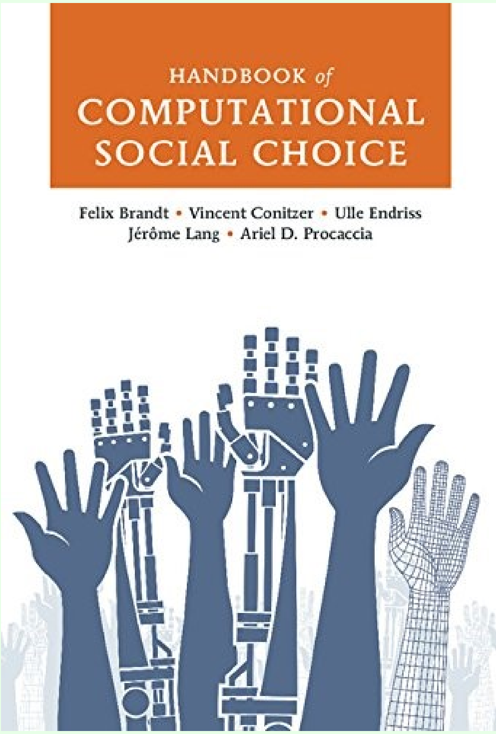
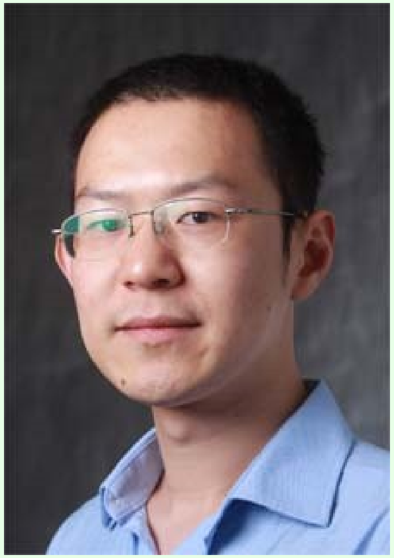
Computational Social Choice and Moral Artificial Intelligence

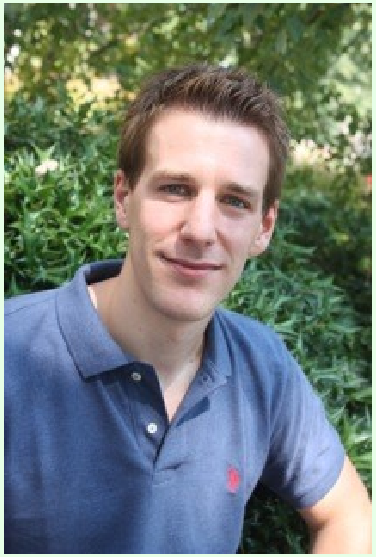
Vincent Conitzer, Duke University

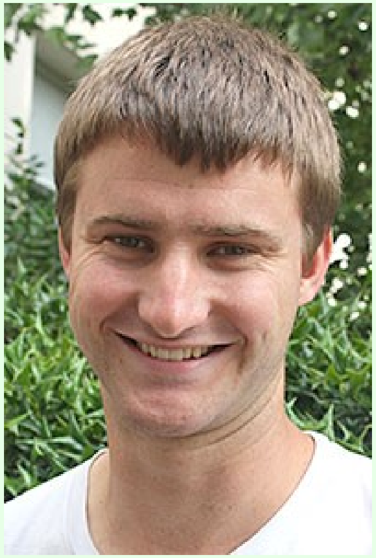
IJCAI-ECAI 2018 tutorial

comsoc mailing list: https://lists.duke.edu/sympa/subscribe/comsoc









Lirong Xia (Ph.D. 2011, now at RPI)

Markus Brill (postdoc 2013-2015, now at TU Berlin)

Rupert Freeman (Ph.D. 2018, joining MSR NYC for postdoc)

some slides based on

Ulle Endriss

Voting

n voters…

… each produce a ranking of *m* alternatives…

$b \succ a \succ c$

$a \succ c \succ b$

$a \succ b \succ c$

… which a social preference function (SPF) maps to one or more aggregate rankings.

