Social Choice and Welfare
https://doi.org/10.1007/s00355-023-01469-z

ORIGINAL PAPER

On the safety of group manipulation

Hans Peters[1] · Yuliya Veselova[2,3]

Received: 28 October 2022 / Accepted: 12 May 2023
© The Author(s) 2023

Abstract
Groups of voters have more possibilities to influence the voting result than separate individuals. However, there is a problem with coordinating their actions. This paper considers manipulation by groups of voters who have the same preferences. If a voting result is more preferable for voters of a particular group provided that all its members use the same strategy (report the same insincere preference), then each of these members has an incentive to manipulate. If there is a chance that they will become worse off in case only a subset of the whole group manipulates, then manipulation is unsafe. For several voting rules we study conditions on the numbers of voters and alternatives which allow for an unsafe manipulation or which make manipulation always safe.

1 Introduction

One of the problems with collective decision making is that voters may submit insincere preferences, aiming to achieve a more preferable result or, in other words, manipulate an election. Manipulation, therefore, is one of the most considered questions in social choice theory. The fundamental result in this direction is the Gibbard-Satterthwaite theorem, which states that every non-dictatorial social choice rule with at least three alternatives in its range, is vulnerable to

The authors thank Ton Storcken and two reviewers for many valuable comments. The article was prepared within the framework of the HSE University Basic Research Program.

✉ Hans Peters
h.peters@maastrichtuniversity.nl

Yuliya Veselova
yul-r@mail.ru

[1] Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands

[2] International Centre of Decision Choice and Analysis, The National Research University Higher School of Economics, Moscow, Russia

[3] Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow 117997, Russia

Published online: 28 June 2023

individual manipulation (Gibbard 1973; Satterthwaite 1975). This result applies to the basic manipulation model, where voters manipulate individually and independently, having complete information about the preferences of the other voters.

While all being vulnerable to manipulation, rules still differ in their degree of manipulability, see for instance (Nitzan 1985; Kelly 1988; Aleskerov and Kurbanov 1999; Maus et al. 2007), and Peters et al. (2012). The present paper studies a somewhat related question. Obviously, groups of voters have even more opportunities to influence the result: they can unite in coalitions and coordinate their manipulation. In many applications, however, this coordination cannot and does not actually take place explicitly. Rather, a voter who aims to manipulate, may take into account that other voters with the same preference may also manipulate in the same way. In fact, this is what we assume in this paper. Given a (sincere) preference within a profile of preferences, we will use the word 'group' to indicate all voters who have this preference. We then say that a(ny) voter in this group has an incentive to manipulate if there is a(n insincere) preference such that, if all voters in this group report this preference, then the election result is better for them according to the true, sincere preference. In that case, a problem may arise if not all voters in the group participate in the manipulation, because if this happens the result may actually be worse than without manipulation. In other words, due to lack of or poor communication within a group of like-minded voters, manipulation may be harmful.

We will call manipulation 'safe' if this does not happen: even if not all voters in a group participate in the manipulation, the result is not worse than without manipulation. Otherwise, manipulation is 'unsafe'. We now provide an example of such an unsafe manipulation for the well-known Borda rule.

Example 1.1 Suppose that there are five alternatives, a, b, c, d, e. A preference profile with seven voters is given in the following table. The first line of the table shows the number of voters for each preference order occurring in the profile.

3	2	1	1
a	d	d	e
b	c	c	d
c	b	e	a
e	e	a	c
d	a	b	b

The Borda rule assigns 4 points to the top alternatives, 3 points to the second-ranked alternatives, etc., until 0 points to the last ranked alternatives, and these points are then added up to obtain the total scores. For the given preference profile, total scores are $S(a) = 15$, $S(b) = 13$, $S(c) = 16$, $S(d) = 15$, $S(e) = 11$. Thus, with sincere preferences, alternative c wins. For the group of three voters $K = \{1, 2, 3\}$ each one having preference (a, b, c, e, d) (i.e., a is preferred to b, b to c, etc.), there is no way to make a win, but they have an incentive to manipulate by reporting preference (b, a, e, c, d). If all voters in K report this preference, then the scores will be $S(a) = 12$, $S(b) = 16$, $S(c) = 13$, $S(d) = 15$, $S(e) = 14$, so that b is the winner of the election, and b is preferred over c by the members of K according to their sincere preference.

Now suppose that only one voter of K decides to manipulate. In this case, the scores are: $S(a) = 14$, $S(b) = 14$, $S(c) = 15$, $S(d) = 15$, $S(e) = 12$. Alternatives c and d have maximal scores. If we assume that d wins against c by tie-breaking, then the final outcome is d, but for the members of K outcome d is worse than c. Therefore, this group manipulation is unsafe. $\lhd$

If manipulation is unsafe, then this fact may prevent voters from voting strategically. However, the possibility of an unsafe manipulation depends on the rule, the number of voters and the number of alternatives. In this paper we consider a collection of well-known rules and investigate for which of these rules group manipulation can be unsafe, and which rules are only safely manipulable.

The concept of (un)safe manipulation has already been considered by Slinko and White (2014). However, their model differs from the one considered in this paper. In their approach, a voter i in a group K has an incentive to manipulate if there is some subset of K, including voter i, such that the election result improves for i if exactly the voters in this subset report a (the same) insincere preference. They call manipulation 'unsafe' if the result can get worse if some other subset, including i, deviates. The main result in Slinko and White (2014) is an extension of the Gibbard-Satterthwaite theorem: for each rule with at least three alternatives in its range, there is a preference profile and a voter who can safely individually manipulate – that is, this voter is not worse off if also some other voters with the same preference manipulate in the same way. We postpone a more elaborate comparison between our paper and Slinko and White (2014) until Sect. 7.1.

Following up on the model of Slinko and White (2014), several papers focus on different aspects of safe manipulation. Computational complexity of finding a safe strategic vote under k-approval and Bucklin rules was studied in Hazon and Elkind (2010). The same question for Borda rule and some classes of scoring rules was considered in Ianovski et al. (2011). The asymptotic probability of a safe manipulation under the IAC assumption (all voting profiles are equally likely) for scoring rules is computed in Wilson and Reyhani Shokat Abad (2010). In an extension of the aforementioned model each manipulator thinks not only about his/her allies, but about all voters having an incentive to manipulate (they are called Gibbard-Satterthwaite-manipulators, or GS-manipulators). Then, a strategy is considered as 'safe' if for any manipulating subset of GS-manipulators, using this strategy is not worse than sincere voting. This kind of model was considered in Elkind et al. (2015) and Grandi et al. (2019). These references are just a few from the strand of literature on voting manipulation games. For a more detailed survey we refer the reader to Slinko (2019).

The rest of the paper is organized as follows. Section 2 presents the formal model and the rules that we consider: scoring rules, in particular Borda; run-off; Copeland; and single transferable vote. Section 3 considers scoring rules in general and Borda in particular, Sect. 4 considers the run-off rule, Sect. 5 the Copeland rule, and Sect. 6 single transferable vote. Section 7 concludes, in particular with a comparison between (Slinko and White 2014) and our approach.

2 Definitions and notations

2.1 The framework

A society of $n \geq 3$ *voters*, $N = \{1, \dots, n\}$, decides which of m *alternatives* from the set X, $|X| = m \geq 3$, to choose.[1] Each voter has a *preference*, i.e., a linear order[2] on X. We denote the set of all preferences by $L(X)$. For $a, b \in X$ and $P \in L(X)$ we write aPb instead of $(a, b) \in P$. Also, we often write $P = (a_1, \dots, a_m)$, meaning that $a_1 P \dots P a_m$, where $X = \{a_1, \dots, a_m\}$. A *preference profile* is a vector $\mathbf{P} = (P_1, \dots, P_n) \in L(X)^N$ of individual preferences.

A *social choice correspondence* (SCC) is a map $C : L(X)^N \to 2^X \backslash \{\emptyset\}$ (where 2^X denotes the set of all subsets of X). A *social choice rule* or simply *rule* is a map $F : L(X)^N \to X$. Thus, a rule can be identified with a single-valued SCC. In this paper we will mainly consider social choice rules derived from social choice correspondences by tie-breaking according to a fixed linear order on X – we will be precise about this whenever this is needed.

For a preference profile $\mathbf{P}$, a preference $P \in L(X)$, and a subset $K \subseteq N$ such that $P_i = P \in L(X)$ for all $i \in K$, we also write $(P_K, \mathbf{P}_{-K})$ instead of $\mathbf{P}$. If, in particular, $K = \{k \in N : P_k = P\}$, then we call K the *group* of (any) voter $i \in K$ at $\mathbf{P}$. Thus, a group collects all voters with the same preference, for some preference in a preference profile.

