Rank Aggregation and Kemeny Voting

Rolf Niedermeier
FG Algorithmics and Complexity Theory
Institut für Softwaretechnik und Theoretische Informatik
Fakultät IV
TU Berlin
Germany

Algorithms & Permutations, Paris, February 20, 2012

MAIN SOURCES OF THIS TALK

- Nadja Betzler, Michael R. Fellows, Jiong Guo, Rolf Niedermeier, Frances A. Rosamond: Fixed-parameter algorithms for Kemeny rankings. **Theoretical Computer Science** 410(45): 4554-4570 (2009).
- Nadja Betzler, Robert Bredereck, Rolf Niedermeier: Partial Kernelization for Rank Aggregation: Theory and Experiments. Proc. of **IPEC** 2010: 26-37.
  (Manuscript of long version available upon request.)
- Nadja Betzler, Jiong Guo, Christian Komusiewicz, Rolf Niedermeier: Average parameterization and partial kernelization for computing medians. **Journal of Computer and System Sciences** 77(4): 774-789 (2011)

Example: Select a Place for PhD Study

Choose between the following places:

- TU Berlin (B),
- MIT (M),
- Oxford University (O),
- Tsinghua University (T),
- ETH Zurich (Z).

Selection based on various criteria, leading to different rankings:

Example: Select a Place for PhD Study

Choose between the following places:

- TU Berlin (B),
- MIT (M),
- Oxford University (O),
- Tsinghua University (T),
- ETH Zurich (Z).

Selection based on various criteria, leading to different rankings:

Criterion	Ranking
Parameterized Complexity	$B \succ O \succ M \succ T \succ Z$
Salary	$Z \succ O \succ M \succ T \succ B$
Practicing English	$M \succ O \succ B \succ Z \succ T$
Cultural activities	$B \succ T \succ Z \succ M \succ O$

Goal: Aggregate the given rankings (that is, permutations) into a median ranking.

Pairwise Comparisons and Voting

Criterion	Ranking
Parameterized Complexity	$B \succ O \succ M \succ T \succ Z$
Salary	$Z \succ O \succ M \succ T \succ B$
Practicing English	$M \succ O \succ B \succ Z \succ T$
Cultural activities	$B \succ T \succ Z \succ M \succ O$

Condorcet and Kemeny:

- **Condorcet Winner**: A candidate who wins against all other candidates in pairwise comparisons. A Condorcet winner does not always exist, but is unique if it exists!

Pairwise Comparisons and Voting

Criterion	Ranking
Parameterized Complexity	B $\succ$ O $\succ$ M $\succ$ T $\succ$ Z
Salary	Z $\succ$ O $\succ$ M $\succ$ T $\succ$ B
Practicing English	M $\succ$ O $\succ$ B $\succ$ Z $\succ$ T
Cultural activities	B $\succ$ T $\succ$ Z $\succ$ M $\succ$ O

Condorcet and Kemeny:

- **Condorcet Winner**: A candidate who wins against all other candidates in pairwise comparisons. A Condorcet winner does not always exist, but is unique if it exists!
- **Kemeny**: Determine consensus ranking that minimizes the total sum of the number of “inversions” to the given rankings...

  Always yields a Condorcet winner if it exists.

On Condorcet Winner Determination

Criterion	Ranking								
Parameterized Complexity	B	$\succ$	O	$\succ$	M	$\succ$	T	$\succ$	Z
Salary	Z	$\succ$	O	$\succ$	M	$\succ$	T	$\succ$	B
Practicing English	M	$\succ$	O	$\succ$	B	$\succ$	Z	$\succ$	T
Cultural activities	B	$\succ$	T	$\succ$	Z	$\succ$	M	$\succ$	O

ON CONDORCET WINNER DETERMINATION

Criterion	Ranking								
Parameterized Complexity	B	$\succ$	O	$\succ$	M	$\succ$	T	$\succ$	Z
Salary	Z	$\succ$	O	$\succ$	M	$\succ$	T	$\succ$	B
Practicing English	M	$\succ$	O	$\succ$	B	$\succ$	Z	$\succ$	T
Cultural activities	B	$\succ$	T	$\succ$	Z	$\succ$	M	$\succ$	O

Pairs of candidates	# votes: $x \succ y$	# votes: $y \succ x$
$(x, y) =$ (B, O)	2	2
$(x, y) =$ (B, M)	2	2
$(x, y) =$ (B, T)	3	1
$(x, y) =$ (B, Z)	3	1
$(x, y) =$ (O, M)	2	2
$(x, y) =$ (O, T)	3	1
$(x, y) =$ (O, Z)	2	2
$(x, y) =$ (M, T)	3	1
$(x, y) =$ (M, Z)	2	2
$(x, y) =$ (T, Z)	2	2

