Ann Math Artif Intell (2013) 68:5–30
DOI 10.1007/s10472-012-9326-6

A behavioral perspective on social choice

Anna Popova · Michel Regenwetter · Nicholas Mattei

Published online: 16 January 2013

Abstract We discuss what *behavioral social choice* can contribute to computational social choice. An important trademark of behavioral social choice is to switch perspective away from a traditional sampling approach in the social choice literature and to ask *inference* questions: Based on limited, imperfect, and highly incomplete observed data, what inference can we make about social choice outcomes at the level of a population that generated those observed data? A second important consideration in theoretical and behavioral work on social choice is *model dependence*: How do theoretical predictions and conclusions, as well as behavioral predictions and conclusions, depend on modeling assumptions about the nature of human preferences and/or how these preferences are expressed in ratings, rankings, and ballots of various kinds? Using a small subcollection from the Netflix Prize dataset, we illustrate these notions with real movie ratings from real raters. We highlight the key roles that inference and behavioral modeling play in the analysis of such data, particularly for sparse data like the Netflix ratings. The social and behavioral sciences can provide a supportive role in the effort to develop behaviorally meaningful and robust studies in computational social choice.

A. Popova
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: apopova2@illinois.edu

M. Regenwetter (✉)
Department of Psychology and Department of Political Science,
University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: regent@illinois.edu

N. Mattei
NICTA and University of New South Wales, Sydney, Australia
e-mail: nicholas.mattei@nicta.com.au

Keywords Behavioral social choice · Consensus methods · Inference · Model dependence · Voting paradoxes

Mathematics Subject Classifications (2010) 91B10 · 91B12 · 91B14

1 Introduction

Voting rules and social choice methods have been used for centuries in order to reach collective decisions. Increasingly, in computer science, data collection and reasoning systems are moving towards distributed and multi-agent design paradigms [49]. With this design shift comes the need to aggregate the (possibly disjoint) observations and preferences of individual agents into an overall partial or complete ordering in order to synthesize knowledge and data.

One of the most common methods of preference aggregation and group decision making in human systems is voting. Many societies, both throughout history and across the planet, use voting to arrive at collective decisions on a range of topics from deciding what to have for dinner in a small group to declaring war as a nation. Unfortunately, mathematical results in the field of social choice prove that there is no perfect voting system and, in fact, voting systems can succumb to a host of problems. Arrow's Theorem demonstrates that any preference aggregation scheme for three or more alternatives will fail to meet a set of simple fairness conditions [1]. Each voting method violates one or more properties that most would consider important for a voting rule (such as non-dictatorship) [5, 10, 50]. Similarly, the Gibbard–Satterthwaite Theorem implies that every non-dictatorial voting rule is manipulable [19, 46]. Moreover, one can easily create an example illustrating how competing voting rules can disagree on winners, losers, and social orders. Questions about voting and preference aggregation have circulated in the mathematics and social choice communities for centuries [2, 6, 17, 24, 45].

Many scholars wish to study how often and under what conditions individual voting rules fall victim to violations of various voting laws and axioms [5, 10]. Due to a lack of large, accurate datasets, many computer scientists, economists, and political scientists have turned towards statistical distributions to generate election scenarios in order to benchmark and analyze voting rules and other decision procedures [16, 39, 40, 54]. Commonly used theoretical assumptions about the distribution of preferences in the electorate such as the Impartial Culture assumption (IC, [16]) and the Impartial Anonymous Culture assumption (IAC, [15]) are extreme symmetry assumptions that represent maximum disagreement among voters. These knife edge distributions lead to pessimistic (and arguably even nonsensical) predictions about voting rules [11, 18, 39] which, in turn, can lead to questionable policy recommendations. For instance, some scholars have concluded one should minimize turnout and minimize the number of candidates running for office, if decisions are to be reached by majority rule [48]. By and large, these approaches take a sampling, not an inference, perspective on social choice.

