the traditional full-list approach: a collection of items may be too large to rank or even present, processing the full list could present a massive computational/cognitive load, and it may be impossible or meaningless to compare and rank items beyond a certain point [7]. Examples of top-k lists are the top-250 movies on IMDB or the top-10 played songs on Spotify [22].

Due to the increased use of such lists, the top-k list aggregation problem (TOP-k-AGG) has attracted considerable attention. TOP-k-AGG seeks to find a top-k list or full list that

best represents the input lists. This problem has been utilized in many different applications, including recommender systems [20], metasearch engines [12], and bioinformatics [17]. TOP-k-AGG is interrelated with many other problems such as top-k recommendation and top-k query.

TOP-k-AGG falls under the umbrella of the more general rank aggregation problem whose objective is to combine individual rankings over a set of items into one representative collective ranking [5]. Variants of this problem have been studied probabilistically [6, 8] and deterministically [10, 12]. In the probabilistic approach, it is assumed that the observed rankings are realizations of a probabilistic model on ranking data, such as Mallows model [16], and the goal is to recover the ground-truth ranking.

Deterministic approaches can be further categorized into score-based and distance-based methods. Approaches in the first category apply relatively simple and efficient functions to calculate the score of each item, and the aggregate ranking is obtained by sorting items based on their total scores. Score-based methods are relatively susceptible to errors and manipulation, and they may violate certain fundamental social choice properties [5]. Conversely, distance-based methods provide more robust aggregation mechanisms. The aim of these approaches is to find a *consensus list* that has the least cumulative disagreement with the input lists. They are typically founded on axiomatic frameworks, from which the aggregate solution is formally guaranteed to satisfy certain desirable properties [9]. However, their aggregation problems tend to be more computationally demanding and are often NP-hard [5].

Distance-based TOP-k-AGG techniques can be divided based on whether the output ranking is considered a full list or another top-k list. Dwork et al. [10], Ailon [1], and Nápoles et al. [19] fall into the first category; Fagin et al. [12] falls into the second category. The works referenced under the first category define TOP-k-AGG as finding a full list with the least cumulative distance to the input lists using the induced Kendall tau, Kendall tau, and Hausdorff distances, respectively. Fagin et al. [12]'s method provides higher flexibility, and it induces a far smaller solution space. Letting n denote the total number of items, there are $\binom{n}{k}k!$ possible top-k lists using the latter approach, which is $(n-k)!$ times smaller than $n!$ (the number of possible full strict lists over n).

There are various distance measures for comparing top-k lists including generalized Kendall tau, generalized Spearman's footrule, Hausdorff [12], and Goodman and Kruskal's gamma [14]. This paper focuses on the distance-based variant of TOP-k-AGG induced by the generalized Kendall tau distance [12]. This focus is motivated by its widespread use for comparing top-k lists, and more importantly, its flexibility at handling partial information from these lists. This distance measure has been used in this capacity for similarity search [21], search engines [18], and influence maximization [4]. Additionally, variants of this distance have been used for comparing and aggregating bucket orders [11, 3] and top-k XML lists [23]. However, to the best of our knowledge, this distance measure has not been utilized for the purpose of aggregating top-k lists since its introduction in Fagin et al. [12], possibly due to a lack of existing exact methods. To facilitate this essential use of the distance measure, this paper studies various exact mathematical formulations.

Contributions. Section 3 introduces a binary nonlinear programming formulation and four mixed integer linear programming (MIP) formulations of TOP-k-AGG under the generalized Kendall tau distance. Two of these formulations result from the introduction of preference cycle-prevention constraints specific to TOP-k-AGG. Section 4 compares the strengths of the MIP formulations using techniques from polyhedral theory. The mathematical formulations and polyhedral analyses presented herein can be extended to TOP-k-AGG using any other distance measure between top-k lists by modifying the objective functions accordingly.

2 Preliminaries

The rank aggregation problem was originally defined over strict rankings. Formally, a strict ranking $\boldsymbol{\pi}$ is a bijection of $[n] = \{1, 2, \ldots, n\}$ onto itself, which represents a strict order of the n items. The Kendall tau distance [15] is one of the most prominent measures of dissimilarity between rankings, which counts the number of distinct item-pairs whose relative order is different in two rankings. The Kendall tau distance between strict rankings $\boldsymbol{\pi}^1, \boldsymbol{\pi}^2$ is given by $K(\boldsymbol{\pi}^1, \boldsymbol{\pi}^2) = \sum_{i \in [n]} \sum_{j \in [n]} K_{i,j}(\boldsymbol{\pi}^1, \boldsymbol{\pi}^2)$, where $K_{i,j}(\boldsymbol{\pi}^1, \boldsymbol{\pi}^2)$ is set to 1 if the relative orderings of i and j are different in $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$, and 0 otherwise. The rank aggregation problem under Kendall tau distance is known alternatively as Kemeny Aggregation (KEMENY-AGG).