The following definition captures the situation where a(ny) voter in a group prefers the alternative which results if all voters in that group vote insincerely using the same preference.

Definition 2.1 Voter $i \in N$ *has an incentive to manipulate* rule F at profile $\mathbf{P} \in L(X)^N$ if there is a $\tilde{P} \in L(X)$ such that $F(\tilde{P}_K, \mathbf{P}_{-K})\ P\ F(\mathbf{P})$, where K is the group of i at $\mathbf{P}$ (i.e., all voters who have common preference $P = P_i$ at $\mathbf{P}$).

Clearly, this definition implies that if voter i has an incentive to manipulate, then all members of i's group have an incentive to manipulate – using the same preference $\tilde{P}$. Therefore, we also say that group K *has an incentive to manipulate*.

We introduce some further terminology. A preference profile $\mathbf{P} \in L(X)^N$ is *manipulable under* rule F if there is a voter who has an incentive to manipulate at $\mathbf{P}$. A rule F is *manipulable* if there is a manipulable preference profile under F.

2.2 Safe and unsafe manipulations

Let F be a rule, and let $\mathbf{P} \in L(X)^N$ be a preference profile. Suppose that voter i belonging to group K has an incentive to manipulate F at $\mathbf{P}$ by preference $\tilde{P}$. We say that *manipulation with* $\tilde{P}$ *is unsafe for* i *at* $\mathbf{P}$ if there exists $M \subsetneq K$ such that $i \in M$ and $F(\mathbf{P})\, P_i\, F(\tilde{P}_M, \mathbf{P}_{-M})$. If such an M does not exist, then *manipulation with* $\tilde{P}$ *is safe for* i *at* $\mathbf{P}$. In words, a manipulation is safe if it never results in a worse

[1] The cases $n < 3$ or $m < 3$ are uninteresting for the purpose of this paper, as can easily be verified in the sequel.

[2] That is, an irreflexive, asymmetric, transitive and complete binary relation.

alternative if not all members of the group join in the manipulation. Clearly, if $K = \{i\}$ then every manipulation is safe.

A preference profile $\mathbf{P} \in L(X)^N$ is *safely manipulable* (given F) if there is a voter for whom manipulation is safe with $\tilde{P}$ for some $\tilde{P} \in L(X)$. It is *unsafely manipulable* if there is a voter for whom manipulation with $\tilde{P}$ is unsafe for some $\tilde{P} \in L(X)$. A preference profile can be both safely and unsafely manipulable, even by the same voter.

The rule F is *safely manipulable* if there is a safely manipulable preference profile, and *unsafely manipulable* (UM) if there is an unsafely manipulable preference profile. Again, F can be both safely and unsafely manipulable. Rule F is *only safely manipulable* (OSM) if for every manipulable profile $\mathbf{P} \in L(X)^N$, $\mathbf{P}$ is not unsafely manipulable. Hence, F is OSM if it is not UM.

2.3 Social choice correspondences

In this subsection we introduce the social choice correspondences from which the rules to be studied in this paper, will be derived by tie-breaking.

2.3.1 Scoring correspondences

A *scoring vector* is a vector $s = (s_1, \dots, s_m) \in \mathbb{R}^m$ such that $s_1 \geq \cdots \geq s_m \geq 0$ and $s_1 > s_m$. For a preference profile $\mathbf{P}$ and an alternative a, let $v_j(a, \mathbf{P})$ denote the number of voters having a at the j-th position (where voter i has a at the j-th position if $|\{b \in X : bP_i a\}| = j - 1$). Then $S(a, \mathbf{P}) = \sum_{j=1}^{m} s_j v_j(a, \mathbf{P})$ is the total score of a at $\mathbf{P}$. The *scoring correspondence* F with scoring vector s assigns to each preference profile $\mathbf{P}$ the set $\{a \in X : S(a, \mathbf{P}) \geq S(a', \mathbf{P}) \text{ for all } a' \in X\}$, i.e., the set of alternatives with maximal total score. Well-known examples are:

- *q-approval*: $s_1 = \cdots = s_q = 1, s_{q+1} = \cdots = s_m = 0$, where $q \in \{1, \dots, m-1\}$; for $q = 1$ this is also called *plurality*, and for $q = m - 1$ this is also called *veto* or *antiplurality*,
- *Borda*: $s = (m-1, m-2, \dots, 1, 0)$.

2.3.2 Run-off

For a preference profile $\mathbf{P}$, two alternatives with maximal plurality scores (see Sect. 2.3.1) are chosen, if necessary using a tie-breaking rule. Among these two, say a and b, we choose the alternative(s) which win in a pairwise contest, that is, a is chosen if $|\{i \in N : aP_i b\}| \geq |\{i \in N : bP_i a\}|$ and b is chosen if $|\{i \in N : bP_i a\}| \geq |\{i \in N : aP_i b\}|$.

2.3.3 Copeland

For a preference profile $\mathbf{P}$, the *Copeland score* of an alternative a is the number

$$|\left\{b \in X : |\{i \in N : aP_i b\}| > \frac{n}{2}\right\}| - |\left\{b \in X : |\{i \in N : bP_i a\}| > \frac{n}{2}\right\}|.$$

Hence, the Copeland score of an alternative a is the number of alternatives beaten by a minus the number of alternatives that beat a, where x beats y if a strict majority of the voters prefers x over y. The Copeland correspondence chooses the alternatives with maximal Copeland score.

2.3.4 Single-transferable-vote, STV

For a preference profile $\mathbf{P}$, for each alternative a determine the number of voters who have a at top position, i.e., determine its plurality score $S(a, \mathbf{P})$ for scoring vector $(1, 0, \ldots, 0)$. If all nonzero plurality scores are equal, then STV assigns the set of all alternatives that occur at top, i.e., that have nonzero plurality score. If not all these nonzero plurality scores are equal then: if there is an alternative a with plurality score strictly higher than $n/2$, then STV assigns $\{a\}$; otherwise, leave out those alternatives that have minimal (possibly zero) plurality score. This results in a restricted preference profile with fewer alternatives. Now repeat this procedure until no more alternatives can be left out: STV assigns the remaining alternatives to $\mathbf{P}$. As an illustration, consider the following two profiles:

$$\begin{array}{ccccc} a & b & b & c & c \\ b & \cdot & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array} \quad \text{and} \quad \begin{array}{cccccc} a & b & b & c & c & c \\ b & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

In the left profile, after eliminating alternatives with zero plurality score (if any), a is eliminated: this results in a profile where b has plurality score 3, so that STV assigns $\{b\}$. In the right profile, again after eliminating alternatives with zero plurality score (if any), also a is eliminated, so that STV assigns $\{b, c\}$.

3 (Un)safe manipulability of scoring rules

In this section we investigate the (un)safe manipulability of rules derived from scoring correspondences, so-called *scoring rules*. Our first main result concerns these rules in general, and our second result focuses on Borda (cf. Section 2.3.1).

As already mentioned, we need to apply tie-breaking in order to derive rules from social choice correspondences. We do this by fixing a linear order on the set of alternatives X and taking the maximal element according to this order from a set assigned by the correspondence. In what follows we will be more precise whenever this is needed.

For a scoring vector s, a *jump* is a non-zero difference between two adjacent scoring values. If s has r jumps, then this means that there are distinct $k_1, \ldots, k_r \in \{1, \ldots, m-1\}$ such that $s_{k_1} > s_{k_1+1}, \ldots, s_{k_r} > s_{k_r+1}$, while all other differences are zero. We use the notation $\Delta_j = s_{k_j} - s_{k_j+1}$ for $j = 1, \ldots, r$ to denote the nonzero differences between scoring values.

Our first result concerns unsafe and only safe manipulability of scoring rules in general.

Theorem 3.1 *Let s be a scoring vector with r jumps, and let F be a scoring rule derived from the scoring correspondence associated with scoring vector s by tie-breaking (Table 1). If $r = 1$, then F is only safely manipulable. If $r \geq 2$, then the results are as in the following table:*

Table 1 The results of Theorem 3.1

	2 jumps			3 or more jumps	
	$\Delta_1 > \Delta_2$	$\Delta_1 \leq \Delta_2$			
		$k_1 = 2$, $k_2 = 4$	Otherwise	$\Delta_1 > \Delta_3$ or $\Delta_2 > \Delta_3$	Otherwise
$m = 3$	$\forall n$: OSM	$\forall n$: OSM		(not applicable)	
$m = 4$	$\exists n$: UM	$\forall n$: OSM		$\exists n$: UM	$\forall n$: OSM
$m = 5$	$\exists n$: UM	$\forall n$: OSM	$\exists n$: UM	$\exists n$: UM	
$m \geq 6$	$\exists n$: UM	$\exists n$: UM		$\exists n$: UM	

Proof (i) In this first part of the proof, we assume that there is an unsafe manipulation and derive conditions implied by this assumption. Let $\mathbf{P}$ be a preference profile and let $a, b, c \in X$ such that for voter i in group K (and, consequently, for all voters in K) we have aP_ibP_ic. Suppose that group K has an incentive to manipulate and manipulation is unsafe with $F(\mathbf{P}) = b$, $F(\tilde{P}_K, \mathbf{P}_{-K}) = a$, and $F(\tilde{P}_M, \mathbf{P}_{-M}) = c$ for some $\tilde{P} \in L(X)$ and $M \subsetneq K$. In words, b is the alternative chosen at $\mathbf{P}$, group K can achieve a by manipulating via $\tilde{P}$, but if only the voters in M deviate, the worse alternative c results.