Another famous but problematic theoretical benchmark is the notion of *Condorcet efficiency* (the probability that a voting rule's winner matches the "Condorcet" winner, given that one exists). A candidate who can beat all other candidates in pairwise elections (the *Condorcet winner*) remains a cornerstone in the normative social choice literature. Low Condorcet efficiency under IC and IAC exacerbates the

gloomy predictions from the axiomatic literature about the inability of an electorate to arrive at a group decision [12, 13, 17, 18]. These statistical models may or may not be grounded in reality and it is an open problem in both the political science and social choice fields as to how, exactly, election data may be modeled realistically [26, 35, 36, 51].

A fundamental problem in empirical and behavioral research into properties of voting rules is the lack of large data sets to run empirical studies [35, 51]. There have been studies of several distinct datasets but these are limited in both number of elections analyzed [5, 33] and size of individual elections within the datasets analyzed [10, 23, 51]. While it is too early to judge the frequency with which different voting paradoxes occur in general, or to judge the consensus between voting methods in general, the existing studies so far [25, 35, 38] have found little evidence of *Condorcet's Voting Paradox* [14, 21] (a cyclical majority ordering). At the same time, *preference domain restrictions* such as *single peakedness* [4, 9, 35, 38] (where one candidate out of a set of three is never ranked last), which is a sufficient conditional to eliminate the Condorcet paradox, also did not account well for real data. Additionally, most of the studies have found a strong consensus between most voting rules except Plurality [5, 10, 35].

2 What is behavioral social choice?

The supreme goal of behavioral social choice is to investigate social choice procedures empirically while avoiding unnecessary and/or unsubstantiated assumptions about human behavior. It is critical, in any fully rigorous behavioral paradigm, that all assumptions about human behavior be stated as explicitly as possible. Ideally, any such assumptions should be tested for their validity. Untested assumptions require especially strong motivation and/or scrutiny. In this spirit, a first step in behavioral social choice is to define individual voter preferences in a general and flexible fashion, and then define consensus methods at a level that is applicable to such general definitions of preference. Our first definition introduces mathematical concepts and terminology as given by Roberts [41], and as commonly used by U.S. scholars (but not as routinely used by European scholars, due to language differences).

Definition 1 Let $\mathscr{C}$ be a finite set of *choice alternatives* or *candidates*. A *binary (preference) relation* R on $\mathscr{C}$ is a collection of ordered pairs of elements of $\mathscr{C}$, i.e., $R \subseteq \mathscr{C} \times \mathscr{C}$. We also write xRy for $(x, y) \in R$. If R and S are two (binary) relations on $\mathscr{C}$, we write $RS = \{(z, y) \in \mathscr{C} \times \mathscr{C} : \exists x \in \mathscr{C}, zRx, xSy\}$. Let $R^{-1} = \{(x, y) \in \mathscr{C} \times \mathscr{C} : yRx\}$, $\overline{R} = (\mathscr{C} \times \mathscr{C}) \setminus R$, and $Id_{\mathscr{C}} = \{(c, c) : c \in \mathscr{C}\}$. A binary relation R on $\mathscr{C}$ is

$$\begin{aligned} \textit{complete} \quad & \text{if} \quad R \cup R^{-1} \cup Id_{\mathscr{C}} = \mathscr{C} \times \mathscr{C}, \\ \textit{asymmetric} \quad & \text{if} \quad R \cap R^{-1} = \varnothing, \\ \textit{negatively transitive} \quad & \text{if} \quad \overline{R}\,\overline{R} \subseteq \overline{R}, \\ \textit{transitive} \quad & \text{if} \quad RR \subseteq R. \end{aligned}$$

A *strict partial order* is an asymmetric and transitive binary relation. An *interval order* is a strict partial order R with the property that $R\overline{R^{-1}}R \subseteq R$. A *semiorder*

is an interval order R with the property that $RR\overline{R^{-1}} \subseteq R$. A *strict weak order* is an asymmetric and negatively transitive binary relation. A *strict linear order* is a transitive, asymmetric, and complete binary relation. If we replace "strict preference" by "preference or indifference" then a strict partial/weak/linear order becomes a "partial/weak/linear order." We will assume asymmetric ("strict") preference without loss of generality.

