Partial Kernelization for Rank Aggregation: Theory and Experiments

Nadja Betzler[1], Robert Bredereck[1], and Rolf Niedermeier

Abstract

RANK AGGREGATION is important in many areas ranging from web search over databases to bioinformatics. The underlying decision problem KEMENY SCORE is NP-complete even in case of four input rankings to be aggregated into a "median ranking". We study efficient polynomial-time data reduction rules that allow us to find *optimal* median rankings. On the theoretical side, we improve a result for a "partial problem kernel" from quadratic to linear size. On the practical side, we provide encouraging experimental results with data based on web search and sport competitions, e.g., computing optimal median rankings for real-world instances with more than 100 candidates within milliseconds.

1 Introduction

We investigate the effectiveness of data reduction for computing *optimal* solutions of the NP-hard RANK AGGREGATION problem. Kemeny's corresponding voting scheme goes back to the year 1959 [14] and was later specified by Levenglick [16]. It can be described as follows. An *election* (V, C) consists of a set V of n votes and a set C of m candidates. A vote or a *ranking* is a total order of all candidates. For instance, in case of three candidates a, b, c, the order $c > b > a$ means that candidate c is the best-liked one and candidate a is the least-liked one. For each pair of votes v, w, the *Kendall-Tau distance* between v and w is defined as

$$\text{KT-dist}(v, w) = \sum_{\{c,d\} \subseteq C} d_{v,w}(c, d),$$

where $d_{v,w}(c, d)$ is set to 0 if v and w rank c and d in the same order, and is set to 1, otherwise. The *score* of a ranking l with respect to an election (V, C) is defined as $\sum_{v \in V} \text{KT-dist}(l, v)$. A ranking l with a minimum score is called a *Kemeny ranking* of (V, C) and its score is the *Kemeny score* of (V, C). The central problem considered in this work is as follows:

RANK AGGREGATION: Given an election (V, C), find a Kemeny ranking of (V, C).

Its decision variant KEMENY SCORE asks whether there is a Kemeny ranking of (V, C) with score at most some additionally given positive integer k. The RANK AGGREGATION problem has numerous applications, ranging from building meta-search engines for the web or spam detection [10] over databases [11] to the construction of genetic maps in bioinformatics [12]. Kemeny rankings are also desirable in classical voting scenarios such as the determination of a president (see, for example, `www.votefair.org`) or the selection of the best qualified candidates for job openings. The wide range of applications is due to the fulfillment of many desirable properties from the social choice point of view [23], including the *Condorcet property*: if there is a candidate (*Condorcet winner*) who is better than every other candidate in more than half of the votes, then this candidate is also ranked first in every Kemeny ranking.

Previous work. First computational complexity studies of KEMENY SCORE go back to Bartholdi et al. [3], showing its NP-hardness. Dwork et al. [10] showed that the problem remains NP-hard even in the case of four votes. Moreover, they identified its usefulness in aggregating web search results

[1] Supported by the DFG, research project PAWS, NI 369/10.

and provided several approximation and heuristic algorithms. Recent papers showed constant-factor approximability [2, 22] and an (impractical) PTAS [15]. Schalekamp and van Zuylen [20] provided a thorough experimental study of approximation and heuristic algorithms. Due to the importance of computing optimal solutions, there have been some experimental studies in this direction [8, 9]: An integer linear program and a branch-and-bound approach were applied to random instances generated under a noise model (motivated by the interpretation of Kemeny rankings as maximum likelihood estimators [8]). From a parameterized complexity perspective, the following is known. First fixed-parameter tractability results have been shown with respect to the single parameters number of candidates, Kemeny score, maximum range of candidate positions, and average KT-distance d_a [4]. The *average KT-distance*

$$d_a := \sum_{v,w \in V, v \neq w} \text{KT-dist}(v,w)\big)/(n(n-1))$$

will also play a central role in this work. Moreover, Kemeny Score remains NP-hard when the average range of candidate positions is two [4], excluding hope for fixed-parameter tractability with respect to this parameterization. Simjour [21] further introduced the parameter "Kemeny score divided by the number of votes" (also showing fixed-parameter tractability) and improved the running times for the fixed-parameter algorithms corresponding to the parameterizations by average KT-distance and Kemeny score. Recently, Karpinski and Schudy [13] devised subexponential-time fixed-parameter algorithms for the parameters Kemeny score, d_a, and Kemeny score divided by the number of votes. Mahajan et al. [17] studied above guarantee parameterization with respect to the Kemeny score. Introducing the new concept of partial kernelization, it has been shown that with respect to the average KT-distance d_a one can compute in polynomial time an equivalent instance where the number of candidates is at most $162d_a^2 + 9d_a$ [5]. This equivalent instance is called *partial kernel*[2] with respect to the parameter d_a because it only bounds the number of candidates but not the number of votes instead of bounding the total instance size (as one has in classical problem kernels). Finally, it is interesting to note that Conitzer [7] developed a powerful preprocessing technique for solving a similar rank aggregation problem (Slater ranking). His concept of similar candidates is related to our approach.

Our contributions. On the theoretical side, we improve the previous partial kernel from $162d_a^2 + 9d_a$ candidates [5] to $11d_a$ candidates. Herein, the central point is to exploit "stronger majorities", going from "$>_{2/3}$-majorities" as used before [5] to "$\geq_{3/4}$-majorities". In this line, we also prove that the consideration of "$\geq_{3/4}$-majorities" is optimal in the sense that "$\geq_s$-majorities" with $s < 3/4$ do not suffice.