Marie Jean Antoine Nicolas Caritat, Marquis de Condorcet 1743-1794

On Condorcet Winner Determination

Criterion	Ranking
Parameterized Complexity	B $\succ$ O $\succ$ M $\succ$ T $\succ$ Z
Salary	Z $\succ$ O $\succ$ M $\succ$ T $\succ$ B
Practicing English	M $\succ$ O $\succ$ B $\succ$ Z $\succ$ T
Cultural activities	B $\succ$ T $\succ$ Z $\succ$ M $\succ$ O

Pairs of candidates	# votes: $x \succ y$	# votes: $y \succ x$
$(x, y) =$ (B, O)	2	2
$(x, y) =$ (B, M)	2	2
$(x, y) =$ (B, T)	3	1
$(x, y) =$ (B, Z)	3	1
$(x, y) =$ (O, M)	2	2
$(x, y) =$ (O, T)	3	1
$(x, y) =$ (O, Z)	2	2
$(x, y) =$ (M, T)	3	1
$(x, y) =$ (M, Z)	2	2
$(x, y) =$ (T, Z)	2	2

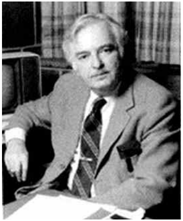
Marie Jean Antoine Nicolas Caritat, Marquis de Condorcet 1743-1794

No Condorcet winner!

Winner Determination in Kemeny voting

Criterion	Ranking								
Parameterized Complexity	B	$\succ$	O	$\succ$	M	$\succ$	T	$\succ$	Z
Salary	Z	$\succ$	O	$\succ$	M	$\succ$	T	$\succ$	B
Practicing English	M	$\succ$	O	$\succ$	B	$\succ$	Z	$\succ$	T
Cultural activities	B	$\succ$	T	$\succ$	Z	$\succ$	M	$\succ$	O

Determine consensus ranking that minimizes the total sum of the number of inversions to the given rankings...



John George Kemeny, 1926-1992.

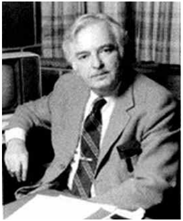
WINNER DETERMINATION IN KEMENY VOTING

Criterion	Ranking								
Parameterized Complexity	B	$\succ$	O	$\succ$	M	$\succ$	T	$\succ$	Z
Salary	Z	$\succ$	O	$\succ$	M	$\succ$	T	$\succ$	B
Practicing English	M	$\succ$	O	$\succ$	B	$\succ$	Z	$\succ$	T
Cultural activities	B	$\succ$	T	$\succ$	Z	$\succ$	M	$\succ$	O

Determine consensus ranking that minimizes the total sum of the number of inversions to the given rankings...

$\rightsquigarrow$ Two (out of 18) optimal consensus ranking with "score" 16:

- $B \succ O \succ M \succ Z \succ T$
- $O \succ M \succ B \succ T \succ Z$



John George Kemeny, 1926-1992.

KEMENY SCORE: KT-DISTANCE

Kendall Tau distance (between two votes v and w)

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

where $d_{v,w}(c,d) = \begin{cases} 0 & \text{if } v \text{ and } w \text{ rank } c \text{ and } d \text{ in the same order,} \\ 1 & \text{otherwise.} \end{cases}$

Example:

v: $a > b > c$
w: $b > c > a$

$$\begin{aligned} \text{KT-dist}(v,w) &= d_{v,w}(a,b) + d_{v,w}(a,c) + d_{v,w}(b,c) \\ &= 1 + 1 + 0 \\ &= 2 \end{aligned}$$

Central Problem: Rank Aggregation

Kemeny Score (Rank Aggregation):

Input: An set of rankings over the same candidate set and a positive integer k.
Question: Is there a ranking r with Kemeny score at most k, that is, the sum of KT-distances of r to all input rankings is at most k?

CENTRAL PROBLEM: RANK AGGREGATION

Kemeny Score (Rank Aggregation):

Input: An set of rankings over the same candidate set and a positive integer k.
Question: Is there a ranking r with Kemeny score at most k, that is, the sum of KT-distances of r to all input rankings is at most k?

Applications:

- Ranking of web sites (using meta search engines)
- Sport competitions
- Databases
- Bioinformatics

Some Results for Kemeny Score

Complexity:

- NP-complete (even for four votes)

  Bartholdi, Tovey and Tick, Social Choice and Welfare 1989,

  Dwork, Kumar, Naor, and Sivakumar, WWW 2001

Algorithms:

- factor $8/5$-approximation, randomized: factor $11/7$

  van Zuylen and Williamson, WAOA 2007,

  Ailon, Charikar, and Newman, JACM 2008

- PTAS

  Kenyon-Mathieu and Schudy, STOC 2007

- exact algorithms, heuristics, branch and bound, and experiments

  Davenport and Kalagnanam, AAAI 2004,

  Conitzer, Davenport, and Kalagnanam, AAAI 2006,

  Schalekamp and van Zuylen, ALENEX 2009,

  Ali and Meilă, Mathematical Social Sciences, 2012

The Leitmotif of Parameterized Algorithmics

Formally: "Two-dimensional analysis of complexity":

NP-hard problem X: Input size n and problem parameter k.