Much of the social choice literature assumes that individual preferences are (strict) linear orders or (strict) weak orders. Within the field of computational social choice there is some use of other information models, specifically (strict) partial orders, where questions of winner determination [55] and manipulation [7] have been addressed. There has also been some work on winner determination and manipulation when voters express probabilities over their preferences [8, 20]. However, despite these forays into more complex information models, the bulk of the work in computational social choice still assumes that strict linear orders are either available, or that they are at least reasonable hypothetical constructs even if not directly observable. The goal of this paper is to highlight, by providing additional references and concrete examples, the pitfalls that may befall scholars, e.g., in computational social choice, as they move from the theoretical to the empirical.

A "profile" in classical voting theory is typically a mapping from the set of individual preferences into the natural numbers, i.e., a vector of voter frequencies or proportions indexed by the appropriate set of binary preferences, such as strict linear orders. We will generalize that definition to include a range of behaviorally important applications. First, it seems reasonable to assume asymmetry because it simply captures "strict" preference (as opposed to "preference or indifference"). The two key generalizations are that preferences can be any asymmetric binary relations of any kind, and that we move from frequencies (or proportions) of binary relations to probabilities of binary relations.

Definition 2 Let $\mathscr{C}$ be a finite set of *choice alternatives* or *candidates*. Let $\mathscr{R}$ denote the collection of all asymmetric binary relations on $\mathscr{C}$. A *profile* P is a probability distribution over $\mathscr{R}$:

$$P : \mathscr{R} \to [0, 1]$$
$$R \mapsto P(R).$$

The classical model where a profile is viewed as proportions of people who hold various strict weak orders is a special case concentrating all probability mass on strict weak orders and where P is just interpreted as a probability measure representing proportions.

In order to define a broad range of consensus methods, such as, e.g., "scoring rules," for such general representations of preferences, we need a mathematical concept of numerical ranks that applies to the general representation. We define the "generalized rank" first axiomatized and discussed in [28].

Definition 3 Let $\mathscr{C}$ be a finite set of n many choice alternatives, i.e., $|\mathscr{C}| = n$. The *differential* $\Delta_R(c)$ of any element $c \in \mathscr{C}$ with respect to a binary relation $R \subseteq \mathscr{C} \times \mathscr{C}$ is

$$\Delta_R(c) = |\{a \in \mathscr{C} : (a, c) \in R\}| - |\{b \in \mathscr{C} : (c, b) \in R\}|.$$

The *generalized rank* $Rank_R(c)$ of c with respect to R is given by

$$Rank_R(c) = \frac{n + 1 + \Delta_R(c)}{2}.$$

Note that generalized ranks are multiples of $\frac{1}{2}$. For strict linear orders, they are the usual integer valued ranks associated with complete rankings without ties. Also, note that, still with $|\mathscr{C}| = n$,

$$Rank_R(c) = 1 \Leftrightarrow [(c, b) \in R, \forall b \in \mathscr{C}, b \neq c]$$

and

$$Rank_R(c) = n \Leftrightarrow [(a, c) \in R, \forall a \in \mathscr{C}, a \neq c].$$

In other words, a candidate has generalized rank 1 if it is strictly preferred to all other candidates, and an option has generalized rank n if all other options are strictly preferred to it. We will utilize the concept of generalized rank both at the individual preference level and at the social welfare level.

We are now ready to define the five social choice procedures we will consider here, Condorcet, Borda, Plurality, Antiplurality, and Plurality Runoff, for general representations of preferences. The definitions of Condorcet, Borda, Plurality, and Antiplurality are from [29, 33], the definition of Plurality Runoff is new.

