How similarity helps to efficiently compute Kemeny rankings.

Nadja Betzler
Friedrich-Schiller-Universität Jena

joint work with

Michael R. Fellows, Jiong Guo, Rolf Niedermeier, and Frances A. Rosamond

AAMAS 2009

Rank aggregation/Kemeny rankings

- **Meta-search engines**
  How to aggregate the results of several search engines into a consensus ranking?
- **Recommendation scenarios**
  How to aggregate viewers' rankings of movies?
  How to aggregate rankings based on different criteria, like price, quality, . . . ?
- **Sports and competitions**
  How to aggregate the results of different competitions to determine the winner of a season?
- **Data base middleware**
  How to aggregate results from multiple databases?
- . . .

Kemeny ranking

Election

Set of votes V, set of candidates C.
A vote is a ranking (total order) over all candidates.

Example: $C = \{a, b, c\}$

vote 1:	a	$>$	b	$>$	c
vote 2:	a	$>$	c	$>$	b
vote 3:	b	$>$	c	$>$	a

How to aggregate the votes into a “consensus ranking”?

KT-distance

KT-distance (between two votes v and w)

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

where $d_{v,w}(c, d)$ is 0 if v and w rank c and d in the same order, 1 otherwise.

Example:

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

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

Kemeny Consensus

Kemeny score of a ranking r:

sum of KT-distances between r and all votes

Kemeny consensus r_{con}:

a ranking that minimizes the Kemeny score

$$
\begin{array}{lll}
v_1: & a > b > c & \text{KT-dist}(r_{con}, v_1) = 0 \\
v_2: & a > c > b & \text{KT-dist}(r_{con}, v_2) = 1 \text{ because of } \{b, c\} \\
v_3: & b > c > a & \text{KT-dist}(r_{con}, v_3) = 2 \text{ because of } \{a, b\} \text{ and } \{a, c\} \\
 & & \\
r_{con}: & \mathbf{a} > \mathbf{b} > \mathbf{c} & \text{Kemeny score: } 0 + 1 + 2 = 3
\end{array}
$$

Motivation

Applications:

- internet: meta search engines, spam detection

  [Dwork et al., WWW 2001]
- databases

  [Fagin et al., SIGMOD, 2003]
- bioinformatics

  [Jackson et al., IEEE/ACM Transactions on Computational Biology and Bioinformatics 2008]

Kemeny is the only voting system that is

- neutral,
- consistent, and
- Condorcet.

Known results

KEMENY SCORE is NP-complete (even for 4 votes)
[DWORK ET AL., WWW 2001]

Algorithms:

- randomized factor 11/7-approximation
  [AILON ET AL., J. ACM 2008]
- factor 8/5-approximation
  [VAN ZUYLEN AND WILLIAMSON, WAOA 2007]
- PTAS [KENYON-MATHIEU AND SCHUDY, STOC 2007]
- Heuristics; greedy, branch and bound
  [DAVENPORT AND KALAGNANAM, AAAI 2004],
  [CONITZER ET AL., AAAI 2006]

Parameterized Complexity

Given an NP-hard problem with input size n and a parameter k
Basic idea: Confine the combinatorial explosion to k

Definition

A problem of size n is called *fixed-parameter tractable* with respect to a parameter k if it can be solved in $f(k) \cdot n^{O(1)}$ time.

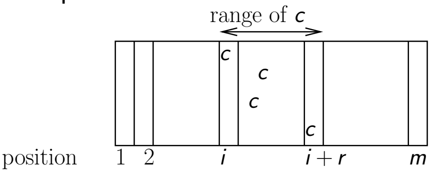
Parameterizations of Kemeny Score

Number of votes n [Dwork et al. WWW 2001]	NP-c for $n = 4$
Number of candidates m [Betzler et al. AAIM 2008]	$O^*(2^m)$
Kemeny score k [Betzler et al. AAIM 2008]	$O^*(1.53^k)$

Parameterizations of Kemeny Score

Number of votes n [Dwork et al. WWW 2001]	NP-c for $n = 4$
Number of candidates m [Betzler et al. AAIM 2008]	$O^*(2^m)$
Kemeny score k [Betzler et al. AAIM 2008]	$O^*(1.53^k)$

Further “structural” parameters:

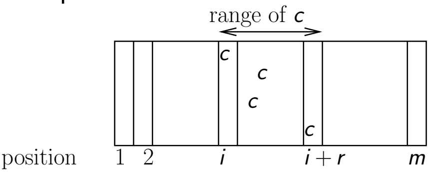


Maximum range $r_m := \max_{c \in C} \text{range}(c)$	$O^*(32^{r_m})$
Average range r_a	NP-c for $r_a \geq 2$

Parameterizations of Kemeny Score

Number of votes n [Dwork et al. WWW 2001]	NP-c for $n = 4$
Number of candidates m [Betzler et al. AAIM 2008]	$O^*(2^m)$
Kemeny score k [Betzler et al. AAIM 2008]	$O^*(1.53^k)$

Further "structural" parameters:



Maximum range $r_m := \max_{c \in C} \text{range}(c)$	$O^*(32^{r_m})$
Average range r_a	NP-c for $r_a \geq 2$

Average KT-distance

Average KT-distance

Recall: The KT-distance between two votes is the number of inversions or "conflict pairs".