On the practical side, we provide strong empirical evidence for the usefulness of data reduction rules associated with the above mentioned kernelization. An essential property of our data reduction rules is that they can break instances into several subinstances to be handled independently, that is, the relative order between the candidates in two different subinstances in a Kemeny ranking is already determined. This also means that for hard instances which we could not completely solve, we were still able to compute "partial rankings" of the top and bottom ranked candidates. Finally, we employ some of the known fixed-parameter algorithms and integer linear programming to solve sufficiently small parts of the instances remaining after data reduction.

Due to the lack of space, several details are deferred to the full version of the paper.

2 Majority-based data reduction rules

We start with some definitions and sketch some relevant previous results [5]. Then we show how to extend the previous results to obtain a linear partial kernel for the parameter average KT-distance by

[2] A formal definition of partial kernels appears in the upcoming journal version of [5].

value of s	**partial kernel result**	**sp. case: no dirty pairs**
$2/3 < s < 3/4$	quadratic partial kernel w.r.t. n_d ([5, Theorem 5])	polynomial-time solvable ([5, Theorem 4])
$3/4 \leq s \leq 1$	linear partial kernel w.r.t. n_d (Theorem 1)	

Table 1: Partial kernelization and polynomial-time solvability. The term dirty refers to the $\geq_s$-majority for the respective values of s. The number of dirty pairs is n_d. A linear partial kernel w.r.t. the average KT-distance follows directly from the linear partial kernel w.r.t. n_d (Theorem 1).

providing a new reduction rule. We also show the "limits" of our new reduction rule. Finally, we provide two more reduction rules of practical relevance.

Definitions and previous results. The data reduction framework from previous work [5] introduces a "dirtiness concept" and shows that one can delete some "non-dirty candidates" by a data reduction rule leading to a partial kernel with respect to the average KT-distance. The "dirtiness" of a pair of candidates is measured by the amount of agreement of the votes for this pair. To this end, we introduce the following notation. For an election (V, C), two candidates $c, c' \in C$, and a rational number $s \in]0.5, 1]$, we write

$$c \geq_s c'$$

if at least $\lceil s \cdot |V| \rceil$ of the votes prefer c to c'. A candidate pair $\{c, c'\}$ is *dirty according to the $\geq_s$-majority* if neither $c \geq_s c'$ nor $c' \geq_s c$. All remaining pairs are *non-dirty according to the $\geq_s$-majority*. This directly leads to the parameter number n_d of dirty pairs according to the $\geq_s$-majority. Previous work only considered $>_{2/3}$-majorities[3] and provided a reduction rule such that the number of candidates in a reduced instance is at most quadratic in n_d as well as in d_a [5]. In this work, we provide a linear partial kernel with respect to n_d according to the $\geq_s$-majority for $s \geq 3/4$ and show that this leads to a linear partial kernel with respect to d_a.

We say that c and c' are *ordered according to the $\geq_s$-majority* in a preference list l if $c \geq_s c'$ and $c > c'$ in l. If all candidate pairs are non-dirty with respect to the $\geq_s$-majority for an $s > 2/3$, then there exists a *$\geq_s$-majority order*, that is, a preference list in which all candidate pairs are ordered according to the $\geq_s$-majority [5]. Furthermore, such a $>_{2/3}$-majority can be found in polynomial time and is a Kemeny ranking [5]. Candidates appearing only in non-dirty pairs are called *non-dirty candidates* and all remaining candidates are *dirty candidates*. Note that with this definition a non-dirty pair can also be formed by two dirty candidates. See Table 1 for an overview of partial kernelization and polynomial-time solvability results.

We end with some notation needed to state our data reduction rules. For a candidate subset $C' \subseteq C$, a ranking fulfills the condition $C' > C \setminus C'$ if every candidate from C' is preferred to every candidate from $C \setminus C'$. A subinstance of (V, C) *induced* by a candidate subset $C' \subseteq C$ is given by (V', C') where every vote in V' one-to-one corresponds to a vote in V keeping the relative order of the candidates from C'.

2.1 New results exploiting $\geq_{3/4}$-majorities

We improve the partial kernel upper bound [5] for the parameter d_a from quadratic to linear, presenting a new data reduction rule. The crucial idea for the new reduction rule is to consider $\geq_{3/4}$-majorities instead of $>_{2/3}$-majorities. We further show that the new reduction rule is tight in the sense that it does not work for $>_{2/3}$-majorities.

[3] To simplify matters, we write "$>_{2/3}$" instead of "$\geq_s$ with $s > 2/3$", and if the value of s is clear from the context, then we speak of "dirty pairs" and omit "according to the $\geq_s$-majority".

value of s	properties
$1/2 \leq s \leq 2/3$	a $\geq_s$-majority order does not necessarily exist (Example 1)
$2/3 < s < 3/4$	a $\geq_s$-majority order exists (follows from [5, Theorem 4]) **but** a non-dirty candidate and a dirty candidate do not have to be ordered according to the $\geq_s$-majority in a Kemeny ranking (Theorem 2)
$3/4 \leq s \leq 1$	a $\geq_s$-majority order exists (follows from [5, Theorem 4]) **and** in every Kemeny ranking every non-dirty candidate is ordered according to the $\geq_s$-majority with respect to all other candidates (Lemma 1)

Table 2: Properties "induced" by $\geq_s$-majorities for different values of s.