Definition 4 Let P be a profile on the collection $\mathscr{R}$ of binary relations on a finite set $\mathscr{C}$ of choice alternatives with $|\mathscr{C}| = n$. Let $c, d \in \mathscr{C}$. *Condorcet* is a pairwise comparison procedure:

$$c \text{ is Condorcet preferred to } d \Leftrightarrow \sum_{\substack{R \in \mathscr{R} \\ (c,d) \in R}} P(R) > \sum_{\substack{R' \in \mathscr{R} \\ (d,c) \in R'}} P(R').$$

Borda, *Plurality*, and *Antiplurality* are *scoring rules* in that they assign scores to choice alternatives as a decreasing function of their generalized ranks in an individual's preference:

$$Borda(c) = \sum_{R \in \mathscr{R}} P(R)\,[n - Rank_R(c)],$$

$$Plurality(c) = \sum_{\substack{R \in \mathscr{R} \\ Rank_R(c) = 1}} P(R),$$

$$Antiplurality(c) = \sum_{\substack{R \in \mathscr{R} \\ Rank_R(c) = n}} P(R).$$

To derive the pairwise preferences for Borda, Plurality, and Antiplurality, we only need to compare scores:

$$c \text{ is Borda preferred to } d \Leftrightarrow Borda(c) > Borda(d),$$

$$c \text{ is Plurality preferred to } d \Leftrightarrow Plurality(c) > Plurality(d),$$

$$c \text{ is Antiplurality preferred to } d \Leftrightarrow Antiplurality(c) < Antiplurality(d).$$

A winner under *Plurality Runoff* first requires that there must be a unique set of two candidates, say $\{x, y\} \subseteq \mathscr{C}$, such that x and y are the two options with the highest plurality scores. If such a set exists, then

$$x \text{ is Plurality Runoff winner if } \sum_{\substack{R \in \mathscr{R} \\ (x,y) \in R}} P(R) > \sum_{\substack{R' \in \mathscr{R} \\ (y,x) \in R'}} P(R'),$$

$$y \text{ is Plurality Runoff winner if } \sum_{\substack{R \in \mathscr{R} \\ (x,y) \in R}} P(R) < \sum_{\substack{R' \in \mathscr{R} \\ (y,x) \in R'}} P(R').$$

In all other cases, Plurality Runoff yields no winner.

Prior work on behavioral social choice has used such generalized definitions, as well as similarly general definitions for various "utility" and "random utility" representations, to compute social choice outcomes from a variety of empirically generated inputs.

In earlier work, a number of papers [5, 10, 26, 27, 30–35, 52, 53] considered general definitions of Condorcet and investigated the empirical prevalence of Condorcet cycles, e.g., where A is Condorcet preferred to B, B is Condorcet preferred to C and C is Condorcet preferred to A. They investigated approval voting ballots from which they inferred probability distributions over strict linear orders [5, 10, 26, 27, 30, 35]. They also analyzed various national election survey data from France, Germany, and the United States, where they interpreted numerical ratings of candidates as strict weak orders or as semiorders [31, 32, 35]. This literature found virtually no evidence for Condorcet cycles in empirical data. They also compared Condorcet and Borda outcomes for strict linear order preferences inferred from approval voting ballots and concluded that Condorcet and Borda led to virtually identical outcomes. More recently, behavioral social choice researchers have found different consensus methods, such as Condorcet, Borda, and Plurality, to agree with each other extensively, especially on candidates with generalized rank 1 or generalized rank n (out of n candidates) [25, 36–38]. All of the empirical studies surveyed [5, 10, 23, 35, 37, 51] came to a similar conclusion: there is scant evidence for occurrences of Condorcet's Paradox [24]. Many of these studies find no occurrence of majority cycles (and those that find cycles find them in fewer than 1 % of elections). Additionally, each of these (with the exception of Niemi and his study of university elections, which he observes is a highly homogeneous population [23]) find almost no occurrences of either single-peaked preferences [4] or the more general value-restricted preferences [35, 47].

Two important concepts have become prominent in prior behavioral social choice analyses:

1. Inference: When investigating social choice outcomes on empirical data, one should evaluate how 'confident' one can be about finding the 'correct' outcomes if one thinks of the data as imperfect and incomplete reflections of the electorate's preference profile. So far, the main tools for evaluating the statistical confidence or *replicability* of social choice outcomes have been a Bayesian inference framework [29, 35] and a bootstrap approach [25, 37, 38]. In the bootstrap, one samples N many observations with replacement from an original data set of N many observations and records the outcomes of the social choice procedures of interest. In our analysis for the results section, we used a pseudo-random sampling procedure in MATLAB to draw such bootstrap samples of size N each. We repeated this process 10,000 times to check what proportion of 10,000 bootstrap samples replicated the social order found in the original data set. The larger the number of bootstrap samples that match a result in the

original profile, the higher the *confidence* in and *replicability* of the finding in the original profile. The idea behind the bootstrap is to quantify how resilient the social choice outcome is to perturbations in the data. Prior analyses of empirical data with these inference tools have suggested that Condorcet paradoxes can be ruled out with high replicability and that different social choice procedures agree with each other on the winner and loser with high replicability.