A top-k list $\boldsymbol{\tau}$ is a bijection from a domain $\boldsymbol{\mathcal{I}}_{\boldsymbol{\tau}}$ (the members of $\boldsymbol{\tau}$) to $[k] = \{1, \ldots, k\}$, where $k < n$. All items in $\boldsymbol{\tau}$ are presumed to be ranked ahead of items not in $\boldsymbol{\tau}$; however, the exact ordering of items not in the list is unknown. Let $i \in \boldsymbol{\tau}$ indicate that item i appears in the top-k list, and let $\boldsymbol{\tau}(i)$ denote the rank or position of i therein. Additionally, let $i \succ_{\boldsymbol{\tau}} j$ denote that item i is rank ahead of item j in $\boldsymbol{\tau}$, that is, if $(i \in \boldsymbol{\tau} \wedge j \notin \boldsymbol{\tau})$ OR $(i, j \in \boldsymbol{\tau} \wedge (\boldsymbol{\tau}(i) < \boldsymbol{\tau}(j)))$. Given top-$k$ lists $\boldsymbol{\tau}^1$ and $\boldsymbol{\tau}^2$, let $\boldsymbol{\Lambda}(\boldsymbol{\tau}^1, \boldsymbol{\tau}^2)$ be the set of all unordered pairs of distinct items in $\boldsymbol{\mathcal{I}}_{\boldsymbol{\tau}^1} \bigcup \boldsymbol{\mathcal{I}}_{\boldsymbol{\tau}^2}$. (TOP-$k$-AGG) Let $\boldsymbol{\mathcal{L}} = \{1, 2, \ldots, m\}$ be the set of indices of the input top-k lists, $\boldsymbol{\tau}^l$ be the input top-k list $l \in \boldsymbol{\mathcal{L}}$, $\boldsymbol{\mathcal{I}} = \bigcup_{l \in \boldsymbol{\mathcal{L}}} \boldsymbol{\mathcal{I}}_{\boldsymbol{\tau}^l}$ be the universe of items, $n := |\boldsymbol{\mathcal{I}}|$ be the number of items in the universe $\boldsymbol{\mathcal{I}}$, $\boldsymbol{\mathcal{T}}$ be the set of all possible top-k lists over $\boldsymbol{\mathcal{I}}$, and $d(.,.)$ be a distance measure between top-k lists. TOP-k-AGG seeks to find a top-k list $\boldsymbol{\tau}^* \in \boldsymbol{\mathcal{T}}$ with the lowest cumulative distance to the input lists; it can be written succinctly as

$$\boldsymbol{\tau}^* = \underset{\boldsymbol{\tau} \in \boldsymbol{\mathcal{T}}}{\operatorname{argmin}} \sum_{l \in \boldsymbol{\mathcal{L}}} d(\boldsymbol{\tau}, \boldsymbol{\tau}^l). \tag{1}$$

The rest of this paper focuses on the generalized Kendall tau distance [12]. Accordingly, the distance is restated in the following. Let p be a fixed parameter, with $0 \leq p \leq 1$, and let $K_{i,j}^{(p)}(\boldsymbol{\tau}^1, \boldsymbol{\tau}^2)$ be the contribution to the distance function, for each item-pair $(i, j) \in \boldsymbol{\Lambda}(\boldsymbol{\tau}^1, \boldsymbol{\tau}^2)$. The generalized Kendall tau distance with penalty parameter p, denoted by

$K^{(p)}$, is defined as

$$K^{(p)}(\boldsymbol{\tau}^1, \boldsymbol{\tau}^2) = \sum_{(i,j)\in\boldsymbol{\Lambda}(\boldsymbol{\tau}^1,\boldsymbol{\tau}^2)} K_{i,j}^{(p)}(\boldsymbol{\tau}^1, \boldsymbol{\tau}^2), \tag{2}$$

where

$$K_{i,j}^{(p)}(\boldsymbol{\tau}^1, \boldsymbol{\tau}^2) = \begin{cases} 1 & (i \succ_{\boldsymbol{\tau}^1} j \ \wedge \ j \succ_{\boldsymbol{\tau}^2} i) \ \vee \ (j \succ_{\boldsymbol{\tau}^2} i \ \wedge \ i \succ_{\boldsymbol{\tau}^1} j) \\ p & (i,j \in \boldsymbol{\tau}^1 \ \wedge \ i,j \notin \boldsymbol{\tau}^2) \ \vee \ (i,j \notin \boldsymbol{\tau}^1 \ \wedge \ i,j \in \boldsymbol{\tau}^2) \\ 0 & \text{otherwise.} \end{cases}$$

$K^{(p)}$ is a *near metric* since it satisfies a relaxed version of the triangle inequality [12]. TOP-k-AGG under $K^{(p)}$ is a combinatorial NP-hard problem [12], which includes KEMENY-AGG as a special case (when $k = n$).

3 Integer Programming Formulations

To the best of our knowledge, no efforts have been made to derive an explicit mathematical model of TOP-k-AGG. This section presents various formulations.

First, we define required parameters for defining the objective functions of the presented formulations. Let μ_{il} be an indicator parameter that is equal to 1 if $i \in \boldsymbol{\tau}^l$, where $l \in \boldsymbol{\mathcal{L}}$. Additionally, let s_{ij} denote the number of input lists where item i is ranked ahead of item j, which can be expressed as

$$\begin{aligned} s_{ij} &= \sum_{l\in\boldsymbol{\mathcal{L}}} \mathbb{1}_{(i,j\in\boldsymbol{\tau}^l \wedge (\boldsymbol{\tau}^l(i) < \boldsymbol{\tau}^l(j)) \vee (i\in\boldsymbol{\tau}^l \wedge j\notin\boldsymbol{\tau}^l)} \\ &= \sum_{l\in\boldsymbol{\mathcal{L}}} \left[\mu_{il}\,\mu_{jl}\,\mathbb{1}_{\boldsymbol{\tau}^l(i) < \boldsymbol{\tau}^l(j)} + \mu_{il}(1-\mu_{jl})\right]. \end{aligned} \tag{3}$$

In words, s_{ij} tallies the number of input lists in which i is ranked ahead of j, that is, the number of input lists in which both items are present and i is ranked ahead of j, plus the number of inputs lists in which i is present but j is not.