Reduction rule. The following lemma allows us to formulate a data reduction rule that deletes all non-dirty candidates and additionally may break the remaining set of dirty candidates into several subsets to be handled independently from each other.

Lemma 1. *Let $a \in C$ be a non-dirty candidate with respect to the $\geq_{3/4}$-majority and $b \in C \setminus \{a\}$. If $a \geq_{3/4} b$, then in every Kemeny ranking one must have "$a > \cdots > b$"; if $b \geq_{3/4} a$, then in every Kemeny ranking one must have "$b > \cdots > a$".*

As a direct consequence of Lemma 1 we can partition the candidates of an election (V, C) as follows. Let $N := \{n_1, \ldots, n_s\}$ denote the set of non-dirty candidates with respect to the $\geq_{3/4}$-majority such that $n_i \geq_{3/4} n_{i+1}$ for $1 \leq i \leq s-1$. Then,

$$D_0 := \{d \in C \setminus N \mid d \geq_{3/4} n_1\},$$
$$D_i := \{d \in C \setminus N \mid n_i \geq_{3/4} d \text{ and } d \geq_{3/4} n_{i+1}\} \text{ for } 1 \leq i \leq s-1, \text{ and}$$
$$D_s := \{d \in C \setminus N \mid n_s \geq_{3/4} d\}.$$

3/4-Majority Rule. *Let (V, C) be an election and N and $D_0, \ldots, D_s$ be the sets of non-dirty and dirty candidates as specified above. Replace the original instance by the $s+1$ subinstances induced by D_i for $i \in \{0, \ldots, s\}$.*

The soundness of the 3/4-Majority Rule follows directly from Lemma 1 and it is straightforward to verify its running time $O(nm^2)$. An instance reduced by the 3/4-Majority Rule contains only dirty candidates with respect to the original instance. Making use of a simple relation between the number of dirty candidates and the average KT-distance as also used previously [5], one can state the following.

Theorem 1. *For* Kemeny Score *a partial kernel with less than $11 \cdot d_a$ candidates and less than $2n_d$ candidates can be computed in $O(nm^2)$ time.*

Tightness results. We investigate to which $\geq_s$-majorities the results obtained for $\geq_{3/4}$-majorities extend. An overview of properties for a Kemeny ranking for different values of s is provided in Table 2.

For the $>_{2/3}$-majority, instances without dirty candidates are polynomial-time solvable [5]. More precisely, the $>_{2/3}$-majority order is a Kemeny ranking. A simple example shows that for any $s \leq 2/3$ a $\geq_s$-majority order does not always exist:

Example 1. Consider the election consisting of the three candidates a, b, and c and the three votes "$a > b > c$", "$b > c > a$", and "$c > a > b$". Here, $a \geq_{2/3} b$, $b \geq_{2/3} c$, and $c \geq_{2/3} a$. Then, no linear order fulfills all three relations.

The existence of a data reduction rule analogously to the $3/4$-Majority Rule for $\geq_s$-majorities for $s < 3/4$ would be desirable since such a rule might be more effective: There are instances for which a candidate is dirty according to the $\geq_{3/4}$-majority but non-dirty according to a $\geq_s$-majority with $s < 3/4$. Hence, for many instances, the number n_d of dirty pairs according to the $\geq_{3/4}$-majority assumes higher values than it does according to smaller values of s. In the following, we discuss why an analogous s-Majority Rule with $s < 3/4$ cannot exist. The decisive point of the $3/4$-Majority Rule is that, in a Kemeny ranking, every non-dirty candidate must be ordered according to the $\geq_{3/4}$-majority with respect to *every* other candidate. The following theorem shows that this is not true for $\geq_s$-majorities with $s < 3/4$.

Theorem 2. *Consider a $\geq_s$-majority for any rational $s \in]2/3, 3/4[$. For a non-dirty candidate x and a dirty candidate y, $x \geq_s y$ does not imply $x > y$ in a Kemeny ranking.*

Proof. Let s_1 and s_2 be two positive integers such that $s = s_1/s_2$. We construct an election such that there is a non-dirty candidate x with $x \geq_s y$ but "$y > \cdots > x$" in every Kemeny ranking. The set of candidates is $\{x, y, a_1, a_2\}$ and there are the following $n = s_1 \cdot s_2$ votes:

- $s_1 \cdot s_2 - s_1^2$ votes of type 1: $x > y > a_1 > a_2$,
- $2s_1^2 - s_1 \cdot s_2$ votes of type 2: $a_1 > a_2 > x > y$,
- $s_1 \cdot s_2 - s_1^2$ votes of type 3: $y > a_1 > a_2 > x$.

We first show that there is a positive number of votes of every type:

Considering the number of votes of types 1 and 3, recall that $3/4 > s_1/s_2$ and thus $s_2 > 4/3 \cdot s_1$. Hence, it is easy to see that their number is $s_1 \cdot s_2 - s_1^2 > s_1 \cdot (4/3 \cdot s_1 - s_1) > 0$. Regarding votes of type 2, we use the trivial bound that $s_1/s_2 > 1/2$ and thus their number is $2s_1^2 - s_1 \cdot s_2 > s_1 \cdot (2s_1 - 2s_1) = 0$.