If there is an algorithm solving X in time

$$f(k) \cdot n^{O(1)},$$

then X is called **fixed-parameter tractable (FPT)**:

The Leitmotif of Parameterized Algorithmics

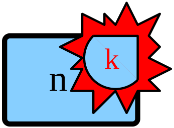
Formally: "Two-dimensional analysis of complexity":

NP-hard problem X: Input size n and problem parameter k.

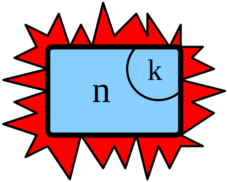
If there is an algorithm solving X in time

$$f(k) \cdot n^{O(1)},$$

then X is called **fixed-parameter tractable (FPT)**:



instead of



Parameterized Complexity Hierarchy

Completeness program developed by Downey and Fellows (1999).

$$\text{FPT} \subseteq \overbrace{\text{W[1]} \subseteq \text{W[2]} \subseteq \ldots \subseteq \text{W[P]} \subseteq \text{XP}}^{\text{Presumably fixed-parameter intractable}}$$

Parameterized Complexity Hierarchy

Completeness program developed by Downey and Fellows (1999).

$$\text{FPT} \subseteq \overbrace{\text{W[1]} \subseteq \text{W[2]} \subseteq \ldots \subseteq \text{W[P]} \subseteq \text{XP}}^{\text{Presumably fixed-parameter intractable}}$$

"Function battle" concerning allowed running time:

FPT: $f(k) \cdot n^{O(1)}$ **vs** **XP:** $f(k) \cdot n^{g(k)}$

Parameterized Complexity Hierarchy

Completeness program developed by Downey and Fellows (1999).

$$\text{FPT} \subseteq \overbrace{\text{W[1]} \subseteq \text{W[2]} \subseteq \ldots \subseteq \text{W[P]} \subseteq \text{XP}}^{\text{Presumably fixed-parameter intractable}}$$

"Function battle" concerning allowed running time:

$$\textbf{FPT: } f(k) \cdot n^{O(1)} \quad \textbf{vs} \quad \textbf{XP: } f(k) \cdot n^{g(k)}$$

Assumption: FPT $\neq$ W[1]

For instance, if W[1]=FPT then 3-SAT for a Boolean formula F with n variables can be solved in $2^{o(n)} \cdot |F|^{O(1)}$ time.

Parameterized Complexity of Kemeny Score

parameter	compl.	comment
number of votes n	NP-c	[1] for $n = 4$
number of candidates m	FPT	[2] $O^*(2^m)$
Kemeny score k	FPT	[3] $O^*(2^{O(\sqrt{k})})$
max. range of cand. pos. r_m	FPT	[2] $O^*(32^{r_m})$
avg. range of cand. pos. r_a	NP-c	[2] for $r_a \geq 2$
avg. KT-distance d_a	FPT	[4] $O^*(5.823^{d_a})$, [3] $O^*(2^{O(\sqrt{d_a})})$
partial kernel:		[5] $\frac{16}{3} \cdot d_a$ candidates
max. KT-distance d_m	FPT	[4] $O^*(4.829^{d_m})$, [3] $O^*(2^{O(\sqrt{d_m})})$

[1] Dwork, Kumar, Naor, Sivakumar, WWW 2001
[2] Betzler, Fellows, Guo, N., and Rosamond, TCS 2009
[3] Karpinski and Schudy, ISAAC 2010
[4] Simjour, IWPEC 2009
[5] Betzler, Guo, Komusiewicz, and N., JCSS 2011;
Betzler, Bredereck, and N., Manuscript of long version of IPEC 2010

PARTIAL KERNELIZATION

View Kemeny Score as a two-dimensional problem with dimensions "number *n* of votes" and "number *m* of candidates.

Basic idea:
Shrink instance into an equivalent smaller instance

- by polynomial-time executable data reduction rules such that
- the size of one "problem dimension" (that is, the number *m* of candidates here) only depends on the parameter.

Partial Kernelization

View Kemeny Score as a two-dimensional problem with dimensions "number n of votes" and "number m of candidates.

Basic idea:
Shrink instance into an equivalent smaller instance

- by polynomial-time executable data reduction rules such that
- the size of one "problem dimension" (that is, the number m of candidates here) only depends on the parameter.

Recall:

- Kemeny Score is NP-hard for $n = 4$ and
- Kemeny Score is fixed-parameter tractable with respect to m. ($O^*(2^m)$ dynamic programming algorithm.)