Using these parameters, the cumulative $K^{(p)}$ distance between a given top-k list $\boldsymbol{\tau} \in \boldsymbol{\mathcal{T}}$ and all of the input top-k lists, i.e., $\sum_{\boldsymbol{\tau}^l\in\boldsymbol{\mathcal{L}}} \sum_{(i,j)\in\boldsymbol{\Lambda}(\boldsymbol{\tau},\boldsymbol{\tau}^l)} K_{ij}^{(p)}(\boldsymbol{\tau}, \boldsymbol{\tau}^l)$, can be expressed as $\sum_{(i,j)\in\boldsymbol{\Lambda}} K_{ij}^{(p)}(\boldsymbol{\tau})$ where $\boldsymbol{\Lambda}$ is set of all unordered pairs of distinct items in $\boldsymbol{\mathcal{I}}$, and

$$K_{ij}^{(p)}(\boldsymbol{\tau}) = \begin{cases} s_{ji} + p\sum_{l\in\boldsymbol{\mathcal{L}}}(1-\mu_{il})(1-\mu_{jl}) & \text{if } i,j \in \boldsymbol{\tau} \ \wedge \ (\boldsymbol{\tau}(i) < \boldsymbol{\tau}(j)), \\ s_{ji} & \text{if } i \in \boldsymbol{\tau} \ \wedge \ j \notin \boldsymbol{\tau}, \\ p\sum_{l\in\boldsymbol{\mathcal{L}}} \mu_{il}\mu_{jl} & \text{if } i,j \notin \boldsymbol{\tau}. \end{cases} \tag{4}$$

Eq. (4) states that, whenever item i and j are both present in $\boldsymbol{\tau}$ (the solution top-k list) and i is ranked ahead of item j, the imposed $K^{(p)}$ distance between $\boldsymbol{\tau}$ and all of the input lists for this pair of items equals the number of input lists where j is ranked ahead of i, plus p-times the number of input lists neither i nor j is present in the same list. Whenever i but not j is present in $\boldsymbol{\tau}$, the imposed $K^{(p)}$ distance equals the number of input lists where j is ranked ahead of i. Finally, whenever neither i nor j is present in $\boldsymbol{\tau}$, the imposed $K^{(p)}$ distance equals p times the number of input lists where i and j are simultaneously present.

The first formulation is an MIP possessing an assignment problem-like structure, with which exactly k items are assigned to the k available positions of the solution top-k list. Its decisions variables are as follows:

$$u_{it} = \begin{cases} 1 & \text{if } i \text{ is assigned to position } t \in [k] \\ 0 & \text{otherwise;} \end{cases}$$

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are in the top-}k \text{ list, and } i \text{ is ranked ahead of } j \\ 0 & \text{otherwise;} \end{cases}$$

$$w'_{ij} = \begin{cases} 1 & \text{if } i \text{ is in the top-}k \text{ list, but not } j \\ 0 & \text{otherwise;} \end{cases}$$

$$w''_{ij} = \begin{cases} 1 & \text{if neither } i \text{ nor } j \text{ is present in the top-}k \text{ list, where } j > i \\ 0 & \text{otherwise.} \end{cases}$$

From the definitions, item i is present in the top-k list if $\sum_{t=1}^{k} u_{it} = 1$, and it is absent if $\sum_{t=1}^{k} u_{it} = 0$. The variables $\boldsymbol{w}, \boldsymbol{w'}$, and $\boldsymbol{w''}$ determine the relative ordering of the items; these are dependent variables, as their exact values are determined by the values of the $\boldsymbol{u}$-variables. The first formulation (MIP#1) is as follows.

$$\min_{u,w,w',w''} \quad \sum_{i\in\boldsymbol{\mathcal{I}}}\sum_{j\in\boldsymbol{\mathcal{I}}}\left[(s_{ji} + p\sum_{l\in\boldsymbol{\mathcal{L}}}(1-\mu_{il})(1-\mu_{jl}))w_{ij} + s_{ji}w'_{ij}\right] + p\sum_{i,j\in\boldsymbol{\mathcal{I}},j>i}\sum_{l\in\boldsymbol{\mathcal{L}}}\mu_{il}\mu_{jl}w''_{ij} \tag{5a}$$

$$\text{s.t.} \quad \sum_{i\in\boldsymbol{\mathcal{I}}} u_{it} = 1 \qquad \forall t \in [k] \tag{5b}$$

$$\sum_{t\in[k]} u_{it} \leq 1 \qquad \forall i \in \boldsymbol{\mathcal{I}} \tag{5c}$$

$$w_{ij} \geq \sum_{t'=1}^{t} u_{it'} + \sum_{t''=t+1}^{k} u_{jt''} - 1 \quad \forall i,j \in \boldsymbol{\mathcal{I}},\, i \neq j;\ \forall t \in [k-1] \tag{5d}$$

$$\sum_{i,j\in\boldsymbol{\mathcal{I}}} w_{ij} \leq \frac{k(k-1)}{2} \tag{5e}$$

$$w'_{ij} \geq \sum_{t\in[k]} u_{it} - \sum_{t\in[k]} u_{jt} \qquad \forall i,j \in \boldsymbol{\mathcal{I}},\ i \neq j \tag{5f}$$

$$\sum_{i,j\in\boldsymbol{\mathcal{I}}} w'_{ij} = k(n-k) \tag{5g}$$

$$w''_{ij} \geq 1 - \sum_{t\in[k]} u_{it} - \sum_{t\in[k]} u_{jt} \qquad \forall i,j \in \boldsymbol{\mathcal{I}},\ i \neq j \tag{5h}$$

$$\sum_{i,j\in\boldsymbol{\mathcal{I}},j>i} w''_{ij} = \frac{(n-k)(n-k-1)}{2} \tag{5i}$$

$$u_{it} \in \{0,1\} \qquad \forall i \in \boldsymbol{\mathcal{I}};\ \ \forall t \in [k] \tag{5j}$$

$$w_{ij}, w'_{ij} \geq 0 \qquad \forall i,j \in \boldsymbol{\mathcal{I}},\ i \neq j \tag{5k}$$

$$w''_{ij} \geq 0 \qquad \forall i,j \in \boldsymbol{\mathcal{I}},\ \ j > i. \tag{5l}$$