2. Model Dependence: Theoretical and empirical analyses of social choice rules can depend to various degrees on the modeling assumptions about individual preferences. In the behavioral analyses we have cited [25, 36–38], the common finding was that the election winners and social orders often depended on modeling assumptions, but the absence of a Condorcet paradox and the agreement among consensus methods did not hinge on a specific model being used.

The rest of this paper offers an illustration of behavioral social choice on new data. We will see whether the earlier inference and model dependence findings appear to extend readily to the much sparser data sets of the Netflix Prize. We will see that the picture for the Netflix data will be more complicated.

3 Data

We have extracted consumer ratings from the Netflix Prize dataset [3]. Netflix is a company based in the USA where users pay a flat monthly fee and either receive DVD's by mail or have video content delivered over the web. A central component of the Netflix service is its recommendation engine. Netflix encourages users to submit ratings (between 1 and 5 stars) of the movie they have just watched or of any other movies, e.g., movies they may have seen on Netflix or elsewhere in the past. Based on these ratings, users receive recommendations for other movies that they may enjoy based on what they have viewed and/or rated thus far.

The Netflix dataset offers a vast amount of rating data; compiled and publicly released by Netflix for its Netflix Prize [3]. There are 100,480,507 distinct ratings in the database. These ratings cover a total of 17,770 movies and 480,189 distinct users. Each user has provided ratings on a five-point scale (the rating ⋆ is the lowest, the rating ⋆⋆⋆⋆⋆ is the highest) for any number of movies, with some raters having rated as many as thousands of movies, while others have rated just a handful. While all movies have at least one score, every user has rated only a small fraction of all the movies. According to Netflix, the dataset contains every movie rating received by Netflix, from its users, between early 2002, when Netflix started tracking the data, and late 2007, when the competition for the Prize was announced. These data have been anonymized to protect privacy and are conveniently coded for use by researchers.

The Netflix data are rare in preference studies: Since users of the Netflix service can expect to receive presumably higher quality recommendations from Netflix if they respond truthfully to the rating prompt, there is an incentive for each user to express sincere preference in their ratings. In the Netflix setup, the user is receiving a tangible benefit (clearer and more accurate recommendations) for providing truthful data. With Netflix's catalog of over 17,000 movies, users need help sorting through all the data, especially if they are interested in discovering great movies that they don't already know. This is in contrast to many other datasets which are compiled through

surveys or other methods where the individuals questioned about their preferences often have little or no stake in providing truthful responses. The Netflix rating system also gives viewers a natural incentive to rate as many movies as possible, as long as they have clearly formed preferences among them, since more information from the user will presumably lead to more accurate and more relevant recommendations.

To illustrate the role of behavioral social choice in empirical studies, we selected three sets of five movies from the dataset. The first two sets were more or less selected at random. These movies had a fairly high number of joint ratings, that is, users who had rated multiple movies out of the set. The third was selected so that all five movies in the set had received a similar and large number of ratings. For this last set, we found five movies that had all received $10{,}040 \pm 10$ user ratings. Brief summaries of the movies we selected can be found in the top panels of Tables 1, 2 and 3. All movie descriptions and genre information are taken from the respective movie page at the Internet Movie Database (www.imbd.com).