For every alternative $x \in X$, let ε_x denote the change in score when a voter $i \in K$ changes from P_i to $\tilde{P}$, hence $\varepsilon_x|G| = S(x, (\tilde{P}_G, \mathbf{P}_{-G})) - S(x, \mathbf{P})$ for every $G \subseteq K$. Since $F(\mathbf{P}) = b$ we have

$$S(b, \mathbf{P}) \geq S(c, \mathbf{P}), \tag{1}$$

and since $F(\tilde{P}_M, \mathbf{P}_{-M}) = c$ we have $S(c, (\tilde{P}_M, \mathbf{P}_{-M})) \geq S(b, (\tilde{P}_M, \mathbf{P}_{-M}))$, hence

$$S(c, \mathbf{P}) + \varepsilon_c|M| \geq S(b, \mathbf{P}) + \varepsilon_b|M|. \tag{2}$$

If $\varepsilon_b \geq \varepsilon_c$, then by (1) and (2), $S(c, \mathbf{P}) = S(b, \mathbf{P})$ and $\varepsilon_b = \varepsilon_c$, which by tie-breaking implies $b = F(\tilde{P}_M, \mathbf{P}_{-M})$, a contradiction. Therefore, $\varepsilon_b < \varepsilon_c$. Similarly, $\varepsilon_c < \varepsilon_a$. Consequently, $\varepsilon_b < \varepsilon_c < \varepsilon_a$. The five possible (sign) cases are given in the following table:

Table 2 Part (i) of the proof of Theorem 3.1

Case	ε_b	ε_c	ε_a
1	$-$	$-$	0
2	$-$	$-, 0, +$	$+$
3	0	$+$	$+$
4	$-$	$-$	$-$
5	$+$	$+$	$+$

In the remainder of the proof, based on Table 2, we derive necessary conditions for an unsafe manipulation as in Part (i) to exist. In the last part, we show that when these conditions are not fulfilled, there can be an unsafe manipulation.

(ii) Suppose that $r = 1$, $\Delta_1 = s_k - s_{k+1}$. Then, for all $x \in X$, $\varepsilon_x = 0$ or $\varepsilon_x = -\Delta_1$ or $\varepsilon_x = \Delta_1$. This and $\varepsilon_b < \varepsilon_c < \varepsilon_a$ imply that in Table 2 the only possible case is Case 2. Hence, $\varepsilon_a = \Delta_1$ and $\varepsilon_b = -\Delta_1$ but this is not possible: indeed, if the score of a increases by Δ_1, then a moves from the bottom $m - k$ alternatives in P_i to the top k alternatives in $\tilde{P}$, but aP_ib, so, b is also among the bottom $m - k$ alternatives in P_i and therefore cannot decrease in score when going to $\tilde{P}$. Thus, in case of precisely one jump a scoring rule is only safely manipulable, and the first claim in the theorem is proved.

(iii) Suppose that $r = 2$, $\Delta_1 = s_{k_1} - s_{k_1+1}$, $\Delta_2 = s_{k_2} - s_{k_2+1}$. We go through all cases in Table 2 and consider all possible combinations of jumps for each ε_x, $x \in \{a, b, c\}$ in each case. Of course, throughout we use that initially b is chosen, then c, and at the end a, but we do not always spell out the details.

First, note that Cases 4 and 5 in Table 2 are not possible since these cases require at least three jumps to occur.

In Case 1, $\varepsilon_b < \varepsilon_c < 0 = \varepsilon_a$, there are two possibilities:

1.1 From P_i to $\tilde{P}$, b goes down one jump and c goes down one jump: this is only possible if $\Delta_1 > \Delta_2$, and then $\varepsilon_b = -\Delta_1, \varepsilon_c = -\Delta_2$, and $\varepsilon_a = 0$. This can be summarized as follows: $ab|c|\circ \to \circ, a|b|c$. [Here, | denotes a jump, $ab|c|\circ$ contains the relevant information about P_i, and $\circ, a|b|c$ contains the relevant information about $\tilde{P}$. The small circles $\circ$ indicate other alternatives that are minimally available.]

1.2 b goes down two jumps and c goes down one jump. Then either $abc|\circ|\circ \to \circ, \circ, a|c|b$, hence $\varepsilon_b = -\Delta_1 - \Delta_2$, $\varepsilon_c = -\Delta_1$, and $\varepsilon_a = 0$; or $ab|c|\circ\circ \to \circ, a|\circ|b, c$, hence $\varepsilon_b = -\Delta_1 - \Delta_2, \varepsilon_c = -\Delta_2$, and $\varepsilon_a = 0$.

In Case 2, $\varepsilon_b < 0$ and $\varepsilon_a > 0$. Then $\varepsilon_b = -\Delta_2$ and $\varepsilon_a = \Delta_1$, For ε_c there are two possibilities:

2.1 $\varepsilon_c = \Delta_2$ and $\circ|ab|c \to a|\circ, c|b$. This is only possible if $\Delta_1 > \Delta_2$.

2.2 $\varepsilon_c = 0$, and $\circ|abc|\circ \to a|\circ\circ, c|b$ or $\circ|ab|\circ, c \to a|\circ\circ|b, c$.

Finally, in Case 3, $\varepsilon_b = 0$ and $\varepsilon_a, \varepsilon_c > 0$. There are again two possibilities:

3.1 $\varepsilon_a = \Delta_1, \varepsilon_c = \Delta_2$, and $\circ|ab|c \to a|b, c|\circ$ or $\circ|a|bc \to a|c|\circ, b$. This is only possible if $\Delta_1 > \Delta_2$.

3.2 $\varepsilon_a = \Delta_1 + \Delta_2, \varepsilon_c = \Delta_2$, and $\circ|\circ|abc \to a|c|\circ\circ, b$.

Based on these six possibilities, we can now examine the $r = 2$ cases in Table 1.

- If $m = 3$, then $P_i = a|b|c$, and therefore none of the Cases 1.1-–3.2 applies. Hence, any manipulation in this case is safe.

- If $m = 4$ and $\Delta_1 > \Delta_2$, then Cases 1.1, 2.1, and 3.1 apply, and so there can be unsafe manipulations.
- If $m = 4$ and $\Delta_1 \leq \Delta_2$, then none of the Cases 1.1-–3.2 applies. Hence, any manipulation in this case is safe.
- If $m = 5$ and $\Delta_1 \leq \Delta_2$, then from Cases 1.2, 2.2, and 3.2, it follows that unsafe manipulation may be possible for the following five jump combinations: $k_1 = 1, k_2 = 2$ (3.2); $k_1 = 1, k_2 = 3$ (2.2); $k_1 = 1, k_2 = 4$ (2.2); $k_1 = 2, k_2 = 3$ (1.2); $k_1 = 3, k_2 = 4$ (1.2). In the remaining case, $k_1 = 2, k_2 = 4$, no unsafe manipulation is possible.
- If $m = 5$ and $\Delta_1 > \Delta_2$, then all Cases 1.1-–3.2 may apply and therefore all six jump combinations are possible, so that unsafe manipulation is possible for any of these combinations.
- If, finally, $m \geq 6$, then it is sufficient to consider the Cases 1.2, 2.2, and 3.2, to conclude that for any jump combination unsafe manipulation is possible.

(iv) We next consider the case $r = 3$. Then there must be at least four alternatives.