Objective function (5a) minimizes the cumulative $K^{(p)}$ distance to the input lists according to Eq. (4). Constraint (5b) enforces that exactly one item must be assigned to each position of the top-k list. Constraint (5c) enforces that every item must be assigned to at most one position of the list. Constraint (5d) determines the respective values of the $\boldsymbol{w}$-variables. More specifically, $w_{ij} = 1$ if i occupies one of the first t positions ($\sum_{t'=t+1}^{t} u_{it'} = 1$) and j occupies position t'', where $t+1 \leq t'' \leq k$ ($\sum_{t''=t+1}^{k} u_{jt''} = 1$); otherwise, this constraint becomes redundant. Constraint (5d) and (5e) together impose preference transitivity (i.e., prevent preference cycles); this means that if h is ranked ahead of i, and i is ranked of j, then h must be ranked ahead of j as well (see Theorem 1). Constraint (5f) determines the respective values of $\boldsymbol{w'}$-variables; it enforces that $w'_{ij} = 1$ if i is present in the top-k list but not j; otherwise, this constraint becomes redundant. Constraint (5g) enforces that at most $k(n-k)$ of the $\boldsymbol{w'}$-variables can take a value of 1 as there are $k(n-k)$ distinct item-pairs where exactly one of the items appears in the list. Constraint (5h) enforces that $w''_{ij} = 1$ if neither i nor j is present in the top-k list; otherwise, this constraint becomes redundant. Constraint (5i) enforces that at most $(n-k)(n-k-1)/2$ of the $\boldsymbol{w''}$-variables can take a value of 1 as this is the number of distinct item-pairs where both items are absent from the list. Constraints (5j)-(5l) specify the domain of the variables.

Taking a closer look at the structure of the constraints, we can observe that even though variables $\boldsymbol{w}, \boldsymbol{w'}$ and $\boldsymbol{w''}$ are specified as binary, they can be treated as non-negative continuous variables since the constraints of the model alone enforce them to only take a value of 0 or 1. It is important also to remark that the reason for including constraints (5f) and (5g) is that the objective function coefficients are not necessarily positive. More specifically, if both i and j are present in the solution top-k list, constraint (5f) implies that $w'_{ij} \geq 0$; however, if the objective function coefficient s_{ij} is 0, then any value of w'_{ij} results in the same objective function value, which is not desirable.

Theorem 1. *Constraints* (5d)-(5e) *impose preference transitivity.*

Proof. Assume that items h, i, j are present in the solution top-k list with h placed in position $t \geq 1$, i in position $t' > t$, and j in position t'', where $k \geq t'' > t'$. Constraint (5d)

enforces that $w_{hi} = w_{hj} = w_{ij} = 1$. However, this constraint only implies that $w_{jh} \geq -1$. In other words, the optimization model may have incentive to assign $w_{jh} = 1$, creating a preference cycle, in order to decrease the objective function value. Hence, Constraint (5d) on its own does not prevent preference cycles.

However, the total number of $\boldsymbol{w}$-variables that *must* take a value of 1 is given by $(k-1) + (k-2) + \cdots + 1 + 0 = k(k-1)/2$—the first-ranked item is ahead of $k-1$ other items in the list, the second-ranked item is ahead of $k-2$ items, ..., and the item at the bottom of the list is not ranked ahead of any other items on the list. For this reason, constraint (5e) allows at most $k(k-1)/2$ of the $\boldsymbol{w}$-variables to take a value of 1, forcing all other variables (including w_{jh}) to equal 0. Therefore, constraints (5d)-(5e) together impose preference transitivity on the solution top-k list returned by solving MIP#1. □

Since KEMENY-AGG is a special case of TOP-k-AGG, MIP#1 provides a novel formulation for that problem as well; however, it does not apply to the variant of the problem with ties (see Yoo and Escobedo [24]). It is important to mention that Cook [9] proposed a binary linear programming formulation of KEMENY-AGG using the structure of the assignment problem; however, their set of preference cycle prevention constraint is different from constraints (5d)-(5e).

Next, we present a binary non-linear programming formulation for TOP-k-AGG. The formulation uses the $\boldsymbol{w}$-variables defined for MIP#1 as well as the following decision variables:

$$z_i = \begin{cases} 1 & \text{if } i \text{ is in the top-}k \text{ list} \\ 0 & \text{otherwise.} \end{cases}$$

The formulation is given by:

$$\min_{\boldsymbol{w},\boldsymbol{z}} \quad \sum_{i\in\boldsymbol{\mathcal{I}}}\sum_{j\in\boldsymbol{\mathcal{I}}}\left[(s_{ji} + p\sum_{l\in\boldsymbol{\mathcal{L}}}(1-\mu_{il})(1-\mu_{jl}))w_{ij} + s_{ji}z_i(1-z_j)\right] + p\sum_{i,j\in\boldsymbol{\mathcal{I}},j>i}\sum_{l\in\boldsymbol{\mathcal{L}}}\mu_{il}\mu_{jl}(1-z_i)(1-z_j) \tag{6a}$$

$$\text{s.t.} \quad \sum_{i\in\boldsymbol{\mathcal{I}}} z_i = k \tag{6b}$$

$$w_{hi} + w_{ij} + w_{jh} \leq 2 \qquad \forall h,i,j \in \boldsymbol{\mathcal{I}},\ i,j > h,\ i \neq j \tag{6c}$$

$$w_{ij} + w_{ji} = z_i z_j \qquad \forall i,j \in \boldsymbol{\mathcal{I}},\ j > i \tag{6d}$$

$$z_i, w_{ij} \in \{0,1\} \qquad \forall i,j \in \boldsymbol{\mathcal{I}},\ i \neq j. \tag{6e}$$