Tables 1–3 provide various types of summary information about the three movie sets and their ratings. For example, Table 1 shows, for Movie Set 1, that only 91 raters offered ratings for movie A, "Bliss: Season 1," and of these, 23 gave a ⋆-rating and 11 gave a ⋆⋆⋆⋆⋆-rating. In contrast, more than 150,000 viewers rated movie E, "Lost in Translation." The table also provides the arithmetic average of the star-ratings for each movie among the raters who rated that given movie. Among those who rated "Jaws," the average rating is 3.89 stars, whereas among those who rated "Bliss," the average rating was 2.56 stars. It is not clear whether it is meaningful to use an arithmetic average: we do not know whether these ⋆-ratings form an *interval scale*, according to which the difference between a ⋆⋆⋆⋆⋆-rating and a ⋆⋆⋆-rating expresses the same "strength of preference" as the difference between a ⋆⋆⋆-rating and a ⋆-rating [41]. In other words, arithmetic averages could be meaningless summary statistics [42–44]. The median rating is perhaps the more appropriate summary statistic, though much less refined. We report the medians for all movies in Tables 1–3. A major rationale behind social choice aggregation methods is to use, as input into the consensus method, only ordinal information from each judge.

In the usual incarnation, social choice theory uses ordinal, rather than quantitative, input about individual preferences as the theoretical primitive. However, much social choice theory is based on the assumption that individual voters/judges have asymmetric, complete and transitive (strict linear order) preferences among the candidates/options. There is very little reason to believe this assumption in the context of Netflix movie viewers, especially that individual preferences ought to be complete. It does not make sense to assume that anyone even knows all of these 17,000+ movies. It also makes little sense to assume that viewers have a strict preference among every two movies, and this is reflected by the fact that Netflix only uses a simple five-point scale for rating the movies. It also may not be legitimate to assume complete preferences over groups of movies, say, if one attempted to reduce the numbers by grouping movies into genres, release dates, and/or other criteria in an effort to sort them into equivalence classes of sorts.

If we just consider the five movies in each of the three sets we have selected for analysis, it is striking from Tables 1–3, that of those few people who rated all movies in a given set not a single one mapped the movies one-to-one into ratings. There is no evidence in these data, not even from a single rater, that would suggest

Table 1 Movie Set 1: Synopsis (*top panel*), summary of ratings (*center panel*), and full ratings of those 15 viewers who rated the entire movie set

Movie Set 1, Description:				
Movie No.	Title	Year	Genre	Synopsis
3462	Bliss: Season 1	2002	Drama-Romance	A Showtime Original Series that explores the desires, passions and fantasies of women.
798	Jaws	1975	Thriller	A giant great white shark threatens a small fishing community and a group of men set out to stop it.
758	Mean Girls	2004	Comedy-Drama	A high school teen drama centering on two girls fighting over a boy.
1145	The Wedding Planner	2001	Comedy-Romance	A wedding planners life is turned upside down when she falls head over heels for a client.
12232	Lost in Translation	2003	Drama	A movie star with a sense of emptiness, and a neglected newlywed meet in Tokyo and form an unlikely bond.

Rating summary:					
Rating	A	B	C	D	E
	Bliss: Season 1	Jaws	Mean Girls	The Wedding Planner	Lost in Translation
⋆	23	1,219	2,899	12,194	15,750
⋆⋆	23	4,527	9,773	23,238	23,578
⋆⋆⋆	27	20,240	38,016	49,351	36,605
⋆⋆⋆⋆	7	31,606	38,099	37,366	41,143
⋆⋆⋆⋆⋆	11	23,686	15,575	18,005	35,330
Number of raters:	91	81,278	104,362	140,154	152,406
Mean rating:	2.56 stars	3.89 stars	3.51 stars	3.18 stars	3.37 stars
Median rating:	⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆

Ratings of those 15 viewers who rated all five movies:							
Number of	Movies					Generalized rank	
raters	A	B	C	D	E	1	5
1	⋆	⋆⋆	⋆⋆⋆	⋆⋆	⋆⋆	C	A
1	⋆	⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	E	A
1	⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆	C	A
1	⋆	⋆⋆⋆⋆	⋆⋆	⋆⋆	⋆	B	–
1	⋆	⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆	⋆⋆⋆⋆	–	A
1	⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	–	A
1	⋆⋆⋆	⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆	C	B
2	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆	–	E
1	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆	C	E
1	⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆	⋆⋆	⋆⋆⋆⋆	–	D
1	⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	E	–
1	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	–	A
1	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆⋆	–	A
1	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆	⋆⋆	B	–