Now, we show that x is non-dirty and $x \geq_s y$. The number of votes with $a > x$ for $a \in \{a_1, a_2\}$ is $2s_1^2 - s_1 \cdot s_2 + s_1 \cdot s_2 - s_1^2 = s_1^2 = s \cdot n$ and the number of votes with $x > y$ is $s_1 \cdot s_2 - s_1^2 + 2s_1^2 - s_1 \cdot s_2 = s_1^2 = s \cdot n$ and thus x is non-dirty according to the $\geq_s$-majority and $x \geq_s y$.

In the following, we show that the score of "$y > a_1 > a_2 > x$" is smaller than the score of every other preference list and, hence, there is no Kemeny ranking in which x and y are ordered according to the $\geq_s$-majority.

Since "$a_1 > a_2$" in every vote, "$a_1 > a_2$" in every Kemeny ranking (see e.g. [4]). Distinguishing three cases, we first show that in every Kemeny ranking "$a_1 > x$" if and only if "$a_2 > x$", and "$a_1 > y$" if and only if "$a_2 > y$". After this, we can treat a_1 and a_2 as one candidate of "weight" two and thus with this argument there remain only six preference lists for which the score has to be investigated to show that "$y > a_1 > a_2 > x$" is the only preference list with minimum score.

Case 1: Consider a preference list with "$a_1 > x > a_2$" where y is placed either before or after all other three candidates. This preference list cannot have minimum score since swapping x and a_2 leads to a preference list with smaller score since $a_2 \geq x$ in more than $sn > 2/3 \cdot n$ votes.

Case 2: Consider a preference list with "$a_1 > y > a_2$" where x is placed either before or after all three other candidates. This preference list cannot have minimum score since swapping a_1 and y leads to a preference list with smaller score. This can be seen as follows. Since $s_1 < 3/4 \cdot s_2$, the number of votes with "$y > a_1$" is

$$2s_1s_2 - 2s_1^2 > 2s_1(s_2 - 3/4 \cdot s_2) = 1/2 \cdot s_1s_2 = n/2.$$

Case 3: Consider the preference list "$a_1 > x > y > a_2$". Note that the same preference list with x and y swapped would clearly have a larger score. We show that "$a_1 > a_2 > x > y$" has a smaller score than "$a_1 > x > y > a_2$". The only pairs that change the score are $\{a_2, y\}$ and $\{a_2, x\}$. These pairs contribute with

$$\#_v(a_2 > y) + \#_v(a_2 > x) = 2s_1^2 - s_1s_2 + 2s_1^2 - s_1s_2 + s_1s_2 - s_1^2 = 3s_1^2 - s_1s_2$$

to the old score and with $2n - \#_v(a_2 > y) - \#_v(a_2 > x)$ to the "new" score. Hence, it remains to show that the difference between the old and new score is positive, that is,

$$3s_1^2 - s_1s_2 - 2s_1s_2 + 3s_1^2 - s_1s_2 = 6s_1^2 - 4s_1s_2 > 6 \cdot 2/3 \cdot s_1s_2 - 4s_1s_2 = 0.$$

Finally, we consider the scores of all possible remaining six preference lists $r_1, \dots, r_6$ with a standing for "$a_1 > a_2$":

$r_1: \quad a > x > y$	$r_3: \quad x > a > y$	$r_5: \quad y > a > x$
$r_2: \quad a > y > x$	$r_4: \quad x > y > a$	$r_6: \quad y > x > a$

Let $t(r)$ denote the score of a preference list r. It is easy to verify that $t(r_1) < t(r_2)$, $t(r_1) < t(r_3)$, and $t(r_4) < t(r_6)$. Hence, it remains to compare the score of r_5 with the score of r_1 and r_4. Since a represents two candidates, we count the corresponding pairs twice in the following computations.

$$\begin{aligned}
&t(r_1) - t(r_5)\\
&= 2\#_v(x > a) + 2\#_v(y > a) + \#_v(y > x) - 2\#_v(a > y) - 2\#_v(x > a) - \#_v(x > y)\\
&= 2s_1s_2 - 2s_1^2 + 4s_1s_2 - 4s_1^2 + s_1s_2 - s_1^2 - 4s_1^2 + 2s_1s_2 - 2s_1s_2 + 2s_1^2 - s_2s_1 + s_1s_2\\
&= 7s_1s_2 - 5s_1^2 > 7s_1 \cdot 4/3 \cdot s_1 - 5s_1^2 = 13/3 \cdot s_1^2 > 0\\
&t(r_4) - t(r_5)\\
&= \#_v(y > x) + 2\#_v(a > x) + 2\#_v(a > y) - 2\#_v(a > y) - 2\#_v(x > a) - \#_v(x > y)\\
&= s_1s_2 - s_1^2 + 2 \cdot s_1^2 - 2 \cdot (s_1s_2) + 2 \cdot s_1^2 - s_1^2\\
&= 2s_1^2 - s_1s_2 > 2/3 \cdot s_1^2 > 0
\end{aligned}$$

This shows that r_5 has a smaller score than r_1 and r_4.

Altogether, we showed that r_5 is the only Kemeny ranking. Thus, there is an election with $x \geq_s y$ for every $s \in]2/3, 3/4[$ such that every Kemeny ranking has $y > x$. □

2.2 Exploiting the Condorcet property

We present a well-known data reduction rule of practical relevance and show that it reduces an instance at least as much as the 3/4-Majority Rule. The reduction rule is based on the following easy-to-verify observation.