Partial Kernel for Kemeny Score

Idea based on $3/4$-majority relations:

- Find candidate pairs that are in the same relative order in at least 3/4 of the votes.
- Their relative order in every Kemeny consensus is then fixed analogously.

PARTIAL KERNEL FOR KEMENY SCORE

Idea based on $3/4$-majority relations:

- Find candidate pairs that are in the same relative order in at least 3/4 of the votes.
- Their relative order in every Kemeny consensus is then fixed analogously.

Definition

A candidate c is **non-dirty** if for every other candidate c' either $c' \geq_{3/4} c$ or $c \geq_{3/4} c'$. Otherwise c is **dirty**.

Partial Kernel for Kemeny Score

Idea based on $3/4$-majority relations:

- Find candidate pairs that are in the same relative order in at least 3/4 of the votes.
- Their relative order in every Kemeny consensus is then fixed analogously.

Definition

A candidate c is **non-dirty** if for every other candidate c' either $c' \geq_{3/4} c$ or $c \geq_{3/4} c'$. Otherwise c is **dirty**.

Lemma

For a non-dirty candidate c and candidate $c' \in C \setminus \{c\}$:
If $c \geq_{3/4} c'$, then $c > c'$ in every Kemeny consensus.
If $c' \geq_{3/4} c$, then $c' > c$ in every Kemeny consensus.

PARTIAL KERNEL FOR KEMENY SCORE

Idea based on $3/4$-majority relations:

- Find candidate pairs that are in the same relative order in at least 3/4 of the votes.
- Their relative order in every Kemeny consensus is then fixed analogously.

Definition

A candidate c is **non-dirty** if for every other candidate c' either $c' \geq_{3/4} c$ or $c \geq_{3/4} c'$. Otherwise c is **dirty**.

Lemma

For a non-dirty candidate c and candidate $c' \in C \setminus \{c\}$:
If $c \geq_{3/4} c'$, then $c > c'$ in every Kemeny consensus.
If $c' \geq_{3/4} c$, then $c' > c$ in every Kemeny consensus.

Data Reduction Rule

If there is a non-dirty candidate c, then delete c and partition the instance into two subinstances accordingly.

Reduction rules using "majority relations"

$a_1 > a_2 > a_3 > c > b_1 > b_2$

$a_3 > a_2 > c > a_1 > b_2 > b_1$

$a_1 > c > a_2 > b_2 > b_1 > a_3$

$a_2 > a_3 > a_1 > b_1 > b_2 > c$

$a_i \geq_{3/4} c$ and $c \geq_{3/4} b_i$

$\Rightarrow$

in every Kemeny consensus:

$\{a_1, a_2, a_3\} > c > \{b_1, b_2\}$

REDUCTION RULES USING "MAJORITY RELATIONS"

$a_1 > a_2 > a_3 > c > b_1 > b_2$

$a_3 > a_2 > c > a_1 > b_2 > b_1$

$a_1 > c > a_2 > b_2 > b_1 > a_3$

$a_2 > a_3 > a_1 > b_1 > b_2 > c$

$a_i \geq_{3/4} c$ and $c \geq_{3/4} b_i$

$\Rightarrow$

in every Kemeny consensus:

$\{a_1, a_2, a_3\} > c > \{b_1, b_2\}$

$a_1 > a_2 > a_3$	c	$b_1 > b_2$
$a_3 > a_2 > a_1$	c	$b_2 > b_1$
$a_1 > a_2 > a_3$	c	$b_2 > b_1$
$a_2 > a_3 > a_1$	c	$b_1 > b_2$

Reduction rules using "majority relations"

$a_1 > a_2 > a_3 > c > b_1 > b_2$

$a_3 > a_2 > c > a_1 > b_2 > b_1$

$a_1 > c > a_2 > b_2 > b_1 > a_3$

$a_2 > a_3 > a_1 > b_1 > b_2 > c$

$a_i \geq_{3/4} c$ and $c \geq_{3/4} b_i$

$\Rightarrow$

in every Kemeny consensus:

$\{a_1, a_2, a_3\} > c > \{b_1, b_2\}$

$a_1 > a_2 > a_3$	c	$b_1 > b_2$
$a_3 > a_2 > a_1$	c	$b_2 > b_1$
$a_1 > a_2 > a_3$	c	$b_2 > b_1$
$a_2 > a_3 > a_1$	c	$b_1 > b_2$

Further (extended) rule:
Data reduction based on non-dirty **sets** of candidates...