Table 2 Movie Set 2: Synopsis (*top panel*), summary of ratings (*center panel*), and full ratings of those 5 viewers who rated the entire movie set

Movie Set 2, Description:				
Movie No.	Title	Year	Genre	Synopsis
15870	Anna Karenina	1967	Drama-Romance	A young wife of an older husband complicates her life by having an affair.
12458	Splendor	1999	Comedy-Romance	A twenty-something starts a romantic affair with two men at the same time.
2568	StarGate SG-1: Season 8	2004	Action-SciFi	A secret military team is formed to explore the StarGate.
2651	Dragon Ball Z: Super Android 13	1992	Animation-Action	A team of super-humans fights an interplanetary force of androids.
197	Taking Lives	2004	Mystery-Thriller	An FBI profiler is called in to catch a serial killer.

Rating summary:					
Rating	A	B	C	D	E
	Anna Karenina	Splendor	StarGate SG-1: Season 8	Dragon Ball Z: Super Android 13	Taking Lives
⋆	29	102	49	460	2,900
⋆⋆	25	188	32	187	9,017
⋆⋆⋆	61	406	129	422	27,651
⋆⋆⋆⋆	36	302	425	567	29,649
⋆⋆⋆⋆⋆	22	127	1,177	790	12,043
Number of raters:	173	1,125	1,812	2,426	81,260
Mean rating:	2.98 stars	3.15 stars	4.46 stars	3.43 stars	3.48 stars
Median rating:	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆

Ratings of those 5 viewers who rated all five movies:							
Number of	Movies					Generalized rank	
raters	A	B	C	D	E	1	5
1	⋆	⋆	⋆	⋆	⋆	–	–
1	⋆	⋆	⋆	⋆	⋆⋆⋆	E	–
1	⋆⋆	⋆⋆	⋆	⋆	⋆⋆	–	–
1	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	–	A
1	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	–	A

that asymmetric, transitive, complete preferences are behaviorally valid. We should indeed be wary of making such an assumption.

The insight that we have detailed information from very few users and the insight that we should not assume preferences to be complete, have important implications that are hard to overstate. In fact, one of the main take-home messages of this paper is that we face two monumental challenges in evaluating consensus outcomes:

1. In any situation like the Netflix data sets, and even in most ballot profiles from real elections, we only have very limited, incomplete, and possibly inaccurate information about each individual's preferences. This forces us to consider consensus as an *inference* problem.

Table 3 Movie Set 3: Synopsis (*top panel*), summary of ratings (*center panel*), and full ratings of those 30 viewers who rated the entire movie set

Movie Set 3, Description:				
Movie No.	Title	Year	Genre	Synopsis
14731	The Good Son	1993	Drama-Thriller	A young boy moves in with his relatives and begins tormenting his young cousins.
17491	Like Mike	2002	Comedy-Family	A young orphan becomes an NBA star after finding a pair of Michael Jordan's shoes.
433	Untamed Heart	1993	Drama-Romance	Girl meets boy, falls in love and into tragedy.
5650	Buena Vista Social Club	1999	Documentary-Music	Documentary about the life and times of aging Cuban musicians.
13244	Striking Distance	1993	Action-Crime	Police officer searches for the true perpetrator of a murder.

Rating summary:					
Rating	A	B	C	D	E
	Striking Distance	Untamed Heart	Buena Vista Social Club	Like Mike	The Good Son
★	418	290	340	621	205
★★	1,354	980	777	1,264	902
★★★	3,831	3,512	2,393	3,547	3,691
★★★★	3,278	3,384	3,938	3,126	3,808
★★★★★	1,153	11,877	2,595	1,488	1,442
Number of raters:	10,034	10,043	10,043	10,046	10,048
Mean rating:	3.38 stars	3.55 stars	3.76 stars	3.36 stars	3.54 stars
Median rating:	★★★	★★★★	★★★★	★★★	★★★★