Objective function (6a) minimizes the cumulative $K^{(p)}$ distance to the input lists. Constraint (6b) restricts k items to be present in the top-k list. Constraint (6c) imposes preference transitivity only whenever items h, i, j all appear in the list; otherwise it becomes redundant, with the help of constraint (6d). Constraint (6d) enforces that, when both i and j are present in the list, one must proceed the other. Constraint (6e) specifies the domains

of the variables. Given a feasible solution, the output top-k items are defined by the set $\overline{\boldsymbol{\tau}} := \{i \in \boldsymbol{\mathcal{I}} | z_i = 1\}$, and the exact rank of item $i \in \overline{\boldsymbol{\tau}}$ is obtained as $\overline{\boldsymbol{\tau}}(i) := k - \sum_{j \in \overline{\boldsymbol{\tau}}} w_{ij}$.

The above non-linear optimization model can be linearized using a technique from Glover and Woolsey [13]. Specifically, constraint (6d) can be replaced with three linear constraints for each distinct item pair (i, j): $w_{ij} + w_{ji} \leq z_i$, $w_{ij} + w_{ji} \leq z_j$, and $w_{ij} + w_{ji} \geq z_i + z_j - 1$. Similarly, the term $z_i(1 - z_j)$ in the objective function is replaced by auxiliary continuous variable x'_{ij} and constraints $x'_{ij} \geq z_i - z_j$ and $x'_{ij} \geq 0$; and the term $(1 - z_i)(1 - z_j)$ in the objective function is replaced by auxiliary continuous variable x''_{ij} and constraints $x''_{ij} \geq 1 - z_i - z_j$ and $x''_{ij} \geq 0$. The latter two cases use the fact the objective function coefficients of $z_i(1 - z_j)$ and $(1 - z_i)(1 - z_j)$ are non-negative, leading to a reduction in the number of constraints required by the linearization. The resulting formulation (MIP#2) is given by:

$$\min_{\boldsymbol{w}, \boldsymbol{x'}, \boldsymbol{x''}, \boldsymbol{z}} \quad \sum_{i \in \boldsymbol{\mathcal{I}}} \sum_{j \in \boldsymbol{\mathcal{I}}} \left[(s_{ji} + p \sum_{l \in \boldsymbol{\mathcal{L}}} (1 - \mu_{il})(1 - \mu_{jl})) w_{ij} + s_{ji} x'_{ij} \right] + p \sum_{i,j \in \boldsymbol{\mathcal{I}}, j > i} \sum_{l \in \boldsymbol{\mathcal{L}}} \mu_{il} \mu_{jl} x''_{ij} \tag{7a}$$

$$\text{s.t.} \quad (6b), (6c), (6e) \tag{7b}$$

$$w_{ij} + w_{ji} \geq z_i + z_j - 1 \qquad \forall i, j \in \boldsymbol{\mathcal{I}},\ j > i \tag{7c}$$

$$w_{ij} + w_{ji} \leq z_i \qquad \forall i, j \in \boldsymbol{\mathcal{I}},\ i \neq j \tag{7d}$$

$$x'_{ij} \geq z_i - z_j \qquad \forall i, j \in \boldsymbol{\mathcal{I}},\ i \neq j \tag{7e}$$

$$\sum_{i,j \in \boldsymbol{\mathcal{I}}} x'_{ij} = k(n - k) \tag{7f}$$

$$x''_{ij} \geq 1 - z_i - z_j \qquad \forall i, j \in \boldsymbol{\mathcal{I}},\ j > i \tag{7g}$$

$$\sum_{i,j \in \boldsymbol{\mathcal{I}}, j > i} x''_{ij} = \frac{(n - k)(n - k - 1)}{2} \tag{7h}$$

$$x'_{ij} \geq 0 \qquad \forall i, j \in \boldsymbol{\mathcal{I}},\ i \neq j, \tag{7i}$$

$$x''_{ij} \geq 0 \qquad \forall i, j \in \boldsymbol{\mathcal{I}},\ j > i. \tag{7j}$$

The rationale behind including constraints (7f) and (7h) is the same as constraints (5g) and (5i) in MIP#1.

Next, we define two variants of the preference transitivity constraints utilized in MIP#2.

Proposition 1. *Constraint* (6c) *can be replaced by non-linear constraints*

$$w_{hi} + w_{ij} + w_{jh} \leq 3 - z_h z_i z_j \qquad \forall i, j > h,\ i \neq j, \quad \textbf{\textit{or}} \tag{8}$$

$$w_{hi} + w_{ij} + w_{jh} \leq 1 + z_h z_i z_j \qquad \forall i, j > h,\ i \neq j. \tag{9}$$

Furthermore, these constraints can be linearized respectively as

$$w_{hi} + w_{ij} + w_{jh} \leq 3 - \frac{1}{3}(z_h + z_i + z_j) \quad \forall h, i, j \in \boldsymbol{\mathcal{I}},\ i, j > h,\ i \neq j, \tag{10}$$

$$w_{hi} + w_{ij} + w_{jh} \leq 1 + \frac{1}{3}(z_h + z_i + z_j) \quad \forall h, i, j \in \boldsymbol{\mathcal{I}},\ i, j > h,\ i \neq j. \tag{11}$$

Proof. The right-hand side of constraints (8)-(11) becomes 2, as desired, when items h, i, j are all in the solution top-k list, i.e., when $z_h = z_i = z_j = 1$. For the remaining cases, these constraints become redundant, with the help of constraint (7d). In particular, assume i is not in the top-k list; constraint (7d) enforces that $w_{ij} + w_{ji} \leq 0$ and $w_{ih} + w_{hi} \leq 0$; hence, constraints (8)-(11) effectively reduce to $w_{jh} \leq 1$, which is redundant. □

Replacing constraint (6c) with constraints (10) and (11), respectively, induces two additional MIPs.