Observation 1. *Let $C' \subseteq C$ be a candidate subset with $c' \geq_{1/2} c$ for every $c' \in C'$ and every $c \in C \setminus C'$. Then there must be a Kemeny ranking fulfilling $C' > C \setminus C'$.*

To turn Observation 1 into a reduction rule, we need a polynomial-time algorithm to identify appropriate "winning subsets" of candidates. We use the following simple strategy, called *winning subset routine*: For every candidate c, compute a minimal winning subset M_c by iteratively adding every candidate c' with $c' >_{1/2} c''$, $c'' \in M_c$, to M_c. After this, we choose a smallest winning subset.

Condorcet-Set Rule. *If the winning subset routine returns a subset C' with $C' \neq C$, then replace the original instance by the two subinstances induced by C' and $C \setminus C'$.*

It is easy to see that the Condorcet-Set Rule can be carried out in $O(nm^3)$ time. The following proposition shows that the Condorcet-Set Rule is at least as powerful as the $3/4$-Majority Rule, implying that the Condorcet-Set Rule provides a partial kernel with less than $11d_a$ candidates.

Proposition 1. *An instance reduced by the Condorcet-Set Rule cannot be further reduced by the* $3/4$-*Majority Rule.*

Proposition 1 shows that the $3/4$-Majority Rule cannot lead to a "stronger" reduction of an instance than the Condorcet-Set Rule does. However, since the Condorcet-Set Rule has a higher running time, that is $O(nm^3)$ compared to $O(nm^2)$, applying the $3/4$-Majority Rule before the Condorcet-Set Rule may lead to an improved running time in practice. This is also true for the consideration of the following "special case" of the Condorcet-Set Rule also running in $O(nm^2)$ time.

Condorcet Rule. *If there is a candidate $c \in C$ with $c \geq_{1/2} c'$ for every $c' \in C \setminus \{c\}$, then delete c.*

Indeed, our experiments will show that combining the Condorcet-Set Rule with the other rules significantly speeds up the practical running times for many instances.

3 Experimental results

To solve sufficiently small remaining parts of the instances left after the application of our data reduction rules, we implemented three exact algorithms. First, an extended version of the search tree algorithm showing fixed-parameter tractability with respect to the Kemeny score [4, 6]. Second, a dynamic programming algorithm running in $O(2^m \cdot nm^2)$ time for m candidates and n votes [4, 19]. Third, the integer linear program [8, Linear Program 3] which was the fastest exact algorithm in previous experimental studies [8, 20]. We use the freely available ILP-solver GLPK[4] to solve the ILP.[5]

Our algorithms are implemented in C++ using several libraries of the boost package. Our implementation consists of about 4000 lines of code. All experiments were carried out on a PC with 3 GHz and 4 GB RAM (CPU: Intel Core2Quad Q9550) running under Ubuntu 9.10 (64 bit) Linux. Source code and test date are available under the GPL Version 3 license under http://theinf1.informatik.uni-jena.de/kconsens/.

We start to describe our results for two different types of web search data (Sections 3.1 and 3.2) followed by instances obtained from sport competitions (Section 3.3).

3.1 Search result rankings

A prominent application of RANK AGGREGATION is the aggregation of search result rankings obtained from different web search engines. We queried the same 37 search terms as Dwork et al. [10] and Schalekamp and van Zuylen [20] to generate rankings. We used the search engines Google, Lycos, MSN Live Search, and Yahoo! to generate rankings of 1000 candidates. We consider two search results as identical if their URL is identical up to some canonical form (cutting after the top-level domain). Results not appearing in all rankings are ignored. Ignoring the term "zen budism" with only 18 candidates, this results in 36 instances having between 55 and 163 candidates. We start with a systematic investigation of the performance of the individual reduction rules followed by describing our results for the web instances.

We systematically applied all combinations of reduction rules, always sticking to the following rule ordering: If applied, the Condorcet-Set Rule is applied last and the $3/4$-Majority Rule is applied

[4] http://www.gnu.org/software/glpk/

[5] We omit a detailed discussion about the performance of the single algorithms. A systematic comparison of the three algorithms will be provided in the full version of this work.

	blues		gardening		classical guitar	
	time	profile	time	profile	time	profile
001	0.03	$1^2 > 5 > 1 > 101 > 1 > 2$	0.01	$1 > 2 > 1 > 102$	0.03	$1 > 114$
010	0.10	$1^{74} > 9 > 1^{29}$	0.05	$1^{54} > 43 > 1^9$	0.06	$1^6 > 92 > 1^{17}$
011	0.10	$1^{74} > 9 > 1^{29}$	0.05	$1^{54} > 43 > 1^9$	0.07	$1^6 > 92 > 1^{17}$
100	0.84	$1^{74} > 9 > 1^{29}$	0.95	$1^{54} > 20 > 1^3 > 9 > 1^{10} > 4 > 1^6$	1.89	$1^6 > 7 > 1^{50} > 35 > 1^{17}$
101	0.10	$1^{74} > 9 > 1^{29}$	1.03	$1^{54} > 20 > 1^3 > 9 > 1^{10} > 4 > 1^6$	2.03	$1^6 > 7 > 1^{50} > 35 > 1^{17}$
110	0.10	$1^{74} > 9 > 1^{29}$	0.10	$1^{54} > 20 > 1^3 > 9 > 1^{10} > 4 > 1^6$	0.19	$1^6 > 7 > 1^{50} > 35 > 1^{17}$
111	0.10	$1^{74} > 9 > 1^{29}$	0.11	$1^{54} > 20 > 1^3 > 9 > 1^{10} > 4 > 1^6$	0.18	$1^6 > 7 > 1^{50} > 35 > 1^{17}$