$a \succ b \succ c$

… or, a social choice function (SCF) just produces one or more winners.

a

Plurality

1 0 0

$b \succ a \succ c$

$a \succ c \succ b$

$a \succ b \succ c$

$a \succ b \succ c$

2 1 0

Borda

$2 \quad 1 \quad 0$

$b \succ a \succ c$

$a \succ c \succ b$

$a \succ b \succ c$

$a \succ b \succ c$

$5 \quad 3 \quad 1$

Instant runoff voting / single transferable vote (STV)

$b \succ a \succ c$

$a \succ b \succ c$

$a \succ b \succ b$

$a \succ b \succ c$

Kemeny

$b \succ a \succ c$

$a \succ b \succ c$

$a \succ c \succ b$

2 disagreements

$\leftrightarrow$

3*3 - 2 = 7 agreements

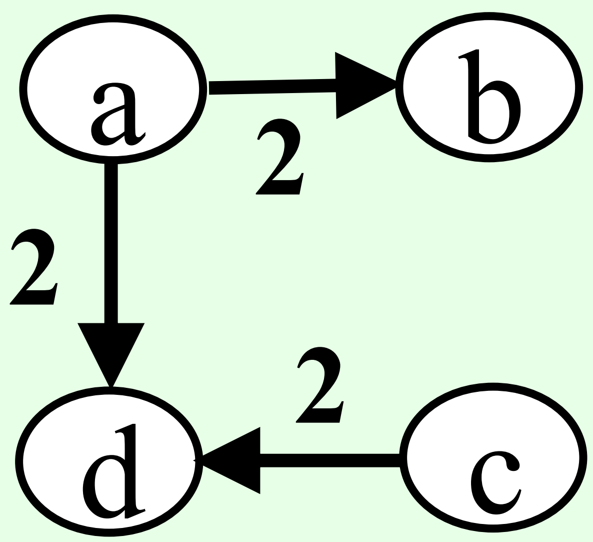
(maximum)

$a \succ b \succ c$

- Natural interpretation as maximum likelihood estimate of the "correct" ranking [Young 1988, 1995]

Pairwise election graphs

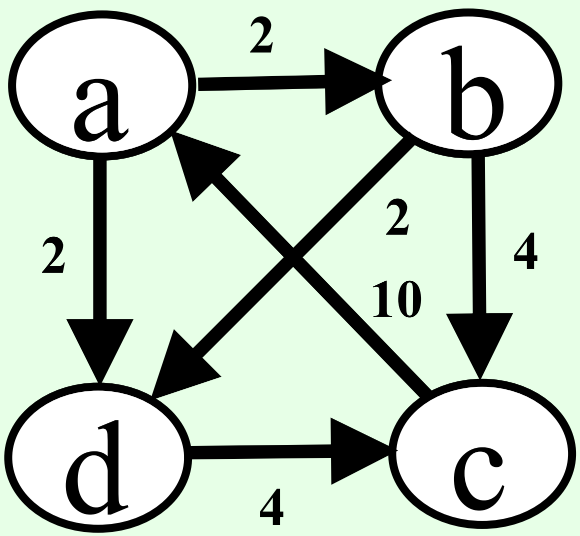
- Pairwise election between *a* and *b*: compare how often *a* is ranked above *b* vs. how often *b* is ranked above *a*
- Graph representation: edge from winner to loser (no edge if tie), weight = margin of victory
- E.g., for votes *a* > *b* > *c* > *d*, *c* > *a* > *d* > *b* this gives



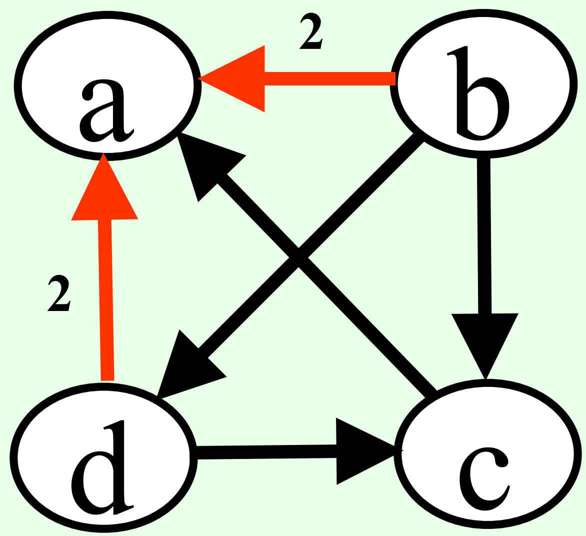
Kemeny on pairwise election graphs

- Final ranking = acyclic tournament graph
  - Edge (a, b) means a ranked above b
  - Acyclic = no cycles, tournament = edge between every pair
- Kemeny ranking seeks to minimize the total weight of the inverted edges