- If $m \geq 5$, then for each combination of three (or more) jumps it is possible to manipulate unsafely by using only two jumps as in Cases 1.2 and 2.2 Thus, unsafe manipulation may be possible for any jump combination.
- Now let $m = 4$. We consider the five cases in the Table 2.
  - Case 1 implies $P_i = a|b|c|\circ$ and $\tilde{P} = a|\circ|b|c$. Since $\varepsilon_b < \varepsilon_c < \varepsilon_a$, this implies $\Delta_2 > \Delta_3$. In this case, unsafe manipulation may be possible.
  - Case 2 implies $P_i = \circ|a|b|c$ and $\tilde{P} = a|\circ|c|b$ or $\tilde{P} = a|c|\circ|b$. Since $\varepsilon_b < \varepsilon_c < \varepsilon_a$, this implies $\Delta_1 > \Delta 3$ or $\Delta_1 > \Delta_2 + \Delta_3$. In turn, this implies that unsafe manipulation may be possible if $\Delta_1 > \Delta_3$.
  - Case 3 implies $P_i = \circ|a|b|c$ and $\tilde{P} = a|c|b|\circ$. Since $\varepsilon_b < \varepsilon_c < \varepsilon_a$, this implies $\Delta_1 > \Delta_2 + \Delta_3$. Under this condition, unsafe manipulation may be possible in this case.
  - Case 4 implies $P_i = a|b|c|\circ$ and $\tilde{P} = \circ|a|b|c$. Since $\varepsilon_b < \varepsilon_c < \varepsilon_a$, this implies $\Delta_2 > \Delta_3 > \Delta_1$. In this case therefore, an unsafe manipulation may exist, but by Case 1, $\Delta_2 > \Delta_3$ is already sufficient for this.
  - Case 5 implies $P_i = \circ|a|b|c$ and $\tilde{P} = a|b|c\circ$. Since $\varepsilon_b < \varepsilon_c < \varepsilon_a$, this implies $\Delta_2 < \Delta_3 < \Delta_1$. In this case therefore, an unsafe manipulation may exist, but by Case 2, $\Delta_1 > \Delta_3$ is already sufficient for this.

  Summarizing, an unsafe manipulation may exist if and only if $\Delta_1 > \Delta_3$ or $\Delta_2 > \Delta_3$.

(v) The OSM cases in Table 1 have now been proved. We complete the proof of the theorem by providing a procedure to construct a preference profile for any kind of unsafe manipulation.

Assume that we have a particular number of alternatives m, a given scoring vector s, and a group K of voters with preferences P s.t. $aPbPc$. Take any way of unsafe manipulation, $\tilde{P}$, corresponding to the given m and s from the previous part of the proof. Then, the chosen way of unsafe manipulation defines score differences for alternatives a, b, and c when one voter manipulates (switches from P to $\tilde{P}$). These score differences are: $\varepsilon_a = \sum_{j=1}^{r} \alpha_j \Delta_j$, $\varepsilon_b = \sum_{j=1}^{r} \beta_j \Delta_j$, and $\varepsilon_c = \sum_{j=1}^{r} \gamma_j \Delta_j$, where the α_j, β_j, γ_j are elements of $\{-1, 0, 1\}$. We need to prove that there exists a preference profile for some n such that members of K have an incentive to manipulate with $\tilde{P}$ and this manipulation is unsafe.

First, without loss of generality we assume tie-breaking according to aP^tc, bP^tc. Let the scores of alternatives be such that $S(b, \mathbf{P}) = S(c, \mathbf{P})$ and $S(c, (\tilde{P}_K, \mathbf{P}_{-K})) = S(a, (\tilde{P}_K, \mathbf{P}_{-K}))$. This, together with $\varepsilon_b < \varepsilon_c < \varepsilon_a$, implies that $F(\mathbf{P}) = b$, $F(\tilde{P}_K, \mathbf{P}_{-K}) = a$, and $F(\tilde{P}_M, \mathbf{P}_{-M}) = c$ for some $M \subseteq K$.

For the difference in scores for alternatives a and c before and after manipulation, we have $S(a, (\tilde{P}_K, \mathbf{P}_{-K})) - S(a, \mathbf{P}) = \varepsilon_a |K|$ and $S(c, (\tilde{P}_K, \mathbf{P}_{-K})) - S(c, \mathbf{P}) = \varepsilon_c |K|$. Since $S(c, (\tilde{P}_K, \mathbf{P}_{-K})) = S(a, (\tilde{P}_K, \mathbf{P}_{-K}))$ we have $S(a, \mathbf{P}) + \varepsilon_a |K| = S(c, \mathbf{P}) + \varepsilon_c |K|$ and, finally, $S(c, \mathbf{P}) - S(a, \mathbf{P}) = \varepsilon_a |K| - \varepsilon_c |K|$. So, $S(c, \mathbf{P}) - S(a, \mathbf{P}) = \sum_{j=1}^{r} \mu_j \Delta_j$ for some integers μ_j.

Summing up, in the required profile $\mathbf{P}$ it is needed that: (a) the score differences between a, b, and c are fixed, $S(b, \mathbf{P}) - S(c, \mathbf{P}) = 0$, $S(c, \mathbf{P}) - S(a, \mathbf{P}) = \sum_{j=1}^{r} \mu_j \Delta_j$ for some integers μ_j; (b) other alternatives do not affect the result; (c) there are exactly $|K|$ voters with preferences P_i.

We now describe a procedure to generate a profile $\mathbf{P}$ with these properties. We first fix a preference profile for some set of voters K, where every member of K has the same preference P, say, $aPbPcPa_1Pa_2P...Pa_{m-3}$. Take any number of voters in K and include their preferences, P_K, in the profile $\mathbf{P}$ that we are constructing. Then we have condition (c) satisfied.

For the voters outside K we consider the following basic profile $B(a)$:

$$\begin{array}{cccc}
a_{m-3} & a_{m-4} & \cdots & a \\
a & a_{m-3} & \cdots & c \\
c & a & \cdots & b \\
b & c & \cdots & a_1 \\
a_1 & b & \cdots & a_2 \\
\vdots & \vdots & \vdots & \vdots \\
a_{m-4} & a_{m-5} & \cdots & a_{m-3}
\end{array}$$

Observe that in $B(a)$ the scores of all alternatives are equal, and that P does not occur. Suppose that we need to increase the score of alternative a by the amount Δ_l, which is the size of the l-th jump, following position k_l. Then we replace column (preference) k_l in $B(a)$ by $\tilde{R}$, where $\tilde{R}$ is obtained by switching positions k_l and $k_l + 1$ in column k_l. This results in a profile $B'(a)$ where the scores of all alternatives except a and a_{m-3} are still equal (and equal to the scores in $B(a)$), the score of a has increased by Δ_l, and the score of a_{m-3} has decreased by Δ_l. Note that $\tilde{P} \neq P$ and, thus, P does not occur in $B'(a)$. So, we include $B'(a)$ in $\mathbf{P}$. If it is needed to increase the score of a by the size of another jump, we include $B''(a)$ constructed similarly, etc.

Similar constructions can be made for b and c, if we need to increase their scores, by starting from the most left columns $(a_{m-3}, b, a, c, a_1, \ldots, a_{m-4})$ and $(a_{m-3}, c, a, b, a_1, \ldots, a_{m-4})$ respectively. Doing this as many times as needed to satisfy conditions (a) and (b), in the end we obtain the required preference profile.

Moreover, notice that we can choose any number of voters in K. So, if an unsafe manipulation exists for some m then it is always possible to find a profile with a group of only two voters having an incentive to manipulate unsafely. □

Observe that, although Theorem 3.1 identifies all scoring rules where an unsafe manipulation exists, it is silent about how many voters are needed to have such a manipulation. It is difficult to derive general results about this for (all) scoring rules and therefore, in the next theorem, we focus on the arguably most famous rules with at least two jumps, namely Borda rules (cf. Sect. 2.3.1). Note that, since all jumps at a Borda rule have equal size, the cases with less than 5 alternatives are covered by Theorem 3.1.

Theorem 3.2 *Let F be a Borda rule. If $m = 5$, then an unsafely manipulable profile exists if and only if $n \geq 4$. If $m \geq 6$, then an unsafely manipulable profile exists if and only if $n \geq 3$.*

Proof

(a) First let $m = 5$, $X = \{a, b, c, d, e\}$, and consider the following profiles $\mathbf{P}'$ and $\mathbf{P}''$ for $n = 4$ and $n = 5$, respectively: . Suppose that tie-breaking is done according to the ordering $T = (e, c, a, b, d)$. Then $F(\mathbf{P}') = c$. If group $K = \{1, 2\}$ changes their preferences to $\tilde{P} = (b, a, d, c, e)$, then $F(\tilde{P}_K, \mathbf{P}'_{-K}) = b$, which is preferred by the members of K to c. However, $F(\tilde{P}_{\{1\}}, \mathbf{P}'_{-\{1\}}) = e$, so that this manipulation is unsafe. As to $\mathbf{P}''$, note that also $F(\mathbf{P}'') = c$ and $K = \{1, 2\}$ can manipulate again by $\tilde{P} = (b, a, d, c, e)$. If only voter 1 manipulates, then again e results, so that also this manipulation is unsafe. Thus, if $m = 5$ and $n = 4$ or $n = 5$, there exists an unsafe manipulation.