Reduction rules using "majority relations"

$a_1 > a_2 > a_3 > c_1 > c_2 > b_1 > b_2$

$a_3 > a_2 > c_2 > c_1 > a_1 > b_2 > b_1$

$a_1 > c_1 > c_2 > a_2 > b_2 > b_1 > a_3$

$a_2 > a_3 > a_1 > b_1 > b_2 > c_2 > c_1$

$a_i \geq_{3/4} c_j$ and $c_j \geq_{3/4} b_i$

$\Rightarrow$

in every Kemeny consensus:

$\{a_1, a_2, a_3\} > \{c_1, c_2\} > \{b_1, b_2\}$

Reduction rules using "majority relations"

$$a_1 > a_2 > a_3 > c_1 > c_2 > b_1 > b_2$$
$$a_3 > a_2 > c_2 > c_1 > a_1 > b_2 > b_1$$
$$a_1 > c_1 > c_2 > a_2 > b_2 > b_1 > a_3$$
$$a_2 > a_3 > a_1 > b_1 > b_2 > c_2 > c_1$$

$a_i \geq_{3/4} c_j$ and $c_j \geq_{3/4} b_i$

$\Rightarrow$

in every Kemeny consensus:

$$\{a_1, a_2, a_3\} > \{c_1, c_2\} > \{b_1, b_2\}$$

Three subinstances (one for the non-dirty set):

$a_1 > a_2 > a_3$	$c_1 > c_2$	$b_1 > b_2$
$a_3 > a_2 > a_1$	$c_2 > c_1$	$b_2 > b_1$
$a_1 > a_2 > a_3$	$c_1 > c_2$	$b_2 > b_1$
$a_2 > a_3 > a_1$	$c_2 > c_1$	$b_1 > b_2$

REDUCTION RULES USING "MAJORITY RELATIONS"

$a_1 > a_2 > a_3 > c_1 > c_2 > b_1 > b_2$

$a_3 > a_2 > c_2 > c_1 > a_1 > b_2 > b_1$

$a_1 > c_1 > c_2 > a_2 > b_2 > b_1 > a_3$

$a_2 > a_3 > a_1 > b_1 > b_2 > c_2 > c_1$

$a_i \geq_{3/4} c_j$ and $c_j \geq_{3/4} b_i$

$\Rightarrow$

in every Kemeny consensus:

$\{a_1, a_2, a_3\} > \{c_1, c_2\} > \{b_1, b_2\}$

Three subinstances (one for the non-dirty set):

$a_1 > a_2 > a_3$	$c_1 > c_2$	$b_1 > b_2$
$a_3 > a_2 > a_1$	$c_2 > c_1$	$b_2 > b_1$
$a_1 > a_2 > a_3$	$c_1 > c_2$	$b_2 > b_1$
$a_2 > a_3 > a_1$	$c_2 > c_1$	$b_1 > b_2$

Such sets can be found in polynomial time.

Average KT-distance as parameter for Kemeny Score

Parameter: average KT-distance between the input votes

$$d_a := \frac{2}{n(n-1)} \cdot \sum_{\{u,v\} \subseteq V} \text{KT-dist}(u, v).$$

Average KT-distance as parameter for Kemeny Score

Parameter: average KT-distance between the input votes

$$d_a := \frac{2}{n(n-1)} \cdot \sum_{\{u,v\} \subseteq V} \text{KT-dist}(u, v).$$

Theorem

A Kemeny Score instance with average KT-distance d_a can be reduced in polynomial time to an equivalent instance with less than $\frac{16}{3} \cdot d_a$ candidates.

In parameterized terms: Kemeny Score yields a partial kernel with $\frac{16}{3} \cdot d_a$ candidates.

What about other majorities, why $3/4$?

Lemma

For a non-dirty candidate c and candidate $c' \in C \setminus \{c\}$:
If $c \geq_{3/4} c'$, then $c > c'$ in every Kemeny consensus.
If $c' \geq_{3/4} c$, then $c' > c$ in every Kemeny consensus.

Observation

Lemma does not hold when we replace $3/4$ by any smaller value. We can construct counterexamples where lemma does not hold.

What about other majorities, why $3/4$?

Lemma

For a non-dirty candidate c and candidate $c' \in C \setminus \{c\}$:
If $c \geq_{3/4} c'$, then $c > c'$ in every Kemeny consensus.
If $c' \geq_{3/4} c$, then $c' > c$ in every Kemeny consensus.

Observation

Lemma does not hold when we replace $3/4$ by any smaller value. We can construct counterexamples where lemma does not hold.

As to $>_{2/3}$**-majorities**...:

- Kemeny Score is polynomial-time solvable if there are no dirty candidates;

WHAT ABOUT OTHER MAJORITIES, WHY $3/4$?

Lemma

For a non-dirty candidate c and candidate $c' \in C \setminus \{c\}$:
If $c \geq_{3/4} c'$, then $c > c'$ in every Kemeny consensus.
If $c' \geq_{3/4} c$, then $c' > c$ in every Kemeny consensus.