pairwise election graph



Kemeny ranking

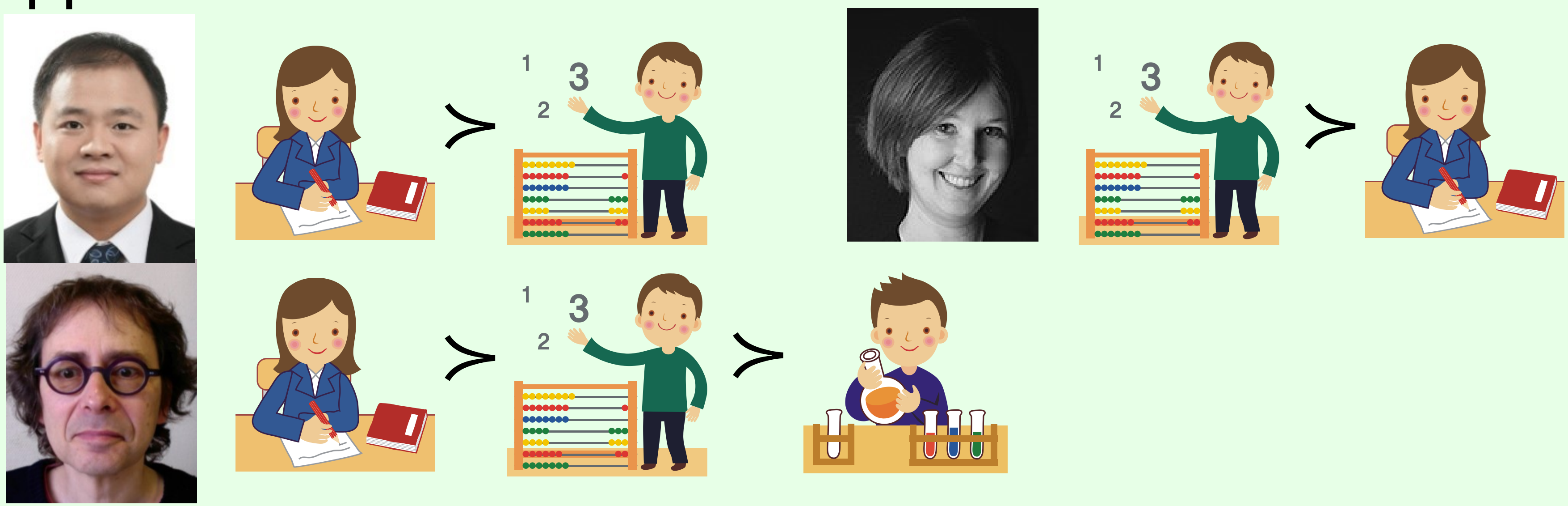


(b > d > c > a)

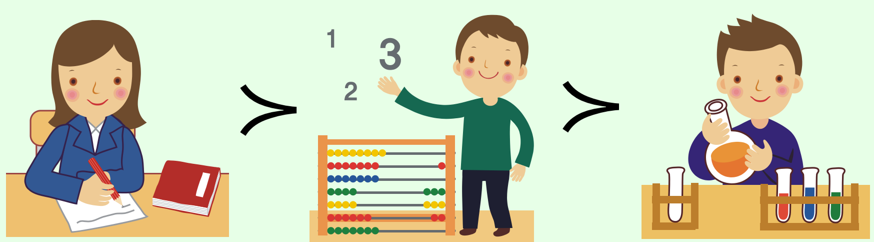
- NP-hard even with 4 voters [Dwork et al. 2001]
- Integer programs scale reasonably [C., Davenport, Kalagnanam 2006]

Ranking Ph.D. applicants
(briefly described in C. [2010])

- Input: Rankings of subsets of the (non-eliminated) applicants



- Output: (one) Kemeny ranking of the (non-eliminated) applicants



Choosing a rule

- How do we choose a rule from all of these rules?
- How do we know that there does not exist another, “perfect” rule?
- Let us look at some criteria that we would like our voting rule to satisfy

Condorcet criterion

- A candidate is the Condorcet winner if it wins all of its pairwise elections
- Does not always exist…
- … but the Condorcet criterion says that if it does exist, it should win

- Many rules do not satisfy this
- E.g., for plurality:
  - b > a > c > d
  - c > a > b > d
  - d > a > b > c
- a is the Condorcet winner, but it does not win under plurality

Consistency (SPF sense)

- An SPF f is said to be consistent if the following holds:
  - Suppose V_1 and V_2 are two voting profiles (multisets) such that f produces the same ranking on both
  - Then f should produce the same ranking on their union.
- Which of our rules satisfy this?