P'_1	P'_2	P'_3	P'_4
a	a	e	e
b	b	c	d
c	c	d	c
d	d	b	a
e	e	a	b

P''_1	P''_2	P''_3	P''_4	P''_5
a	a	e	e	d
b	b	c	c	e
c	c	b	d	c
d	d	a	b	b
e	e	d	a	a

(b) We next show that no unsafe manipulation exists if $m = 5$ and $n = 3$. In this case, a possibly unsafely manipulating group can only consist of 2 members, say $K = \{1, 2\}$. Suppose, indeed, that for some $a, b, c \in X$, $aP_i bP_i c$ for all $i \in K$, and that there is a preference P_3 for voter 3 and a preference $\tilde{P}$ such that $F(\mathbf{P}) = b$, $F(\tilde{P}_{\{1,2\}}, P_3) = a$, and $F(\tilde{P}_{\{1\}}, \mathbf{P}_{\{2,3\}}) = c$. Note that, at $\mathbf{P}$, the Borda score of c must be strictly larger than the Borda score of a: if not, then the score of c should increase more than the score of a after manipulation by just one member of K, but then c would still win after manipulation by both members of K, a contradiction. Further, the score of a contributed by P_1 and P_2 is at least 4 more than the the score of c contributed by P_1 and P_2, since $aP_i bP_i c$ for $i = 1, 2$. In turn, these facts imply that the score of c contributed by P_3 is at least five more than the score of a contributed by P_3, which is impossible with five alternatives.

(c) Consider the case $m = 6$, $X = \{a, b, c, d, e, f\}$, and $n = 3$, and the profile $\mathbf{P}$ with $P_1 = P_2 = (a, b, c, d, e, f)$ and $P_3 = (c, b, f, d, e, a)$. Let $\tilde{P} = (a, e, d, c, b, f)$. Then $F(\mathbf{P}) = b$, $F(\tilde{P}_{\{1,2\}}, P_3) = a$, and (assuming that c beats a by tie-breaking) $F(\tilde{P}_{\{1\}}, \mathbf{P}_{\{2,3\}}) = c$, so that an unsafe manipulation exists in this case.

(d) Finally, the hitherto constructed profiles where an unsafe manipulation exists, can be extended with any number of alternatives, simply by adding those alternatives at the bottom of the preferences. Also, each of the manipulable profiles can be extended by any even number of agents 2ℓ: add ℓ times the pair of preferences $(a_1, \ldots, a_m)$ and $(a_m, \ldots, a_1)$, where $X = \{a_1, \ldots, a_m\}$, and note that this just adds equal scores for all alternatives. The proof of the theorem is now complete. □

4 (Un)safe manipulability of run-off

For the definition of the run-off correspondence, see Sect. 2.3.2.

We start by observing that at a run-off rule it is impossible to manipulate in favor of the most preferable alternative.

Lemma 4.1 *Let F be a run-off rule, let $\mathbf{P}$ be a preference profile, let $i \in N$ and $a \in X$ such that aP_ix for all $x \in X \setminus \{a\}$ and $a \neq F(\mathbf{P})$, and let K be the group of i. Then there is no $\tilde{P} \in L(X)$ such that $a = F(\tilde{P}_K, \mathbf{P}_{-K})$.*

Proof If a does not survive the first stage of the run-off procedure at $\mathbf{P}$, then it will also not survive the first stage at any $(\tilde{P}_K, \mathbf{P}_{-K})$. If a survives the first stage but not the second stage of the run-off procedure at $\mathbf{P}$, then for any $\tilde{P} \in L(X)$, either a does not survive the first stage at $(\tilde{P}_K, \mathbf{P}_{-K})$, or it does. In the latter case, since for every $x \in X$ we have $|\{j \in K : aP_jx\}| \geq |\{j \in K : a\tilde{P}_jx\}|$, it follows that a does not survive the second stage at $(\tilde{P}_K, \mathbf{P}_{-K})$. □

Our results for run-off rules are as follows.

Theorem 4.2 *Let F be a run-off rule.*

(a) *If $m = 3$, then F is only safely manipulable.*
(b) *If $m = 4$, then F is only safely manipulable if and only if $n \leq 5$.*
(c) *If $m \geq 5$, then F is only safely manipulable if and only if $n \leq 4$.*

Proof We will prove the theorem for seven specific cases, depending on the numbers m and n of alternatives and voters, and then summarize how the theorem follows from these cases.

(1) Let $m = 3$, $X = \{a, b, c\}$, $\mathbf{P} \in L(X)^N$, and let K be a group with common preference aP_ibP_ic for every $i \in K$. If K can manipulate unsafely by $\tilde{P}$, then we must have $F(\mathbf{P}) = b$, $F(\tilde{P}_K, \mathbf{P}_{-K}) = a$, and $F(\tilde{P}_M, \mathbf{P}_{-M}) = c$ for some $M \subseteq K$. Such a

manipulation, however, is excluded by Lemma 4.1. This proves Part (a) of the theorem.

(2) If $n = 3$ then for an unsafe manipulation a group of at least two members is required, but then their common top alternative is chosen by F. So F is only safely manipulable. From now on, we assume that $m, n \geq 4$, $a, b, c, d \in X$, and the members of group K have a preference P with top alternative a and with bPc, cPd.

(3) Let $n = 4$. Assume, contrary to what we want to prove, that K has an unsafe manipulation at $\mathbf{P}$ via $\tilde{P}$. Then by Lemma 4.1 we may assume that $F(\mathbf{P}) = c$, $F(\tilde{P}_K, \mathbf{P}_{-K}) = b$, and $F(\tilde{P}_M, \mathbf{P}_{-M}) = d$ for some $M \subseteq K$. Since $|K| = 2$, the plurality score of a at $\mathbf{P}$ is 2, and the plurality score of c at $\mathbf{P}$ is 1 or 2. In the latter case, the plurality score of d at $\mathbf{P}$ is 0, but then $F(\tilde{P}_M, \mathbf{P}_{-M}) \neq d$ since $|M| = 1$, a is the top alternative of P, and b is the top alternative of $\tilde{P}$, and so d does not survive the first stage at $(\tilde{P}_M, \mathbf{P}_{-M})$, contradicting that $F(\tilde{P}_M, \mathbf{P}_{-M}) = d$. Therefore, we have that the plurality score of c at $\mathbf{P}$ is 1. If the plurality score of d at $\mathbf{P}$ is 0, then as before, $F(\tilde{P}_M, \mathbf{P}_{-M}) \neq d$, a contradiction. Thus, the plurality score of d at $\mathbf{P}$ is 1. Since $F(\mathbf{P}) = c$, a and c survive the first stage at $\mathbf{P}$, which implies that the tie-breaking order P^t satisfies cP^td and cP^ta. Since the top alternative of $\tilde{P}$ is b, at $(\tilde{P}_M, \mathbf{P}_{-M})$ the alternatives a, b, c, d all have equal plurality score 1, and since cP^td and cP^ta, we have that b and c survive the first round, contradicting again that $F(\tilde{P}_M, \mathbf{P}_{-M}) = d$. Hence, we have proved that for $m \geq 4$ and $n = 4$ there is no unsafe manipulation.

(4) Let $n = 5$ and $m = 4$. As in Part (3), assume that K has an unsafe manipulation at $\mathbf{P}$ via $\tilde{P}$, with $F(\mathbf{P}) = c$, $F(\tilde{P}_K, \mathbf{P}_{-K}) = b$, and $F(\tilde{P}_M, \mathbf{P}_{-M}) = d$ for some $M \subseteq K$. Clearly, the plurality score of a at $\mathbf{P}$ cannot be larger than 2, and therefore is equal to 2. In particular, $|K| = 2$, say $K = \{1, 2\}$. Since $F(\mathbf{P}) = c$ and $F(\tilde{P}_K, \mathbf{P}_{-K}) = b$, the plurality score of c at $\mathbf{P}$ is 1 or 2. In the latter case, say that P_3 and P_4 have top alternative c. Since $F(\tilde{P}_{\{1\}}, \mathbf{P}_{-\{1\}}) = d$ and the top alternative of $\tilde{P}$ is b, we have that a, b, and d each have plurality score 1 at $(\tilde{P}_{\{1\}}, \mathbf{P}_{-\{1\}})$, and c and d survive the first stage. However, cP_jd for $j = 2, 3, 4$, so that c finally wins, a contradiction. Hence, the plurality score of c at $\mathbf{P}$ is 1. Since $F(\tilde{P}_{\{1\}}, \mathbf{P}_{-\{1\}}) = d$, the plurality score of d at $\mathbf{P}$ is at least 1, and since $F(\mathbf{P}) = c$, it is exactly 1. It follows that the plurality score of b at $\mathbf{P}$ is also 1. In turn, for the tie-breaking order P^t, this implies that cP^td. But then, at $(\tilde{P}_{\{1\}}, \mathbf{P}_{-\{1\}})$, d does not survive the first round, contradicting that $F(\tilde{P}_{\{1\}}, \mathbf{P}_{-\{1\}}) = d$. Hence, we have proved that for $m = 4$ and $n = 5$ there is no unsafe manipulation.