Observation

Lemma does not hold when we replace $3/4$ by any smaller value. We can construct counterexamples where lemma does not hold.

As to $>_{2/3}$**-majorities**...:

- Kemeny Score is polynomial-time solvable if there are no dirty candidates;
- quadratic partial kernel with respect to the number of dirty candidates;

WHAT ABOUT OTHER MAJORITIES, WHY $3/4$?

Lemma
For a non-dirty candidate c and candidate $c' \in C \setminus \{c\}$:
If $c \geq_{3/4} c'$, then $c > c'$ in every Kemeny consensus.
If $c' \geq_{3/4} c$, then $c' > c$ in every Kemeny consensus.

Observation
Lemma does not hold when we replace $3/4$ by any smaller value. We can construct counterexamples where lemma does not hold.

As to $>_{2/3}$**-majorities**...:

- Kemeny Score is polynomial-time solvable if there are no dirty candidates;
- quadratic partial kernel with respect to the number of dirty candidates;
- open: is there a partial linear kernel with respect to the number of dirty candidates?

Counterexample Against Using $5/7$-Majorities

2 votes: $x > y > a > b > c > d > e > f$

3 votes: $a > b > c > d > e > f > x > y$

2 votes: $y > a > b > c > d > e > f > x$

Counterexample Against Using $5/7$-Majorities

2 votes: $x > y > a > b > c > d > e > f$

3 votes: $a > b > c > d > e > f > x > y$

2 votes: $y > a > b > c > d > e > f > x$

- x is non-dirty according to the $\geq_{5/7}$-majority, since "$x > y$" in five out of seven votes and "{a,b,c,d,e,f}$>x$" in five out of seven votes

Counterexample Against Using $5/7$-Majorities

2 votes: $\boxed{x} > \boxed{y} > a > b > c > d > e > f$

3 votes: $a > b > c > d > e > f > \boxed{x} > \boxed{y}$

2 votes: $\boxed{y} > a > b > c > d > e > f > \boxed{x}$

- $\boxed{x}$ is non-dirty according to the $\geq_{5/7}$-majority, since "$\boxed{x} > \boxed{y}$" in five out of seven votes and "{a,b,c,d,e,f}$> \boxed{x}$" in five out of seven votes
- Although $\boxed{x} \geq_{5/7} \boxed{y}$, the only ranking with minimum Kemeny score is: $\boxed{y} > a > b > c > d > e > f > \boxed{x}$

Counterexample Against Using 5/7-Majorities

2 votes: $x > y > a > b > c > d > e > f$

3 votes: $a > b > c > d > e > f > x > y$

2 votes: $y > a > b > c > d > e > f > x$

- x is non-dirty according to the $\geq_{5/7}$-majority, since "$x > y$" in five out of seven votes and "{a,b,c,d,e,f}$>$ x" in five out of seven votes
- Although $x \geq_{5/7} y$, the only ranking with minimum Kemeny score is: $y > a > b > c > d > e > f > x$

Remarks:

- Similar (a little more technical) counterexamples can be found for every majority ratio in $]2/3, 3/4[$.
- For majority ratios $s \leq 2/3$, the $\geq_s$-majority relation is not necessarily transitive...

Data Reduction Based on Condorcet

Definition
A candidate c beating every other candidate in at least half of the votes, that is, $c \geq_{1/2} c'$ for every candidate $c' \neq c$, is called **weak Condorcet winner**.

Data Reduction Based on Condorcet

Definition

A candidate c beating every other candidate in at least half of the votes, that is, $c \geq_{1/2} c'$ for every candidate $c' \neq c$, is called **weak Condorcet winner**.

A weak Condorcet winner takes the first position in at least one Kemeny consensus (Condorcet property).

Data Reduction Based on Condorcet

Definition

A candidate c beating every other candidate in at least half of the votes, that is, $c \geq_{1/2} c'$ for every candidate $c' \neq c$, is called **weak Condorcet winner**.

A weak Condorcet winner takes the first position in at least one Kemeny consensus (Condorcet property).

Reduction Rule

If there is a weak Condorcet winner in an election provided by a Kemeny Score instance, then delete this candidate.

DATA REDUCTION BASED ON CONDORCET

Definition

A candidate c beating every other candidate in at least half of the votes, that is, $c \geq_{1/2} c'$ for every candidate $c' \neq c$, is called **weak Condorcet winner**.

A weak Condorcet winner takes the first position in at least one Kemeny consensus (Condorcet property).

Reduction Rule

If there is a weak Condorcet winner in an election provided by a Kemeny Score instance, then delete this candidate.

A **Condorcet loser** is defined analogously. Again, this rule can be extended to a rule searching for "Condorcet winner/loser sets"...