Consistency (S**C**F sense)

- An SCF f is said to be consistent if the following holds:
  - Suppose V_1 and V_2 are two voting profiles (multisets) such that f produces the same **winner** on both
  - Then f should produce the same winner on their union.
- Which of our rules satisfy this?
- Consistency properties are closely related to interpretability as MLE of the truth [C., Rognlie, Xia 2009]

Some axiomatizations

- **Theorem** [Young 1975]. An SCF is symmetric, consistent, and continuous if and only if it is a positional scoring rule.
- **Theorem** [Young and Levenglick 1978]. An SPF is neutral, consistent, and Condorcet if and only if it is the Kemeny SPF.
- **Theorem** [Freeman, Brill, C. 2014]. An SPF satisfies independence of bottom alternatives, consistency at the bottom, independence of clones (& some minor conditions) if and only if it is the STV SPF.

Manipulability

- Sometimes, a voter is better off revealing her preferences insincerely, AKA manipulating
- E.g., plurality
  - Suppose a voter prefers a > b > c
  - Also suppose she knows that the other votes are
    - 2 times b > c > a
    - 2 times c > a > b
  - Voting truthfully will lead to a tie between b and c
  - She would be better off voting, e.g., b > a > c, guaranteeing b wins

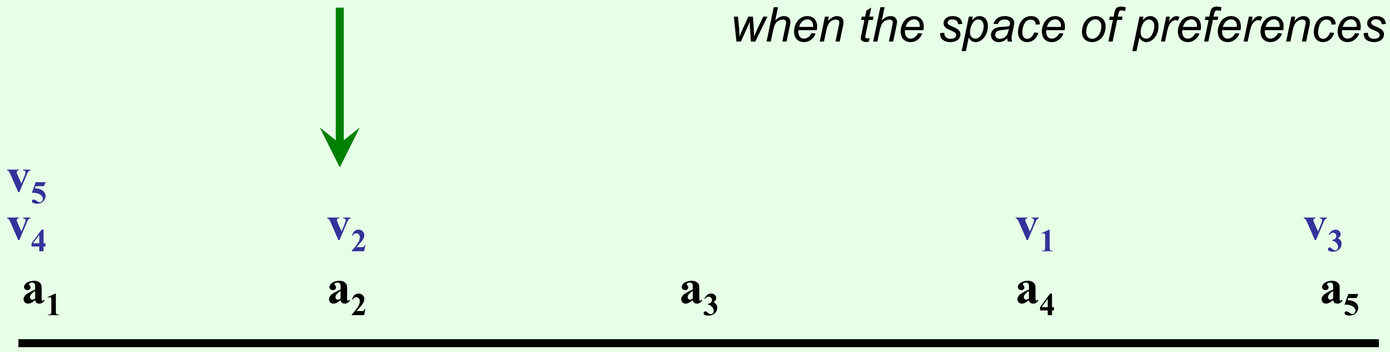
Gibbard-Satterthwaite impossibility theorem

- Suppose there are at least 3 alternatives
- There exists no rule that is simultaneously:
  - non-imposing/onto (for every alternative, there are some votes that would make that alternative win),
  - nondictatorial (there does not exist a voter such that the rule simply always selects that voter's first-ranked alternative as the winner), and
  - nonmanipulable/strategy-proof