(5) Let $n = 5$ and $m = 5$, $X = \{a, b, c, d, e\}$. We show that there is an unsafe manipulation. Let $K = \{1, 2\}$, and let $\mathbf{P}$ be a preference profile with $P_1 = P_2 = (a, b, c, d, e)$, and such that c, d, and e each have plurality score of 1 at $\mathbf{P}$, cP_ja for $j = 3, 4, 5$, and bP_5dP_5c. Let the tie-breaking order be $P^t = (c, d, e, a, b)$. Then $F(\mathbf{P}) = c$. If $\tilde{P}$ has top alternative b and $d\tilde{P}c$, then $F(\tilde{P}_{\{1,2\}}, \mathbf{P}_{-\{1,2\}}) = b$, and $F(\tilde{P}_{\{1\}}, \mathbf{P}_{-\{1\}}) = d$. So K has an unsafe manipulation.

(6) For $n \geq 6$ and $m = 4$ we construct unsafely manipulable profiles based on the following preferences: $P^1 = (a, b, c, d)$, $P^2 = (c, a, b, d)$, $P^3 = (d, b, c, a)$, $P^4 = (b, a, d, c)$. Let $\mathbf{P}_n$ denote a preference profile with n voters. For

Table 3 Proof summary for Theorem 4.2

	$n = 3$	$n = 4$	$n = 5$	$n \geq 6$
$m = 3$	OSM Part 1	OSM Part 1	OSM Part 1	OSM Part 1
$m = 4$	OSM Part 2	OSM Part 3	OSM Part 4	UM Part 6
$m \geq 5$	OSM Part 2	OSM Part 3	UM Parts 5, 7	UM Parts 6, 7

$j = 0, 1, 2, \ldots$ let $\mathbf{P}_{6+3j}$ such that it contains P^1, P^2, and P^3 each $2+j$ times: $\mathbf{P}_{6+3j} = ((2+j)P^1, (2+j)P^2, (2+j)P^3)$. Assume that the tie-breaking order is $P^t = (b, d, c, a)$. Then $F(\mathbf{P}_{6+3j}) = c$. If the voters with preference P^1 change to P^4, then b wins. If only one voter manipulates, then d wins. Similarly, we consider preference profiles $\mathbf{P}_{7+3j} = ((2+j)P^1, (2+j)P^2, (3+j)P^3)$ for $j = 0, 1, 2, \ldots$; with the same tie-breaking rule, the same kind of unsafe manipulation exists. Finally, we consider profiles $\mathbf{P}_{8+3j} = ((2+j)P^1, (2+j)P^2, (4+j)P^3)$ for $j = 0, 1, 2, \ldots$, and tie-breaking order $P^t = (b, c, d, a)$: again the same kind of unsafe manipulation exists.

(7) Finally, we observe that for any unsafely manipulable profile we obtain an unsafely manipulable profile for more alternatives by simply adding those additional alternatives at the bottom of the preferences in the given profile. Table 3 summarizes how the theorem follows from the seven parts of the proof.

□

5 (Un)safe manipulability of Copeland

Recall (Sect. 2.3.3) that the Copeland correspondence chooses the alternatives with maximal Copeland score, where the Copeland score of an alternative a at a preference profile $\mathbf{P}$ is the number $|\left\{b \in X : |\{i \in N : aP_ib\}| > \frac{n}{2}\right\}| - |\left\{b \in X : |\{i \in N : bP_ia\}| > \frac{n}{2}\right\}|$.

Theorem 5.1 *Let F be a Copeland rule. Then F is only safely manipulable if and only if $m = 3$ or $n = 3$.*

Proof The proof proceeds in five parts.

(1) Suppose $n = 3$. Since $|K| \geq 2$ is required for an unsafe manipulation, but in that case the top alternative of the voters in K is chosen, there exists no unsafe manipulation.

(2) Suppose $m = 3$, $X = \{a, b, c\}$, let $\mathbf{P}$ be a preference profile, and let the voters in a group K have preferences $aPbPc$. For an unsafe manipulation by K we must have $F(\mathbf{P}) = b$, $F(\tilde{P}_K, \mathbf{P}_{-K}) = a$, and $F(\tilde{P}_M, \mathbf{P}_{-M}) = c$ for some $M \subseteq K$ and $\tilde{P} \in L(X)$. We show that this is impossible. For alternative x denote by $S(x)$, $\tilde{S}(x)$, and $\bar{S}(x)$ the Copeland scores of x at $\mathbf{P}$, $(\tilde{P}_K, \mathbf{P}_{-K})$, and $(\tilde{P}_M, \mathbf{P}_{-M})$,

respectively. Clearly we have $S(a) \geq \bar{S}(a) \geq \tilde{S}(a)$ and $S(c) \leq \bar{S}(c) \leq \tilde{S}(c)$. Since $F(\tilde{P}_K, \mathbf{P}_{-K}) = a$ and $F(\tilde{P}_M, \mathbf{P}_{-M}) = c$ we have $\bar{S}(c) \geq \bar{S}(a)$ and $\tilde{S}(a) \geq \tilde{S}(c)$. Hence $\tilde{S}(a) \geq \tilde{S}(c) \geq \bar{S}(c) \geq \bar{S}(a) \geq \tilde{S}(a)$, and therefore $\tilde{S}(a) = \tilde{S}(c) = \bar{S}(c) = \bar{S}(a)$. This, however, is inconsistent with (any order of) tie-breaking. Thus, for $m = 3$ there exists no unsafe manipulation.

(3) We exhibit an unsafely manipulable preference profile for $m = 4$ and $n = 4$. Consider the profile $\mathbf{P}$ given by

P_1	P_2	P_3	P_4
c	c	b	b
a	a	a	a
b	b	c	c
d	d	d	d

The Copeland scores of a, b, c, d at $\mathbf{P}$ are, respectively, $1, 1, 1, -3$. With tie-breaking order $P^t = (a, b, c, d)$ we have $F(\mathbf{P}) = a$. If $K = \{3, 4\}$ changes preferences to $\tilde{P} = (b, d, c, a)$, then the scores are $-1, 1, 1, -1$, so that $F(\tilde{P}_K, \mathbf{P}_{-K}) = b$. If $M = \{3\}$, then the scores are $0, 1, 2, -3$, so that $F(\tilde{P}_M, \mathbf{P}_{-M}) = c$. Hence, $\mathbf{P}$ is an unsafely manipulable preference profile.

(4) We exhibit an unsafely manipulable preference profile for $m = 4$ and $n = 5$. Consider the profile $\mathbf{P}$ given by

P_1	P_2	P_3	P_4	P_5
c	c	a	b	b
b	a	d	a	a
a	b	c	c	c
d	d	b	d	d

In this case, it is easy to verify that, with the same tie-breaking as in Part (3), $K = \{4, 5\}$ can unsafely manipulate by $\tilde{P} = (b, d, c, a)$.

(5) Finally, if there are more than five agents then these agents can be added in pairs with opposite preferences to the unsafely manipulable profiles in Parts (3) and (4), to obtain such profiles with more than five agents. If there are more than four alternatives, then the additional alternatives can be added at the bottom of the unsafely manipulable profiles for four alternatives. This concludes the proof of the theorem.

□

6 (Un)safe manipulability of single-transferable-vote

For the definition of the Single-Transferable-Vote (STV) correspondence, see Sect. 2.3.4.

As for all the previous rules, group manipulation for STV with $m = 3$ is always safe:

Lemma 6.1 *Let F be an STV rule, and let* $m = 3$. *Then F is only safely manipulable.*

Proof Let $X = \{a, b, c\}$, let $\mathbf{P}$ be a preference profile, and suppose there is an unsafe manipulation by group K, who have preference $P = (a, b, c)$, via $\tilde{P} \in L(X)$. Then we have $F(\mathbf{P}) = b$, $F(\tilde{P}_K, \mathbf{P}_{-K}) = a$, and $F(\tilde{P}_M, \mathbf{P}_{-M}) = c$ for some $M \subseteq K$. Then the top alternative of $\tilde{P}$ cannot be a, since this would not change the outcome, nor b, since this would either not change the outcome or lead to the elimination of a. Hence, the top alternative of $\tilde{P}$ is c. Denote by $S(x)$ the plurality score (i.e., number of top positions) of $x \in X$ at $\mathbf{P}$, and by $\tilde{S}(x)$ the plurality score of x at $(\tilde{P}_K, \mathbf{P}_{-K}))$. Then $S(a) > \tilde{S}(a)$, $S(c) < \tilde{S}(c)$, and $S(b) = \tilde{S}(b)$.