Definition

For an election (V, C) the average KT-distance d_a is defined as

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

In the following, we show that Kemeny Score is fixed-parameter tractable with respect to the "average KT-distance".

Complementarity of parameterizations

- Number of candidates m: $O^*(2^m)$
- Maximum range r of candidate positions in the input votes: $O^*(32^r)$
- Average distance of the input votes: $O^*(16^{d_a})$

($m \geq r$, but corresponding algorithm has a better running time)

Complementarity of parameterizations

- Number of candidates m: $O^*(2^m)$
- Maximum range r of candidate positions in the input votes: $O^*(32^r)$
- Average distance of the input votes: $O^*(16^{d_a})$

($m \geq r$, but corresponding algorithm has a better running time)

Example 1: small range, large number of candidates and average distance

a	$>$	c	$>$	b	$>$	e	$>$	d	$>$	f	$\dots$
b	$>$	a	$>$	c	$>$	d	$>$	e	$>$	f	$\dots$
b	$>$	c	$>$	a	$>$	e	$>$	f	$>$	d	$\dots$

Example 2: small average distance, large number of candidates and range

a	$>$	b	$>$	c	$>$	d	$>$	e	$>$	f	$\dots$	
b	$>$	c	$>$	d	$>$	e	$>$	f	$>$	$\dots$		a
a	$>$	b	$>$	c	$>$	d	$>$	e	$>$	f	$\dots$	

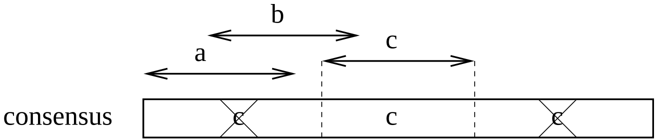
$\Rightarrow$ check size of parameter and then use appropriate strategy

Basic idea

Average distance d_a.

Crucial observation

In every Kemeny consensus every candidate can only assume a number of consecutive positions that is bounded by $2 \cdot d_a$.



Dynamic programming

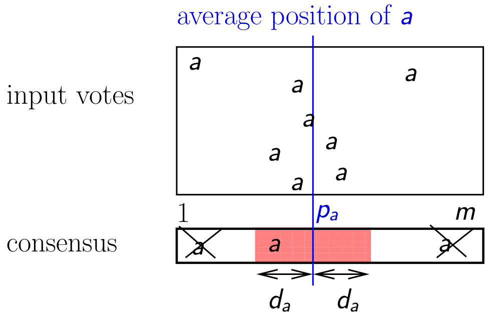
making use of the fact that every candidate can be "forgotten" or "inserted" at a certain position.

Crucial observation

Let the average position of a candidate c be $p_a(c)$.

Lemma

Let d_a be the average KT-distance of an election (V, C). Then, in every optimal Kemeny consensus r_{con}, for every candidate $c \in C$ we have $p_a(c) - d_a < r_{con}(c) < p_a(c) + d_a$.



Crucial observation

Let the average position of a candidate c be $p_a(c)$.

Lemma

Let d_a be the average KT-distance of an election (V, C). Then, in every optimal Kemeny consensus r_{con}, for every candidate $c \in C$ we have $p_a(c) - d_a < r_{con}(c) < p_a(c) + d_a$.

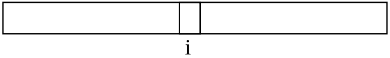
Idea of proof:

1. "The Kemeny score of (V, C) is smaller than $d_a \cdot |V|$."
   We show that one of the input votes has this Kemeny score.
2. Contradiction: Assume a candidate has a position outside the given range. Then, we can show that the Kemeny score is greater than $d_a \cdot |V|$, a contradiction.

Dynamic programming

One can show that the set P_i of candidates that can take a position i has size at most $4d_a$.

consensus



$P_i = \{a, b, c, d, e, f\}$

Observation:

For any position i and a subset P_i of candidates that can assume i:

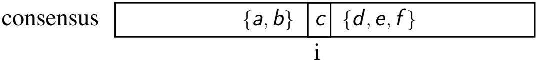
- One candidate of P_i must assume position i in a consensus.
- Every other candidate of P_i must be either left or right of i.

Running time

n votes

m candidates

$P_i = \{a, b, c, d, e, f\}$



We have $|P_i| \leq 4d_a$, thus there are at most 2^{4d_a} subsets of P_i.

$\Rightarrow$ Table size is bounded by $16^{d_a} \cdot \text{poly}(n, m)$.

Theorem

Kemeny Score can be solved in $O(16^d \cdot \text{poly}(n, m))$ time with average KT-distance d_a and $d := \lceil d_a \rceil$.

Overview of parameterized complexity

	Kemeny Score
Number of votes n [Dwork et al. WWW 2001]	NP-c for $n = 4$
Kemeny score k	$O^*(1.53^k)$
Number of candidates m	$O^*(2^m)$
Maximum range of candidate positions r	$O^*(32^r)$
Average range of candidate positions r_a	NP-c for $r_a \geq 2$
Average KT-distance d_a	$O^*(16^{d_a})$

Outlook

- Average distance: investigate typical values.
- Improve the running time for the parameterizations "average distance" and "maximum candidate range".
- Implementation of the algorithms is under way.
- Consider generalizations like incomplete votes and ties.
- NP-completeness of KEMENY SCORE with 3 votes?