Effectiveness of Condorcet rules

Example:

$a_1 > a_2 > a_3 > c_1 > c_2 > b_1 > b_2$

$a_3 > a_2 > c_2 > c_1 > a_1 > b_2 > b_1$ $\qquad a_i \geq_{3/4} c_j$ and $c_j \geq_{3/4} b_i$

$a_1 > c_1 > c_2 > a_2 > b_2 > b_1 > a_3$ $\qquad \Rightarrow$

$a_2 > a_3 > a_1 > b_1 > b_2 > c_2 > c_1$

Effectiveness of Condorcet rules

Example:

$a_1 > a_2 > a_3 > c_1 > c_2 > b_1 > b_2$

$a_3 > a_2 > c_2 > c_1 > a_1 > b_2 > b_1$ $\quad a_i \geq_{3/4} c_j$ and $c_j \geq_{3/4} b_i$

$a_1 > c_1 > c_2 > a_2 > b_2 > b_1 > a_3$ $\quad \Rightarrow$

$a_2 > a_3 > a_1 > b_1 > b_2 > c_2 > c_1$

$\{a_1, a_2, a_3\}$ is a Condorcet winner set, and

$\{b_1, b_2\}$ is a Condorcet loser set.

Effectiveness of Condorcet rules

Example:

$a_1 > a_2 > a_3 > c_1 > c_2 > b_1 > b_2$

$a_3 > a_2 > c_2 > c_1 > a_1 > b_2 > b_1$ $\quad$ $a_i \geq_{3/4} c_j$ and $c_j \geq_{3/4} b_i$

$a_1 > c_1 > c_2 > a_2 > b_2 > b_1 > a_3$ $\quad$ $\Rightarrow$

$a_2 > a_3 > a_1 > b_1 > b_2 > c_2 > c_1$

$\{a_1, a_2, a_3\}$ is a Condorcet winner set, and

$\{b_1, b_2\}$ is a Condorcet loser set.

Fact:

The rule searching for Condorcet sets are at least as effective as the majority-based rules.

Such sets can be found in polynomial time.

DATA REDUCTION RULES APPLIED

Running time comparison for four data reduction rules:
non-dirty candidates $<^1$ Condorcet candidates $<^2$ non-dirty sets $<^1$ Condorcet sets

Heuristic combination of the data reduction rules

1. If exists, eliminate a non-dirty candidate.
2. Otherwise, if exists, eliminate a Condorcet candidate.
3. Otherwise, if exists, eliminate a non-dirty set.
4. Otherwise, if exists, eliminate a Condorcet set.

---

[1] empirical, [2] provable

Data Reduction Rules Applied

Running time comparison for four data reduction rules:
non-dirty candidates $<^1$ Condorcet candidates $<^2$ non-dirty sets $<^1$ Condorcet sets

Heuristic combination of the data reduction rules

1. If exists, eliminate a non-dirty candidate.
2. Otherwise, if exists, eliminate a Condorcet candidate.
3. Otherwise, if exists, eliminate a non-dirty set.
4. Otherwise, if exists, eliminate a Condorcet set.

instance	Condorcet set alone	heuristic combination above
blues	0.84 sec	0.10 sec
gardening	0.95 sec	0.11 sec
classical guitar	1.89 sec	0.18 sec

[1]empirical, [2]provable

Reduction of Metasearch engine data

Four votes: Google, Lycos, MSN Live Search, and Yahoo!

top 1000 hits each, candidates that appear in all four rankings

search term	cand.	sec.	red. inst. solved/unsol.
aff. action	127	0.41	[27] > **41** > [59]
alcoholism	115	0.21	[115]
architecture	122	0.47	[36] > **12** > [30] > **17** > [27]
blues	112	0.16	[74] > **9** > [29]
cheese	142	0.39	[94] > **6** > [42]
class. guitar	115	1.12	[6] > **7** > [50] > **35** > [17]
Death Valley	110	0.25	[15] > **7** > [30] > **8** > [50]
field hockey	102	0.21	[37] > **26** > [20] > **4** > [15]
gardening	106	0.19	[54] > **20** > [2] > **9** > [8] > **4** > [9]
HIV	115	0.26	[62] > **5** > [7] > **20** > [21]
lyme disease	153	2.61	[25] > **97** > [31]
mutual funds	128	3.33	[9] > **45** > [9] > **5** > [1] > **49** > [10]
rock climbing	102	0.12	[102]
Shakespeare	163	0.68	[100] > **10** > [25] > **6** > [22]
telecomm.	131	2.28	[9] > **109** > [13]

Exact Solutions using Data Reduction & ILPs

Strongest (fastest) empirical results with combination of our data reduction rules and an ILP formulation...

Exact Solutions using Data Reduction & ILPs

Strongest (fastest) empirical results with combination of our data reduction rules and an ILP formulation...