We claim that $\tilde{S}(a) \geq \tilde{S}(b)$. Suppose not, i.e., $\tilde{S}(a) < \tilde{S}(b)$. Then, if $\tilde{S}(c) \geq \tilde{S}(a)$, alternative a will be eliminated at $(\tilde{P}_K, \mathbf{P}_{-K}))$, contradicting $F(\tilde{P}_K, \mathbf{P}_{-K}) = a$. If $\tilde{S}(c) < \tilde{S}(a)$, then $S(c) < S(a)$ and $S(c) < S(b)$, and $\tilde{S}(c) < \tilde{S}(a) < \tilde{S}(b)$. This means that c is eliminated first, both at $\mathbf{P}$ and at $(\tilde{P}_K, \mathbf{P}_{-K}))$; since, however, $F(\mathbf{P}) = F(P_K, \mathbf{P}_{-K}) = b$, this implies that also $F(\tilde{P}_K, \mathbf{P}_{-K}) = b$ as $P = (a, b, c)$ and $\tilde{P} = (c, \cdot, \cdot)$. This is a contradiction, and thus the claim is proved.

We next consider $\tilde{S}(c)$. If $\tilde{S}(a) = \tilde{S}(b)$, then $\tilde{S}(c) = \tilde{S}(a) = \tilde{S}(b)$, otherwise $a \neq F(\tilde{P}_K, \mathbf{P}_{-K})$. If $\tilde{S}(a) > \tilde{S}(b)$, then $\tilde{S}(c) \geq \tilde{S}(b)$, because otherwise again $a \neq F(\tilde{P}_K, \mathbf{P}_{-K})$ by a similar argument as in the second case of the preceding paragraph.

Finally, if $\tilde{S}(c) = \tilde{S}(a) = \tilde{S}(b)$, then for all $M \subseteq K$, c is eliminated in $(\tilde{P}_M, \mathbf{P}_{-M})$, a contradiction. If $\tilde{S}(c) > \tilde{S}(b)$, then there is some $M' \subseteq K$, such that $S(c) + |M'| - 1 < S(b)$ and $S(c) + |M'| \geq S(b)$. For all G with $|G| < |M'|$, c is eliminated; and for all G with $|G| \geq |M'|$ the winner is a. This is again a contradiction, which concludes the proof of the lemma. □

Theorem 6.2 *Let F be an STV rule. Then F is only safely manipulable if and only if $m = 3$ or $n \leq 7$.*

Proof The proof proceeds in several parts.

(1) By Lemma 6.1, if $m = 3$ then F is only safely manipulable.

(2) For $n \geq 8$ and $m = 4$, $X = \{a, b, c, d\}$, we construct unsafely manipulable profiles based on the following preferences: $P^1 = (a, b, c, d)$, $P^2 = (a, b, d, c)$, $P^3 = (b, a, c, d)$, $P^4 = (b, a, d, c)$, $P^5 = (b, c, d, a)$, $P^6 = (c, a, b, d)$, $P^7 = (d, b, c, a)$, and $P^8 = (d, c, a, b)$. Let $\mathbf{P}_n$ denote a preference profile of n voters.

  (2.1) For $j = 0, 1, 2, \ldots$ consider a profile $\mathbf{P}_{8+4j} = (2P^1, (1+j)P^2, (1+j)P^5, (2+j)P^6, (2+j)P^8)$, meaning that preference P^1 occurs 2 times etc. The plurality scores (of the first round) are: $S_1(a) = 3 + j$, $S_1(b) = 1 + j$, $S_1(c) = 2 + j$, $S_1(d) = 2 + j$. Since b has minimal score, it is deleted. At the second round: $S_2(a) = 3 + j$, $S_2(c) = 2j + 3$, $S_2(d) = 2 + j$, so that d is deleted. At the third round: $S_3(a) = 3 + j$, $S_3(c) = 3j + 5$, so that c wins. Suppose that the group of voters with preferences P^1 switch to P^3. Then $S_1(a) = 1 + j$, $S_1(b) = 3 + j$, $S_1(c) = 2 + j$, $S_1(d) = 2 + j$, so that a is deleted; $S_2(b) = 2j + 4$, $S_2(c) = 2 + j$, $S_2(d) = 2 + j$, so that c and d are deleted, and thus b wins. If only one voter of the group manipulates, then the scores are: $S_1(a) = 2 + j$, $S_1(b) = 2 + j$, $S_1(c) = 2 + j$, $S_1(d) = 2 + j$, so, there is a complete tie and d wins provided that the tie-breaking order satisfies dP^ta, dP^tb, dP^tc.

(2.2) In a profile $\mathbf{P}_{9+4j} = (2P^1, (1 + j)P^2, (1 + j)P^5, (2 + j)P^6, (3 + j)P^8)$ with $j = 0, 1, 2, \ldots$ the same kind of manipulation by the same group leads to the same results, for any tie-breaking order.

(2.3) In a profile $\mathbf{P}_{14+4j} = (2P^1, (2 + j)P^2, P^4, (1 + j)P^5, (4 + j)P^6, (4 + j)P^8)$ with $j = 0, 1, 2, \ldots$ the result is c. Switching to P^3 for the group of voters having P^1 leads to b. When only one voter manipulates, d wins provided that dP^tc.

(2.4) For a profile $\mathbf{P}_{15+4j} = (2P^1, (2 + j)P^2, P^4, (1 + jP^5, (4 + j)P^6, (5 + j)P^8)$ with $j = 0, 1, 2, \ldots$ the result is c if cP^td and cP^ta. Switching to P^3 for the group of voters having P^1 leads to b and manipulation of only one member leads to d.

(2.5) Only two cases are left: $n = 10$ and $n = 11$. Let $\mathbf{P}_{10} = (2P^1, P^2, P^5, 3P^6, 3P^8)$ and $\mathbf{P}_{11} = (2P^1, P^2, P^4, 4P^6, 3P^7)$. In these profiles, the result is c, but if the voters having preferences P^1 switch to P^4, then the result changes to b, and in case of manipulation of one voter it changes to d.

Summing up, for $m = 4$ and $n \geq 8$ there exist unsafely manipulable profiles. This result also holds for $m > 4$: additional alternatives can be added at the bottom of all preferences and will not change the result.

(3) In this part of the proof we show that for $m \geq 4$ and $n \leq 7$ unsafely manipulable profiles do not exist. Let $\mathbf{P}$ be a preference profile. Take four alternatives $a, b, c, d \in X$, and let there be a group K with preference $aPbPcPd$ restricted to these alternatives, and with a their top alternative. Let $M \subseteq K$ and $\tilde{P} \in L(X)$. We use the notation $S(x)$ for the plurality score of x at $\mathbf{P}$, $\tilde{S}(x)$ for the plurality score of x at $(\tilde{P}_K, \mathbf{P}_{-K})$, and $\bar{S}(x)$ for the plurality score of x at $(\tilde{P}_M, \mathbf{P}_{-M})$. Assume that K has a manipulation via $\tilde{P}$.

(3.1) Since $|K| \geq 2$, $S(a) \geq 2$. Let $c = F(\mathbf{P})$. Then $S(c) \geq 2$. Hence $n > 3$. Consider the case $n = 4$, $S(a) = 2$ and $S(c) = 2$. Members of K cannot manipulate in favor of a (since voting for another alternative will lead to elimination of a), but they can make b winning by voting for b, which is better than c. This manipulation is safe, since $\bar{S}(a) = 1$, $\bar{S}(b) = 1$, and $\bar{S}(c) = 2$ and d cannot win in $(\tilde{P}_M, \mathbf{P}_{-M})$.

(3.2) Consider the case $n = 5$. The first-round scores are $S(a) = 2$, $S(c) = 2$, and there is some $x \in X$, $x \neq a$, $x \neq c$, s.t. $S(x) = 1$. Manipulation in favor of a is also impossible, so members of K vote for b. Again, d cannot be the winner at $(\tilde{P}_M, \mathbf{P}_{-M})$, even if $x = d$, since $\bar{S}(a) = \bar{S}(b) = \bar{S}(d) = 1$ and all these alternatives are eliminated in the first round.

(3.3) Consider the case $n = 6$. Again we have $S(a) = 2$, $S(c) = 2$, and members of K manipulating by voting for b. If $S(d) = 1$, then again d will be eliminated at $(\tilde{P}_M, \mathbf{P}_{-M})$. So, $S(d) = 2$. Let c be the STV winner at $\mathbf{P}$. Since $S(a) = S(c) = S(d) = 2$ it follows that cP^td (where P^t is the tie-breaking order). So, $\bar{S}(a) = 1$, $\bar{S}(b) = 1$, $\bar{S}(c) = 2$, and $\bar{S}(d) = 2$. Therefore in round 1 at $(\tilde{P}_M, \mathbf{P}_{-M})$ alternatives a and b are eliminated. But as $aPbPcPd$, the score of c in round 2 at $(\tilde{P}_M, \mathbf{P}_{-M})$ is at least 3. By $cP^t d$, it follows that d cannot be the STV winner at $(\tilde{P}_M, \mathbf{P}_{-M})$.