Single-peaked preferences

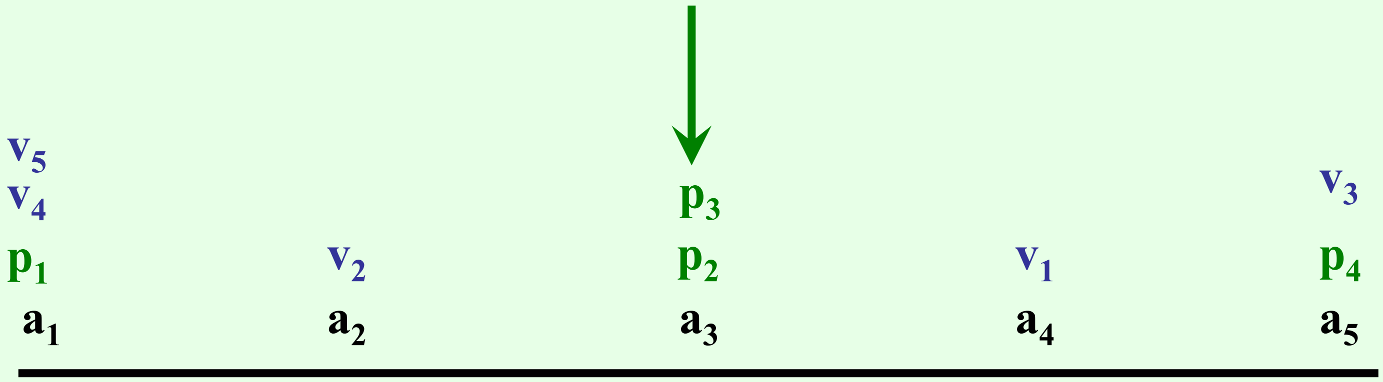
- Suppose candidates are ordered on a line
- Every voter prefers candidates that are closer to her most preferred candidate
- Let every voter report only her most preferred candidate ("peak")
- Choose the median voter's peak as the winner
  - This will also be the Condorcet winner
- Nonmanipulable!

Impossibility results do not necessarily hold when the space of preferences is restricted



Moulin's characterization

- Slight generalization: add phantom voters, then choose the median of real+phantom voters
- **Theorem** [Moulin 1980]. Under single-peaked preferences, an SCF is strategy-proof, Pareto efficient, and anonymous if and only if it is such a generalized median rule.



Computational hardness as a barrier to manipulation

- A (successful) manipulation is a way of misreporting one’s preferences that leads to a better result for oneself
- Gibbard-Satterthwaite only tells us that for some instances, successful manipulations exist
- It does not say that these manipulations are always easy to find
- Do voting rules exist for which manipulations are computationally hard to find?

A formal computational problem

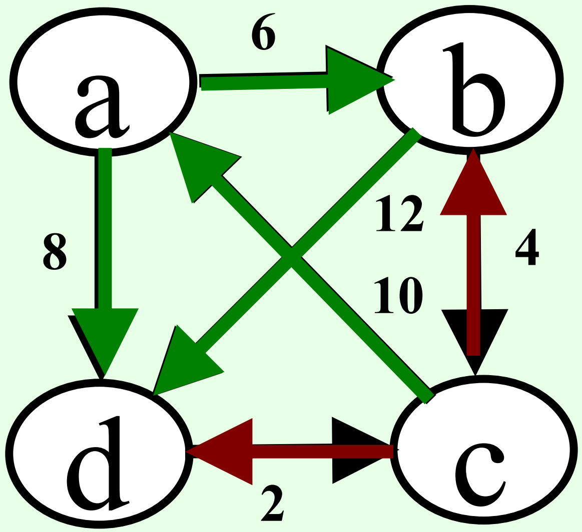
- The simplest version of the manipulation problem:
- CONSTRUCTIVE-MANIPULATION:
  - We are given a voting rule *r*, the (unweighted) votes of the other voters, and an alternative *p*.
  - We are asked if we can cast our (single) vote to make *p* win.
- E.g., for the Borda rule:
  - Voter 1 votes A > B > C
  - Voter 2 votes B > A > C
  - Voter 3 votes C > A > B
- Borda scores are now: A: 4, B: 3, C: 2
- Can we make B win?
- Answer: YES. Vote B > C > A (Borda scores: A: 4, B: 5, C: 3)

Early research

- **Theorem.** CONSTRUCTIVE-MANIPULATION is NP-complete for the second-order Copeland rule. [Bartholdi, Tovey, Trick 1989]
  - Second order Copeland = alternative's score is sum of Copeland scores of alternatives it defeats
- **Theorem.** CONSTRUCTIVE-MANIPULATION is NP-complete for the STV rule. [Bartholdi, Orlin 1991]
- Most other rules are easy to manipulate (in P)