ILP formulation of Kemeny Score[1] using

- C for the set of candidates;
- coefficients $\#_{a>b}$ for the number of rankings having "$a > b$";
- binary variables $x_{a>b}$ if "$a > b$" in a Kemeny consensus.

minimize $\Sigma_{\{a,b\}\subseteq C}\ \#_{a>b} \cdot x_{a>b} + \#_{b>a} \cdot x_{b>a}$

Exact Solutions using Data Reduction & ILPs

Strongest (fastest) empirical results with combination of our data reduction rules and an ILP formulation...

ILP formulation of Kemeny Score[1] using

- C for the set of candidates;
- coefficients $\#_{a>b}$ for the number of rankings having "$a > b$";
- binary variables $x_{a>b}$ if "$a > b$" in a Kemeny consensus.

minimize $\Sigma_{\{a,b\} \subseteq C} \#_{a>b} \cdot x_{a>b} + \#_{b>a} \cdot x_{b>a}$

subject to
for all $\{a, b\} \subseteq C$: $x_{a>b} + x_{b>a} = 1$

Exact Solutions using Data Reduction & ILPs

Strongest (fastest) empirical results with combination of our data reduction rules and an ILP formulation...

ILP formulation of Kemeny Score[1] using

- C for the set of candidates;
- coefficients $\#_{a>b}$ for the number of rankings having "$a > b$";
- binary variables $x_{a>b}$ if "$a > b$" in a Kemeny consensus.

minimize $\Sigma_{\{a,b\}\subseteq C}\ \#_{a>b} \cdot x_{a>b} + \#_{b>a} \cdot x_{b>a}$

subject to

for all $\{a,b\} \subseteq C$: $x_{a>b} + x_{b>a} = 1$

for all $\{a,b,c\} \subseteq C$: $x_{a>b} + x_{b>c} + x_{c>a} \geq 1$

- First conditions are to ensure that either "$a > b$" or "$b > a$" (for fixed a and b);
- second conditions are to ensure transitivity.

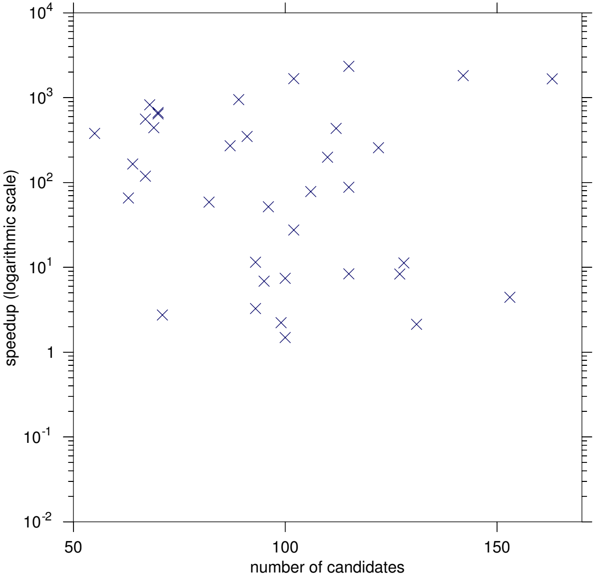
[1] [Conitzer, Davenport, Kalagnanam, AAAI 2006]

Exactly Solving Real-World Instances

Observations:

- In our experiments, no combinatorial (fixed-parameter) algorithm for exactly solving Kemeny Score could compete with the ILP-based solver (gurobi).
- Instances with hundreds of candidates can be solved within few seconds.
- Data reduction used as preprocessing led to siginificant speedups when compared to using the ILP alone.

SPEEDUP OF ILP THROUGH DATA REDUCTION



CONCLUDING REMARKS

- Key feature of our data reduction:
  Break instances into smaller, independent parts.
- Execution order of data reduction rule execution has significant impact on efficiency.
- “Cascading effects” of data redution rules not well understood.

Concluding Remarks

- Key feature of our data reduction:
  Break instances into smaller, independent parts.
- Execution order of data reduction rule execution has significant impact on efficiency.
- “Cascading effects” of data redution rules not well understood.

Some Challenges:

- Improved analysis using Condorcet rules?
- Linear partial kernel for “s-majorities” with $s < 3/4$?
- Our data redcution rules do not apply to “**constraint rankings**” where the input also contains some canadiate pairs whose relative ordering in the consensus ranking is already fixed...

CONCLUDING REMARKS

- Key feature of our data reduction:
  Break instances into smaller, independent parts.
- Execution order of data reduction rule execution has significant impact on efficiency.
- "Cascading effects" of data redution rules not well understood.

Some Challenges:

- Improved analysis using Condorcet rules?
- Linear partial kernel for "s-majorities" with $s < 3/4$?
- Our data redcution rules do not apply to "**constraint rankings**" where the input also contains some canadiate pairs whose relative ordering in the consensus ranking is already fixed...

Merci!