Figure 1: The first column encodes the combination of reduction rules used: the first digit is “1” if the Condorcet-Set Rule is applied, the second if the Condorcet Rule is applied and the last digit is “1” if the $3/4$-Majority Rule is applied. For the three instances corresponding to the search terms “blues”, “gardening”, and “classical guitar” we give the running times in seconds and the profiles describing the result of the data reduction process.

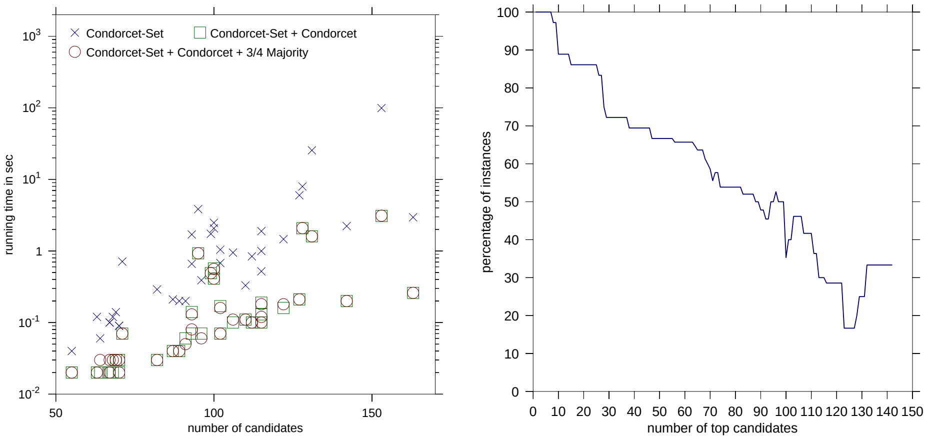

Figure 2: Left: Running times of different combinations of reduction rules. To improve readability, we omitted the data points for the Condorcet-Set Rule combined with the $3/4$-Majority Rule which was usually worse and in no case outperformed the best running times for the other combinations. Right: Percentage of the web search instances for which the x top candidates could be determined by data reduction and dynamic programming within five minutes. For a given number x of top positions, we only considered instances with at least x candidates.

first. After a successful application of the Condorcet-Set Rule, we “jump” back to the other rules (if “activated”). Examples are given in Fig. 1. This led to the following observations.

First, surprisingly, the Condorcet Rule alone led to a stronger reduction than the $3/4$-Majority Rule in most of the instances whereas the $3/4$-Majority Rule never led to a stronger reduction than the Condorcet Rule. Second, for several instances the Condorcet-Set Rule led to a stronger reduction than the other two rules, for example, for gardening and classical guitar (see Fig. 1). It led to a stronger reduction for 14 out of the 36 instances and restricted to the 15 instances with more than 100 candidates (given in Table 3), it led to a stronger reduction for eight of them. Finally, the running times for the Condorcet-Set Rule in combination with the other rules are given in the left part of Fig. 2. Applying the Condorcet Rule before the Condorcet-Set Rule led to a significant speed-up. Additionally applying the $3/4$-Majority Rule changes the running time only marginally. Note that jumping back to the “faster” rules after applying the Condorcet-Set Rule is crucial to obtain the given running times. In the following, by “our reduction rules”, we refer to all three rules applied in the order: Condorcet Rule, $3/4$-Majority Rule, and Condorcet-Set Rule.

Table 3: Web data instances with more than 100 candidates. The first column denotes the search term, the second the number of candidates, the third the running time in seconds, and the last column the "profiles" remaining after data reduction to read as follows. Every "1" stands for a position for which a candidate was determined in a Kemeny ranking and higher numbers for groups of candidates whose "internal" order could not be determined by the data reduction rules. Sequences of i ones are abbreviated by 1^i. For example, for the search term "architecture", we know the order of the best 36 candidates, then we know the set of candidates that must assume positions 37– 48 without knowledge of their relative orders, and so on.

search term	# cand.	time	structure of reduced instance		
affirmative action	127	0.21	1^{27}	$> 41 >$	1^{59}
alcoholism	115	0.10	1^{115}		
architecture	122	0.16	1^{36}	$> 12 > 1^{30} > 17 >$	1^{27}
blues	112	0.10	1^{74}	$> 9 >$	1^{29}
cheese	142	0.20	1^{94}	$> 6 >$	1^{42}
classical guitar	115	0.19	1^{6}	$> 7 > 1^{50} > 35 >$	1^{17}
Death+Valley	110	0.11	1^{15}	$> 7 > 1^{30} > 8 >$	1^{50}
field hockey	102	0.17	1^{37}	$> 26 > 1^{20} > 4 >$	1^{15}
gardening	106	0.10	1^{54}	$> 20 > 1 > 1 > 9 > 1^{8} > 4 >$	1^{9}
HIV	115	0.13	1^{62}	$> 5 > 1^{7} > 20 >$	1^{21}
lyme disease	153	3.08	1^{25}	$> 97 >$	1^{31}
mutual funds	128	2.08	1^{9}	$> 45 > 1^{9} > 5 > 1 > 49 >$	1^{10}
rock climbing	102	0.07	1^{102}		
Shakespeare	163	0.26	1^{100}	$> 10 > 1^{25} > 6 >$	1^{22}
telecommuting	131	1.60	1^{9}	$> 109 >$	1^{13}