MIP#3:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{x'},\boldsymbol{x''},\boldsymbol{z}} \quad & (7a) \\ \text{s.t.} \quad & (6b), (6e), (7c) - (7g) \\ & w_{hi} + w_{ij} + w_{jh} \leq 3 - \frac{1}{3}(z_h + z_i + z_j) \quad \forall h, i, j \in \boldsymbol{\mathcal{I}},\ i, j > h,\ i \neq j. \end{aligned}$$

MIP#4:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{x'},\boldsymbol{x''},\boldsymbol{z}} \quad & (7a) \\ \text{s.t.} \quad & (6b), (6e), (7c) - (7g) \\ & w_{hi} + w_{ij} + w_{jh} \leq 1 + \frac{1}{3}(z_h + z_i + z_j) \quad \forall h, i, j \in \boldsymbol{\mathcal{I}},\ i, j > h,\ i \neq j. \end{aligned}$$

4 Polyhedral Comparison

Next, we compare the strength of the proposed MIPs based on their linear programming (LP) relaxation models. First, we compare the strength of MIPs #2, #3, and #4. To that end, notice that these three MIPs become equivalent when $k \leq 2$—when the preference transitivity relations are irrelevant—or when $n = k$—when all items appear in the solution top-k list. Afterwards, we show that each of these formulations is stronger than MIP#1. For the remainder of the paper, let $\mathcal{P}^1, \mathcal{P}^2, \mathcal{P}^3, \mathcal{P}^4$ be the polyhedral corresponding to the LP relaxations of MIPs #1, #2, #3, #4, respectively.

Theorem 2. *For any instance of TOP-k-AGG, $\mathcal{P}^4 \subseteq \mathcal{P}^2 \subseteq \mathcal{P}^3$, and these inclusions can be strict.*

Proof. Note that MIPs #2, #3, and #4 differ only in their preference transitivity constraints. First, we show that $\mathcal{P}^4 \subseteq \mathcal{P}^2 \subseteq \mathcal{P}^3$.

Since $0 \leq z_i \leq 1\ \forall i \in \boldsymbol{\mathcal{I}}$, for every feasible solution in $\mathcal{P}^2, \mathcal{P}^3, \mathcal{P}^4$, we have that $(z_h + z_i + z_j)/3 \leq 1\ \forall h, i, j \in \boldsymbol{\mathcal{I}},\ i, j > h,\ i \neq j$. Letting $(\boldsymbol{w}, \boldsymbol{x'}, \boldsymbol{x''}, \boldsymbol{z})^{(4)} \in \mathcal{P}^4$ be a feasible

solution to MIP#4, we have that

$$w_{hi}^{(4)} + w_{ij}^{(4)} + w_{jh}^{(4)} \leq 1 + \frac{1}{3}(z_i^{(4)} + z_j^{(4)} + z_h^{(4)}) \leq 2 \leq 3 - \frac{1}{3}(z_i^{(4)} + z_j^{(4)} + z_h^{(4)}).$$

Therefore, all feasible solutions to MIP#4 are also feasible to MIPs #2 and #3. Using the same logic, all feasible solutions to MIP#2 are feasible to MIP#3. This gives that $\mathcal{P}^4 \subseteq \mathcal{P}^2 \subseteq \mathcal{P}^3$.

To show that the inclusion $\mathcal{P}^4 \subseteq \mathcal{P}^2$ can be strict, consider a small instance with $\boldsymbol{\mathcal{I}} = \{1, 2, 3, 4\}$ and $k = 3$. Fix the solution $(\boldsymbol{w}, \boldsymbol{x'}, \boldsymbol{x''}, \boldsymbol{z})^{(2)} \in \mathcal{P}^2$ as

$$x_{14}'^{(2)} = x_{24}'^{(2)} = x_{34}'^{(2)} = 0.24, \quad w_{12}^{(2)} = w_{23}^{(2)} = w_{31}^{(2)} = 0.62, \quad w_{14}^{(2)} = w_{24}^{(2)} = w_{34}^{(2)} = 0.38,$$
$$z_1^{(2)} = z_2^{(2)} = z_3^{(2)} = 0.81, \qquad z_4^{(2)} = 0.57;$$

with all other variables equal to 0. By inspection, this solution satisfies all constraints of MIP#2. However, we have that

$$w_{12}^{(2)} + w_{23}^{(2)} + w_{31}^{(2)} = 1.86 \nleq 1 + \frac{0.81 + 0.81 + 0.81}{3} = 1.81.$$

This indicates that this solution does not satisfy the preference transitivity constraints of MIP#4.