Ranked pairs rule [Tideman 1987]

- Order pairwise elections by decreasing strength of victory
- Successively "lock in" results of pairwise elections unless it causes a cycle



Final ranking:
c>a>b>d

- **Theorem.** CONSTRUCTIVE-MANIPULATION is NP-complete for the ranked pairs rule [Xia et al. IJCAI 2009]

Many manipulation problems…

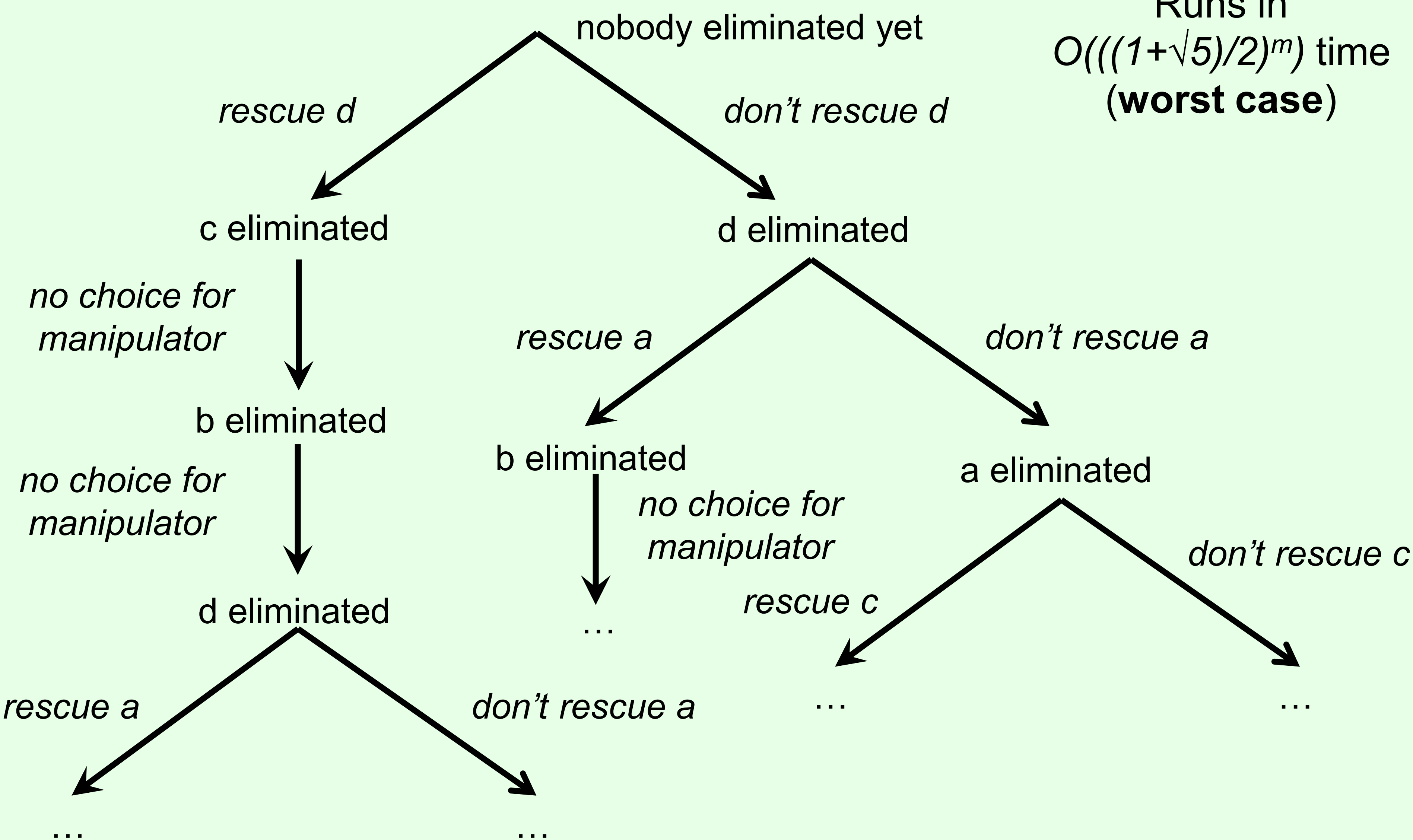
	unweighted votes, constructive manipulation		weighted votes,						
			constructive				destructive		
# alternatives			2	3	4	≥5	2	3	≥4
# manipulators	1	≥ 2							
plurality	P	P	P	P	P	P	P	P	P
plurality with runoff	P	P	P	NP-c	NP-c	NP-c	P	NP-c	NP-c
veto	P	P	P	NP-c	NP-c	NP-c	P	P	P
cup	P	P	P	P	P	P	P	P	P
Copeland	P	P	P	P	NP-c	NP-c	P	P	P
Borda	P	NP-c	P	NP-c	NP-c	NP-c	P	P	P
Nanson	NP-c	NP-c	P	P	NP-c	NP-c	P	P	NP-c
Baldwin	NP-c	NP-c	P	NP-c	NP-c	NP-c	P	NP-c	NP-c
Black	P	NP-c	P	NP-c	NP-c	NP-c	P	P	P
STV	NP-c	NP-c	P	NP-c	NP-c	NP-c	P	NP-c	NP-c
maximin	P	NP-c	P	P	NP-c	NP-c	P	P	P
Bucklin	P	P	P	NP-c	NP-c	NP-c	P	P	P
fallback	P	P	P	P	P	P	P	P	P
ranked pairs	NP-c	NP-c	P	P	P	NP-c	P	P	?
Schulze	P	P	P	P	P	P	P	P	P