For all instances with more than 100 candidates, the results of our reduction rules are displayed in Table 3: the data reduction rules are not only able to reduce candidates at the top and the last positions but also partition some instances into several smaller subinstances. Out of the 36 instances, 22 were solved directly by the reduction rules and one of the other algorithms in less than five minutes. Herein, the reduction rules always contributed with less than four seconds to the running time. For all other instances we still could compute the "top" and the "flop" candidates of an optimal ranking. For example, for the search term "telecommuting" there remains a subinstance with 109 candidates but we know the best nine candidates (and their order). The effectiveness in terms of top candidates of our reduction rules combined with the dynamic programming algorithm is illustrated in Fig. 2. For example, we were able to compute the top seven candidates for all instances and the top 40 candidates for 70 percent of the instances.

3.2 Impact rankings

We generated rankings that measure the "impact in the web" of different search terms. For a search engine, a list of search terms is ranked according to the number of the hits of each single term. We used Ask, Google, MSN Live Search, and Yahoo! to generate rankings for all capitals (240 candidates), all nations (242 candidates), and the 103 richest people of the world.[6] Our biggest instance is built from a list of 1349 mathematicians.[7]

As to the capitals, in less than a second, our algorithms (reduction rules and any of the other algorithms for solving subinstances up to 11 candidates) computed the following "profile" of a Kemeny ranking: $1^{45} > 34 > 1^{90} > 43 > 1^{26}$ (see Table 3 for a description of the profile concept). The final Kemeny ranking starts as follows: London $>$ Paris $>$ Madrid $>$ Singapore $>$ Berlin $> \cdots$.

[6] http://en.wikipedia.org/wiki/List_of{capitals_by_countries, richest_people}

[7] http://aleph0.clarku.edu/∼djoyce/mathhist/chronology.html

For aggregating the nation rankings, our algorithms were less successful. However, we could still compute the top 6 and the flop 12 candidates. Surprisingly, the best represented nation in the web seems to be Indonesia, followed by France, the United States, Canada, and Australia. The instance consisting of the 103 richest persons could be solved exactly in milliseconds by the data reduction rules. In contrast, for the mathematicians we could only compute the top 31 and flop 31 candidates but could not deal with a subinstance of 1287 candidates between. For the mathematicians instance, the search strategy for minimal subsets for the Condorcet-Set Rule as given in Section 2 led to a running time of more than a day. Hence, we used a cutoff of 20 candidates for the size of the minimal subsets. This decreased the running time to less than one hour.

3.3 Sport competitions

Formula 1. The winner determination of a Formula 1 season can be considered as an election where the candidates are the drivers and the votes are the single races. Currently, the winner determination is based on a "scoring rule", that is, in a single race every candidate gets some points depending on the outcome and the candidate with highest total score wins. We computed Kemeny winners for the seasons from 1970 till 2008. Since currently our implementation cannot handle ties, we only considered candidates that have competed in all races. Candidates that dropped out of a race are ordered according to the order determined by how long the drivers participated in the race. The generated instances have about 16 votes and up to 28 candidates.

Without data reduction, the ILP-approach was the most successful algorithm. It could solve all instances in less than 31 seconds whereas the dynamic programming algorithm could not solve the two instances with the highest number of candidates within 5 minutes. All search tree variants performed even worse. The Condorcet and the Condorcet-Set Rule partitioned nearly all instances in very small components such that a Kemeny ranking could be computed for all years except 1983 in few milliseconds. For 1983 (24 candidates), a remaining component with 19 candidates could be solved in less than one minute by the dynamic programming algorithm.

The Kemeny winner in most of the considered seasons is the same as the candidate selected by the used scoring rule. However, in 2008, Lewis Hamilton was elected as world champion (beating Felipe Massa by only one point) whereas Massa was the "Condorcet driver" and thus the first candidate in every Kemeny ranking. Since in contrast to Kemeny's voting system there is no scoring rule fulfilling the Condorcet property [23], this is no complete surprise.

Winter sport competitions. For ski jumping and cross skiing, we considered the world cup rankings from the seasons 2005/2006 to 2008/2009,[8] ignoring candidates not appearing in all four rankings. Without data reduction, the ski jumping instance, consisting of 33 candidates, was solved by the ILP-solver GLPK in 103 seconds whereas the search tree and dynamic programming algorithms did not find a solution within five minutes. In contrast, the instance was solved in milliseconds by only applying the reduction rules. The cross skiing instance, consisting of 69 candidates, could not be solved without data reduction within five minutes by any of our algorithms but was reduced in 0.04 seconds such that one component with 12 and one component with 15 candidates were left while all other positions could be determined by the reduction rules. The remaining parts could be solved, for example by the dynamic programming algorithm, within 0.12 and 0.011 seconds.