Next, we use a similar process to show that the inclusion $\mathcal{P}^2 \subseteq \mathcal{P}^3$ can be strict. Consider a small instance with $\boldsymbol{\mathcal{I}} = \{1, 2, 3, 4\}$ and $k = 3$. Fix the solution $(\boldsymbol{w}, \boldsymbol{x'}, \boldsymbol{x''}, \boldsymbol{z})^{(3)} \in \mathcal{P}^3$ as

$$x_{14}'^{(3)} = x_{24}'^{(3)} = x_{34}'^{(3)} = 0.4, \qquad w_{12}^{(3)} = w_{23}^{(3)} = w_{31}^{(3)} = 0.7, \qquad w_{14}^{(3)} = w_{24}^{(3)} = w_{34}^{(3)} = 0.3,$$
$$z_1^{(2)} = z_2^{(3)} = z_3^{(3)} = 0.85, \qquad z_4^{(3)} = 0.45;$$

with all other variables equal to 0. By inspection, this solution satisfies all constraints of MIP#3. However, we have that

$$w_{12}^{(3)} + w_{23}^{(3)} + w_{31}^{(3)} = 2.1 \nleq 2.$$

This indicates that this solution does not satisfy the preference transitivity constraints of MIP#2. □

Theorem 3. *For any instance of TOP-k-AGG, $proj_{\boldsymbol{w}}\, \mathcal{P}^2, proj_{\boldsymbol{w}}\, \mathcal{P}^3, proj_{\boldsymbol{w}}\, \mathcal{P}^4 \subseteq proj_{\boldsymbol{w}}\, \mathcal{P}^1$, and these inclusions can be strict.*

Proof. First, we prove that $\text{proj}_{\boldsymbol{w}}\mathcal{P}^3 \subseteq \text{proj}_{\boldsymbol{w}}\mathcal{P}^1$. We show that, starting from an arbitrary solution $(\boldsymbol{w}, \boldsymbol{x'}, \boldsymbol{x''}, \boldsymbol{z}) \in \mathcal{P}^3$, we can deduce a solution $(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{w'}, \boldsymbol{w''}) \in \mathcal{P}^1$. To this end, we define the following affine mappings of variables from $\mathcal{P}^3$ to $\mathcal{P}^1$:

$$u_{it} = \frac{z_i}{k} \quad \forall i \in \boldsymbol{\mathcal{I}}, \ \forall t \in \{1, \ldots, k\} \rightarrow \sum_{t=1}^{k} u_{it} = z_i \quad \forall i \in \boldsymbol{\mathcal{I}}, \tag{13a}$$

$$w'_{ij} = x'_{ij} \quad \forall i, j \in \boldsymbol{\mathcal{I}},\ i \neq j, \tag{13b}$$
$$w''_{ij} = x''_{ij} \quad \forall i, j \in \boldsymbol{\mathcal{I}},\ j > i. \tag{13c}$$

Mapping (13b)-(13c) guarantees that the objective function values achieved by the respective feasible points are equal. To establish that $\text{proj}_{\boldsymbol{w}}\mathcal{P}^3 \subseteq \text{proj}_{\boldsymbol{w}}\mathcal{P}^1$, it is sufficient to show that, given a feasible solution in $\mathcal{P}^3$, the mapped variables are guaranteed to satisfy all constraints of MIP#1 (i.e., this point belongs to $\mathcal{P}^1$).

Consider constraint (5b). For any $t \in \{1, \ldots, k\}$, we have

$$\sum_{i \in \boldsymbol{\mathcal{I}}} u_{it} = \sum_{i \in \boldsymbol{\mathcal{I}}} \frac{z_i}{k} = \frac{\sum_{i \in \boldsymbol{\mathcal{I}}} z_i}{k} \xrightarrow{\sum_{i \in \boldsymbol{\mathcal{I}}} z_i = k} \sum_{i \in \boldsymbol{\mathcal{I}}} u_{it} = 1.$$

Therefore, mapping (13a) provides a solution that is guaranteed to satisfy constraint (5b).

Consider constraint (5c). For every $i \in \boldsymbol{\mathcal{I}}$, we have

$$\sum_{t=1}^{k} u_{it} = \sum_{t=1}^{k} \frac{z_i}{k} = \frac{k z_i}{k} = z_i \leq 1.$$

The last inequality follows from the fact that the $\boldsymbol{z}$-variables are binary. Therefore, mapping (13a) provides a solution that is guaranteed to satisfy constraint (5c).

Next, consider constraint (5d); we focus on the maximum value of the right-hand side of this constraint given mapping (13a). For any arbitrary item-pair (i, j) and any $t \in \{1, \ldots, k-1\}$ we have

$$\begin{aligned}
\sum_{t'=1}^{t} u_{it'} + \sum_{t''=t+1}^{k} u_{jt''} - 1 &= \sum_{t'=1}^{t} \frac{z_i}{k} + \sum_{t''=t+1}^{k} \frac{z_j}{k} - 1 \\
&= \frac{t z_i}{k} + \frac{(k-t) z_j}{k} - 1 \\
&\leq \frac{t}{k} + \frac{k-t}{k} - 1 = \frac{k}{k} - 1 = 1 - 1 = 0.
\end{aligned}$$

The above equation states that using mapping (13a), the left-hand side values of constraint (5d) will be non-positive. Since $w_{ij} \geq 0$, mapping (13a) provides a solution that is guaranteed to satisfy constraint (5d).

Next, consider constraint (5e). By summing over constraint (7d), we have

$$\begin{aligned}
2 \sum_{i,j \in \boldsymbol{\mathcal{I}}} w_{ij} &\leq (k-1) \sum_{i \in \boldsymbol{\mathcal{I}}} z_i = k(k-1) \\
\rightarrow \sum_{i,j \in \boldsymbol{\mathcal{I}}} w_{ij} &\leq \frac{k(k-1)}{2},
\end{aligned}$$

which is exactly constraint (5e).

Finally, consider constraints (5f)-(5i). Mappings (13a)-(13c) imply that all feasible solutions to constraints (7e)-(7h) are feasible to constraints (5f)-(5i). Putting all pieces together, we have $\text{proj}_{\boldsymbol{w}} \mathcal{P}^3 \subseteq \text{proj}_{\boldsymbol{w}} \mathcal{P}^1$.