Table from: C. & Walsh, Barriers to Manipulation, Chapter 6 in *Handbook of Computational Social Choice*

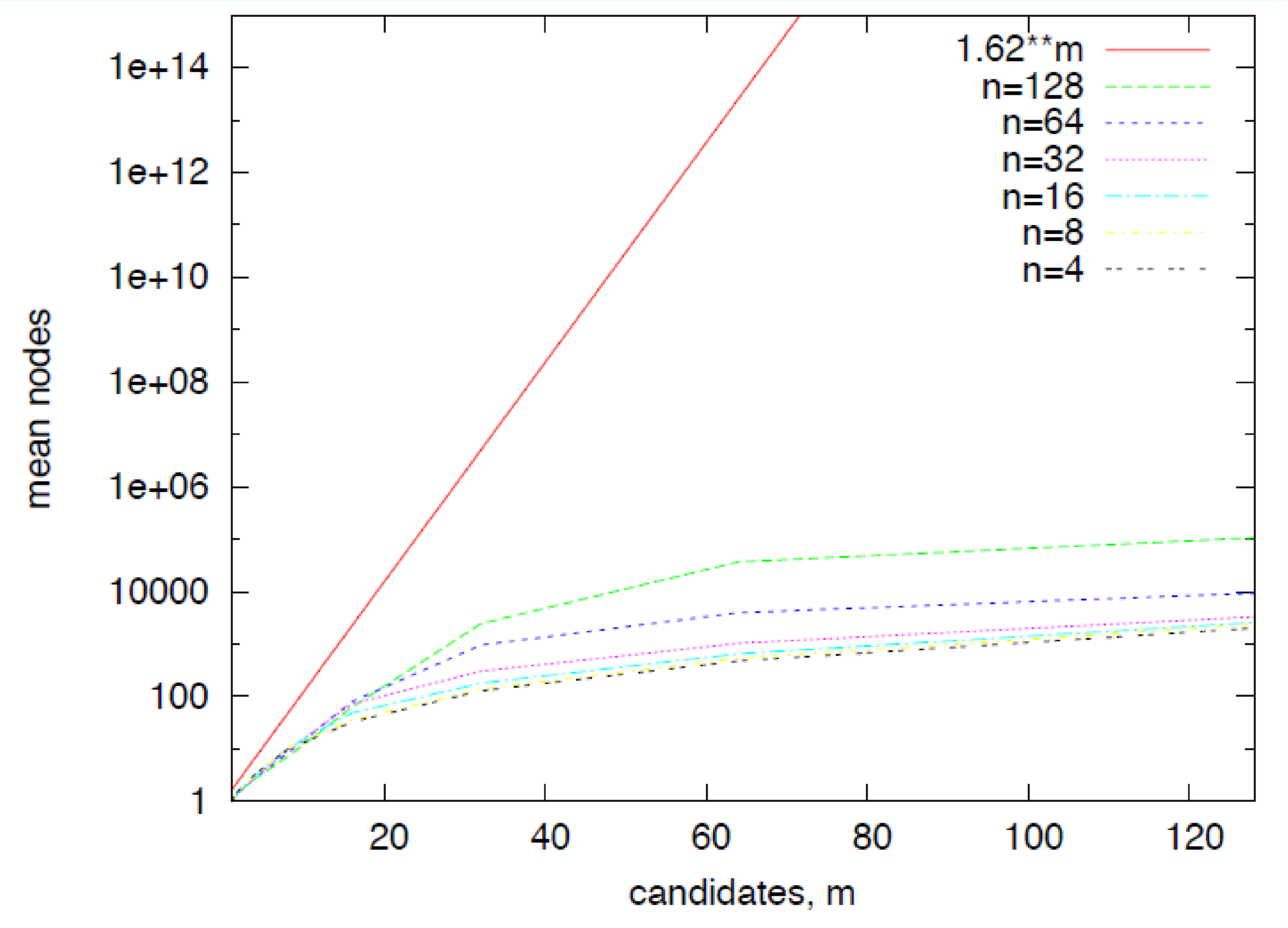
STV manipulation algorithm

[C., Sandholm, Lang JACM 2007]

Runs in
$O(((1+\sqrt{5})/2)^m)$ time
(**worst case**)



Runtime on random votes [Walsh 2011]



Fine – how about another rule?

- Heuristic algorithms and/or experimental (simulation) evaluation [C. & Sandholm 2006, Procaccia & Rosenschein 2007, Walsh 2011, Davies, Katsirelos, Narodytska, Walsh 2011]

- Quantitative versions of Gibbard-Satterthwaite showing that under certain conditions, for some voter, even a random manipulation on a random instance has significant probability of succeeding [Friedgut, Kalai, Nisan 2008; Xia & C. 2008; Dobzinski & Procaccia 2008; Isaksson, Kindler, Mossel 2010; Mossel & Racz 2013]

"for a social choice function f on k≥3 alternatives and n voters, which is ϵ-far from the family of nonmanipulable functions, a uniformly chosen voter profile is manipulable with probability at least inverse polynomial in n, k, and ϵ^{-1}."

Judgment aggregation

[for an overview, see Ulle Endriss' chapter 17 in the computational social choice handbook]

- Three judges have to decide on a case of an alleged breach of contract
- They need to decide (a) whether the contract is **valid** and (b) whether the contract has been **breached**.
- Legal doctrine stipulates that the defendant is **liable** if and only if (a) and (b) hold.