4 Conclusion

Our experiments showed that the described data reduction rules allow for the computation of optimal Kemeny rankings for real-world instances of non-trivial sizes within seconds. For instance, all of our larger now solved instances (with more than 50 candidates) could not be solved by the ILP,

[8] Obtained from `http://www.sportschau.de/sp/wintersport/`

the previously fastest exact algorithm [8], or the two other implemented fixed-parameter algorithms directly. A key-feature of the data reduction rules is to break instances into smaller, independent instances. A crucial observation in the experiments with the different data reduction rules regards certain cascading effects, that is, jumping back to the faster-to-execute rules after a successful application of the Condorcet-Set Rule significantly improves the running time. This shows that the order of applying data reduction rules is important. We could not observe a specific behavior of our data reduction rules for the different types of data under consideration. However, a further extension of the data sets and experiments in this direction are clearly of interest.

On the theoretical side, we improved the previous partial kernel [5] with respect to the parameter average KT-distance from quadratic to linear size. Despite the negative results from Theorem 2, there is still room for improving the $>_{2/3}$-majority based results. In particular, is there a linear partial kernel with respect to the $\geq_s$-majority for any $s < 3/4$? A natural step in answering this question seems to investigate whether for two *non-dirty* candidates a, b, there must be a Kemeny ranking with $a > b$ if $a \geq_s b$. An important extension of RANK AGGREGATION is to consider "constraint rankings", that is, the problem input additionally contains a prespecified order of some candidate pairs in the consensus list [22]. Here, our data reduction rules cannot be applied anymore. New reduction rules for this scenario could also be used in "combination" with the search tree algorithm [4] in an "interleaving mode" [18]. Other challenging variants of RANK AGGREGATION of practical interest are investigated by Ailon [1].

References

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

[2] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM*, 55(5), 2008. Article 23 (October 2008).

[3] J. Bartholdi III, 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*, 6:157–165, 1989.

[4] N. Betzler, M. R. Fellows, J. Guo, R. Niedermeier, and F. A. Rosamond. Fixed-parameter algorithms for Kemeny rankings. *Theoretical Computer Science*, 410(45):4554–4570, 2009.

[5] N. Betzler, J. Guo, C. Komusiewicz, and R. Niedermeier. Average parameterization and partial kernelization for computing medians. In *Proc. of 9th LATIN*, volume 6034 of *LNCS*, pages 60–71. Springer, 2010. Long version to appear in *Journal of Computer and System Sciences*.

[6] R. Bredereck. *Fixed-Parameter Algorithms for Computing Kemeny scores – Theory and Practice*. Studienarbeit, Universität Jena, 2010. CoRR abs/1001.4003.

[7] V. Conitzer. Computing Slater rankings using similarities among candidates. In *Proc. of 21st AAAI '06*, pages 613–619. AAAI Press, 2006.

[8] V. Conitzer, A. Davenport, and J. Kalagnanam. Improved bounds for computing Kemeny rankings. In *Proc. of 21st AAAI*, pages 620–626. AAAI Press, 2006.

[9] A. Davenport and J. Kalagnanam. A computational study of the Kemeny rule for preference aggregation. In *Proc. of 19th AAAI*, pages 697–702. AAAI Press, 2004.

[10] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the Web. In *Proc. of 10th WWW*, pages 613–622, 2001.

[11] R. Fagin, R. Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *Proc. of 22nd ACM SIGMOD*, pages 301–312. ACM, 2003.

[12] 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*, 5(2):161–171, 2008.

[13] M. Karpinski and W. Schudy. Approximation schemes for the betweenness problem in tournaments and related ranking problems. Technical report, CoRR abs/0911.2214, 2009.

[14] J. Kemeny. Mathematics without numbers. *Daedalus*, 88:571–591, 1959.

[15] C. Kenyon-Mathieu and W. Schudy. How to rank with few errors. In *Proc. of 39th STOC*, pages 95–103. ACM, 2007.

[16] A. Levenglick. Fair and reasonable election systems. *Behavioral Science*, 20(1):34–46, 1975.

[17] M. Mahajan, V. Raman, and S. Sikdar. Parameterizing above or below guaranteed values. *Journal of Computer and System Sciences*, 75:137–153, 2009.

[18] R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.

[19] V. Raman and S. Saurabh. Improved fixed parameter tractable algorithms for two "edge" problems: MAXCUT and MAXDAG. *Information Processing Letters*, 104(2):65–72, 2007.

[20] F. Schalekamp and A. van Zuylen. Rank aggregation: Together we're strong. In *Proc. of 11th ALENEX*, pages 38–51. SIAM, 2009.

[21] N. Simjour. Improved parameterized algorithms for the Kemeny aggregation problem. In *Proc. of 4th IWPEC*, volume 5917 of *LNCS*, pages 312–323. Springer, 2009.

[22] A. van Zuylen and D. P. Williamson. Deterministic pivoting algorithms for constrained ranking and clustering problems. *Mathematics of Operations Research*, 34:594–620, 2009.

[23] H. Young and A. Levenglick. A consistent extension of Condorcet's election principle. *SIAM J. Appl. Math.*, 35(2):285–300, 1978.

Nadja Betzler, Robert Bredereck, and Rolf Niedermeier
Institut für Informatik
Friedrich-Schiller-Universität Jena
Ernst-Abbe-Platz 2
D-07743 Jena, Germany
Email: `{nadja.betzler, robert.bredereck, rolf.niedermeier}@uni-jena.de`