(3.4) Finally, consider the case $n = 7$. Since c is the STV winner at $\mathbf{P}$, $S(a) \neq 3$ and $S(d) \neq 3$. Otherwise c would be eliminated in the first round at $\mathbf{P}$. If $S(c) = 3$, then $\tilde{S}(a) = 0$, $\tilde{S}(b) = \tilde{S}(d) = 2$ and $\tilde{S}(c) = 3$. So, b is eliminated at $(\tilde{P}_K, \mathbf{P}_{-K})$ in the first round. Therefore, $S(a) = S(c) = S(d) = 2$ and $S(b) = 1$. Also, $|M| = 1$. Hence, $\bar{S}(a) = 1$ and $\bar{S}(b) = \bar{S}(c) = \bar{S}(d) = 2$. therefore, in round 1 at $(\tilde{P}_M, \mathbf{P}_{-M})$ alternative a is eliminated. Since all agents in K have preference $aPbPcPd$ it follows that in round two (after eliminating a) the score of b has increased by one to 3 whereas the scores of c and d are unchanged. Therefore, in round 2 at $(\tilde{P}_M, \mathbf{P}_{-M})$ alternatives c and d are eliminated. This contradicts that d is the STV winner in profile $(\tilde{P}_M, \mathbf{P}_{-M})$.

Thus, if $m \geq 4$ and $n \leq 7$, then there are no unsafe manipulations. This concludes the proof of the theorem. □

7 Concluding remarks

7.1 Relation with Slinko and White (2014)

In this subsection we compare our work with Slinko and White (2014) – henceforth SW.

For a rule F and a preference profile $\mathbf{P}$, according to SW a voter i with group K has an incentive to manipulate if there is a preference $\tilde{P} \in L(X)$ and a set $G \subseteq K$ with $i \in G$ such that $F(\tilde{P}_G, \mathbf{P}_{-G}) P_i F(\mathbf{P})$. Observe that SW do not require that all voters in K deviate to $\tilde{P}$. Clearly, if i has an incentive to manipulate in our sense (Definition 2.1), then i has an incentive to manipulate according to SW (simply take $G = K$), but the converse is not necessarily true.

Next, SW call such a manipulation by $\tilde{P}$ unsafe if there exists $M \subseteq K$ with $i \in M$ such that all members of M have an incentive to manipulate by $\tilde{P}$, but $F(\mathbf{P}) P_i F(\tilde{P}_M, \mathbf{P}_{-M})$; and safe if for all $U \subseteq K$ with $i \in U$, we have $F(\tilde{P}_U, \mathbf{P}_{-U}) P_i F(\mathbf{P})$ or $F(\tilde{P}_U, \mathbf{P}_{-U}) = F(\mathbf{P})$. Hence, if i has an incentive to manipulate by $\tilde{P}$ in our sense, so that, by the preceding paragraph, i also has an incentive to manipulate by $\tilde{P}$ according to SW, then if this manipulation is (un)safe in our sense (see Sect. 2.2), it is also (un)safe according to SW.

The definitions of (un)safely manipulable preference profiles and rules in SW are similar to ours (Sect. 2.2), so that we obtain the following corollary.

Corollary 7.1 *If a rule is safely (unsafely) manipulable for some m and n, then is is also safely (unsafely) manipulable according to SW.*

Thus, results about the (un)safety of manipulation in our sense are applicable to the model of SW. Unfortunately, we cannot directly adapt results in our paper about cases where we have only safe manipulations, to the model of SW in the same way. Indeed, in preference profiles where there is no manipulation in our sense there could still be voters having an incentive to manipulate according to SW, and this manipulation could be unsafe.

The main result in SW, their Theorem 3.2, says that for every onto and non-dictatorial rule F with range at least three there is a preference profile $\mathbf{P}$, a voter i, and a preference $\tilde{P}$, such that i has an incentive to manipulate and this manipulation is safe.

In the SW model, if voter i has an incentive to manipulate safely by $\tilde{P}$ in $\mathbf{P}$, this does not necessarily imply that the same voter has an incentive to manipulate in our model, since this safe manipulation according to SW still allows for $F(\tilde{\mathbf{P}}_K, \mathbf{P}_{-K}) = F(\mathbf{P})$. Thus, if a rule is safely manipulable according to SW, it does not follow directly from the definitions that the same holds in our model and for this reason Theorem 3.2 of SW does not carry over directly to our model. However, we can prove that if a rule F is manipulable in our model, then a safe manipulation also exists.

Theorem 7.2 *If a rule is manipulable, then it is also safely manipulable in our model.*

Proof Let $\mathbf{P}$ be a manipulable profile, hence there are $\tilde{P} \in L(X)$ and $i \in N$ such that $F(\tilde{P}_K, \mathbf{P}_{-K})P_iF(\mathbf{P})$. If this manipulation is safe, then we are done. If this manipulation is unsafe, then there is an $M \subset K$ with $i \in M$ such that $F(\mathbf{P})P_iF(\tilde{P}_M, \mathbf{P}_{-M})$ and, consequently, $F(\tilde{P}_K, \mathbf{P}_{-K})P_iF(\tilde{P}_M, \mathbf{P}_{-M})$. Consider the profile $\mathbf{P}' = (P_{K\setminus M}, \tilde{P}_M, \mathbf{P}_{-K})$. Now $K' = K \setminus M$ is a group and members of K' have an incentive to manipulate with $\tilde{P}$. Again, if this manipulation is safe, we are done. Otherwise, by the same reasoning there is $M' \subset K'$ that $F(\mathbf{P}')P_iF(\tilde{P}_{M'}, \mathbf{P}'_{-M'})$; and so on. This way we either find a safe manipulation or end up with a group of size one, and the single member of this group has a trivially safe manipulation.

7.2 Further remarks

We have considered the safety of group manipulation for several rules, and established conditions for the existence of safe and unsafe manipulations. Theorem 7.2 says that if a rule is manipulable (by a group), then it is safely manipulable. The situation is different for unsafe manipulation. For instance, scoring rules with one jump in a scoring vector turn out to be only safely manipulable, which means that they do not allow for unsafe manipulations at all. The other rules that we considered, are manipulable in an unsafe way. A more detailed analysis, however, shows that even for unsafely manipulable rules the existence of an unsafe manipulation depends on the number of voters and alternatives. For the rules under consideration in this paper, we have established exact bounds for these values. Moreover, even if we know that for the given values of m and n a social choice rule allows for an unsafe manipulation, this does not mean that any group manipulation is unsafe and, thus, risky. It only means that in some preference profile unsafe manipulation is possible. We do not have a general picture of how often unsafe manipulations occur.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aleskerov F, Kurbanov E (1999) Degree of manipulability of social choice procedures. Curr Trends Econ 13–27

Elkind E, Grandi U, Rossi F, Slinko A (2015) Gibbard-satterthwaite games. In: IJCAI

Gibbard A (1973) Manipulation of voting schemes: a general result. Econometrica 41:587–601

Grandi U, Hughes D, Rossi F, Slinko A (2019) Gibbard-Satterthwaite games for k-approval voting rules. Math Soc Sci 99:24–35

Hazon N, Elkind E (2010) Complexity of safe strategic voting. In: International symposium on algorithmic game theory, Springer, 210–221

Ianovski E, Yu L, Elkind E, Wilson MC (2011) The complexity of safe manipulation under scoring rules. IJCAI 11:246–251

Kelly JS (1988) 4. minimal manipulability and local strategy-proofness. Soc Choice Welf 5:81–85

Maus S, Peters H, Storcken T (2007) Anonymous voting and minimal manipulability. J Econ Theory 135:533–544

Nitzan S (1985) The vulnerability of point-voting schemes to preference variation and strategic manipulation. Public Choice 47:349–370

Peters H, Roy S, Storcken T (2012) On the manipulability of approval voting and related scoring rules. Soc Choice Welf 39:399–429

Satterthwaite MA (1975) Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. J Econ Theory 10:187–217

Slinko A (2019) Beyond Gibbard and Satterthwaite: Voting ma- nipulation games. Fut Econ Design, Springer, 131–138

Slinko A, White S (2014) Is it ever safe to vote strategically? Soc Choice Welf 43:403–427

Wilson M, Reyhani Shokat Abad R (2010) The probability of safe manipulation. COMSOC 2010

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.