	Valid?	*Breach?*	*Liable?*
Judge 1	**Yes**	**Yes**	**Yes**
Judge 2	**Yes**	**No**	**No**
Judge 3	**No**	**Yes**	**No**

Why is this considered a paradox?

	p	q	$p \wedge q$
Judge 1	**Yes**	**Yes**	**Yes**
Judge 2	**Yes**	**No**	**No**
Judge 3	**No**	**Yes**	**No**
Majority	**Yes**	**Yes**	**No**

- **Reason 1: Premise-based procedure** and **conclusion-based procedure** produce different outcomes.
- **Reason 2:** Even though each individual judgment is logically consistent, the **majority outcome** is not.

Distance-based rules

- **Idea:** Find a **consistent** judgment set that minimizes the "distance" to the profile
- **Hamming distance** between two judgment sets is given by the number of disagreements
  - distance to a profile given by sum of distances to individual judgment sets in the profile

p	*q*	*r*
No	**No**	**Yes**
Yes	**Yes**	**Yes**

- Two ways to define aggregation rule based on Hamming distance:
  - minimize Hamming distance to profile ← generalized Kemeny rule
  - minimize Hamming distance to majority outcome ← generalized Slater rule

Example

ϕ_1 and ϕ_2 are both equivalent to
$p \vee (q_1 \wedge q_2) \vee (r_1 \wedge r_2 \wedge r_3)$

	p	q_1	q_2	r_1	r_2	r_3	ϕ_1	ϕ_2
1 agent	Yes	No	No	No	No	No	Yes	Yes
10 agents	No	Yes	Yes	No	No	No	Yes	Yes
10 agents	No	No	No	Yes	Yes	Yes	Yes	Yes

Kemeny	No	Yes	Yes	No	No	No	Yes	Yes
Slater	Yes	No	No	No	No	No	Yes	Yes