Note that the preference cycle-prevention constraints of MIP#3 have no counterpart in MIP#1. Therefore, we can show that the inclusion $\text{proj}_{\boldsymbol{w}} \mathcal{P}^3 \subseteq \text{proj}_{\boldsymbol{w}} \mathcal{P}^1$ can be strict by providing a solution that satisfies constraints (7c)-(7f) but violates preference cycle-prevention constraint (10), as this solution satisfies all constraints of MIP#1. There is an infinite number of such solutions; for example, consider a small instance with $\boldsymbol{\mathcal{I}} = \{1, 2, 3, 4\}$ and $k = 3$. Fix the solution $(\boldsymbol{w}, \boldsymbol{x'}, \boldsymbol{x''}, \boldsymbol{z})^{(3)}$ as

$$x'^{(3)}_{14} = x'^{(3)}_{24} = x'^{(3)}_{34} = 0.44, \quad w^{(3)}_{12} = w^{(3)}_{23} = w^{(3)}_{31} = 0.72, \quad w^{(3)}_{14} = w^{(3)}_{24} = w^{(3)}_{34} = 0.28,$$
$$z^{(2)}_1 = z^{(3)}_2 = z^{(3)}_3 = 0.86, \qquad z^{(3)}_4 = 0.42;$$

with all other variables equal to 0. By inspection, this solution satisfies constraints (7c)-(7f); however, it violates the preference transitivity constraints involved in MIP#3, as we have

$$w_{12} + w_{23} + w_{31} = 2.16 \not\leq 3 - (0.86 + 0.86 + 0.86)/3 = 2.14.$$

Finally, from Theorem 2, we have that $\mathcal{P}^4 \subseteq \mathcal{P}^2 \subseteq \mathcal{P}^3$; therefore, we can conclude that $\text{proj}_{\boldsymbol{w}} \mathcal{P}^2, \text{proj}_{\boldsymbol{w}} \mathcal{P}^4 \subseteq \text{proj}_{\boldsymbol{w}} \mathcal{P}^1$, and these inclusions can be strict. □

5 Concluding Remarks

This paper studies the top-k list aggregation problem, which includes Kemeny aggregation as a special case. It presents a binary non-linear and four mixed-integer linear programming formulations. Furthermore, it studies the strength of the four mixed-integer linear programming formulations using polyhedral analysis. Our findings shows that the presented formulations can be ordered based on the strength of their LP relaxations. The strongest formulation is induced by a novel set of preference cycle-prevention constraints tailored to the specific structure of the top-k list aggregation problem introduced herein.

Future research will explore heuristic and approximation algorithms for this problem. Additionally, investigating whether lower bounding techniques of Kemeny aggregation [2] can be modified for the top-k list aggregation problem can be another avenue of research.

References

[1] Ailon, N. (2010). Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica*, 57(2):284–300.

[2] Akbari, S. and Escobedo, A. R. (2021). Lower bounds on kemeny rank aggregation with non-strict rankings. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE.

[3] Akbari, S. and Escobedo, A. R. (2022). Beyond kemeny aggregation: Theoretical and computational insights for robust ranking aggregation. *under review*.

[4] Aslay, C., Barbieri, N., Bonchi, F., and Baeza-Yates, R. (2014). Online topic-aware influence maximization queries. In *EDBT*, pages 295–306.

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

[6] Chen, Y., Fan, J., Ma, C., and Wang, K. (2019). Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204.

[7] Chierichetti, F., Dasgupta, A., Haddadan, S., Kumar, R., and Lattanzi, S. (2018). Mallows models for top-k lists. In *Advances in Neural Information Processing Systems*, pages 4382–4392.

[8] Collas, F. and Irurozki, E. (2021). Concentric mixtures of mallows models for top-k rankings: sampling and identifiability. In *International Conference on Machine Learning*, pages 2079–2088. PMLR.

[9] Cook, W. D. (2006). Distance-based and ad hoc consensus models in ordinal preference ranking. *European Journal of operational research*, 172(2):369–385.

[10] Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622.

[11] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. (2004). Comparing and aggregating rankings with ties. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 47–58.

[12] Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160.

[13] Glover, F. and Woolsey, E. (1974). Converting the 0-1 polynomial programming problem to a 0-1 linear program. *Operations research*, 22(1):180–182.

[14] Goodman, L. A. and Kruskal, W. H. (1959). Measures of association for cross classifications. ii: Further discussion and references. *Journal of the American Statistical Association*, 54(285):123–163.

[15] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

[16] Mallows, C. L. (1957). Non-null ranking models. i. *Biometrika*, 44(1/2):114–130.

[17] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804.

[18] McCown, F. and Nelson, M. L. (2007). Agreeing to disagree: search engines and their public interfaces. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital Libraries*, pages 309–318.

[19] Nápoles, G., Falcon, R., Dikopoulou, Z., Papageorgiou, E., Bello, R., and Vanhoof, K. (2017). Weighted aggregation of partial rankings using ant colony optimization. *Neurocomputing*, 250:109–120.

[20] Oliveira, S. E., Diniz, V., Lacerda, A., Merschmann, L., and Pappa, G. L. (2020). Is rank aggregation effective in recommender systems? an experimental analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(2):1–26.

[21] Pal, K. and Michel, S. (2016). Efficient similarity search across top-k lists under the kendall's tau distance. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management*, pages 1–12.

[22] Pedroche, F. and Conejero, J. A. (2020). Corrected evolutive kendall's τ coefficients for incomplete rankings with ties: Application to case of spotify lists. *Mathematics*, 8(10):1828.

[23] Varadarajan, R., Farfán, F., and Hristidis, V. (2013). Comparing top-k xml lists. *Information Systems*, 38(6):820–834.

[24] Yoo, Y. and Escobedo, A. R. (2021). A new binary programming formulation and social choice property for kemeny rank aggregation. *Decision Analysis*, 18(4):296–320.