Ratings of those 30 viewers who rated all five movies:							
Number of raters	Movies					Generalized rank	
	A	B	C	D	E	1	5
2	★	★	★	★	★	–	–
1	★	★★★	★★★	★	★	–	–
1	★★	★★★	★	★	★★★★	E	–
1	★★	★★★	★★★	★★★	★★	–	–
1	★★	★★	★★	★★	★★	–	–
1	★★★	★★	★	★★	★★★	–	C
1	★★★	★★	★	★★★	★	–	–
1	★★★	★★	★★★	★★★	★★★	–	B
1	★★★	★★★	★★	★★	★★	–	–
1	★★★	★★★	★★	★★★	★★★	–	C
1	★★★	★★★	★★★	★★	★★★★	E	D
1	★★★	★★★	★★★	★★★	★★★	–	–
1	★★★	★★★	★★★★	★★	★★★★	–	D
1	★★★	★★★	★★★★	★★★★	★	–	E
1	★★★	★★★★	★	★	★	B	–
1	★★★	★★★★	★★	★★★★	★★	–	–
1	★★★	★★★★	★★★	★	★★★	B	D
1	★★★	★★★★★	★★★★	★★	★★★	B	D
1	★★★★	★	★	★★★	★★★★	–	–
1	★★★★	★	★★	★★★	★★★	A	B
1	★★★★	★★★	★★	★★★★	★★★★	–	C
1	★★★★	★★★	★★	★★★★★	★★★	D	C
1	★★★★	★★★	★★★	★★★	★★★	A	–
1	★★★★	★★★★	★★★	★★	★★★	–	D
1	★★★★	★★★★★	★★★	★★★	★★★★	B	–
1	★★★★	★★★★★	★★★★★	★★★★★	★★★★★	–	A
1	★★★★★	★★	★★★	★★★★★	★★★	–	B
1	★★★★★	★★★★★	★★	★★★★	★★★★	–	C
1	★★★★★	★★★★★	★★★★★	★★★★★	★★★★★	–	–

2. When we attempt to interpret data as partial indicators of preferences, we must be highly attentive to the modeling assumptions we make and how they may affect our substantive conclusion, such as, e.g., our inferences about the consen-

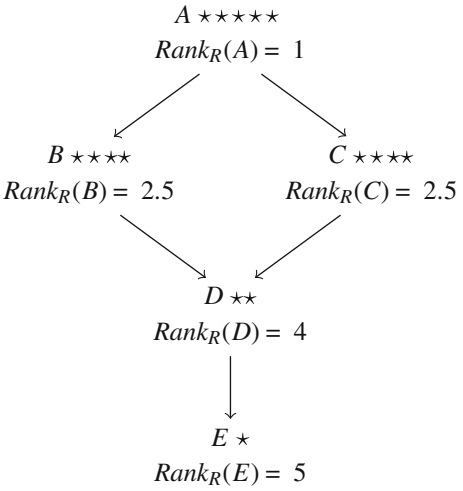

Fig. 1 Hasse diagram of the binary preference relation of a hypothetical viewer who rated all five movies. Arrows indicate strict preference, with arrows implied by transitivity omitted

sus outcomes. In other words, we face a problem of potential *model dependence* of our analyses and conclusions.

We highlight these two problems, because such concerns are second-nature to quantitative or mathematical behavioral scientists, but, not being questions of computational complexity per se, they may not be quite so salient in the computational social choice community at large.

The goal of this paper is not to develop and find the most accurate and refined model of movie rating behavior. That appears like a daunting task. Rather, we illustrate the role of any such model in the analysis of social choice procedures. For purposes of illustration, in this paper, we will thus use three simple models of how binary preferences may be expressed in Netflix movie ratings. More precisely, the three models specify how preferences can be inferred from movies ratings. One model takes an "agnostic" view in that it specifically avoids assuming preferences that involve unrated choice options. This model is based on the "strict partial order" or "Zwicker" model of prior analyses of partial ranking ballots [38]. The second model takes the "pessimistic" view, according to which each rater dislikes unrated movies more than any movies she has rated so far. This model is motivated by the "strict weak order" model used previously for the analysis of partial ranking ballots [37, 38] according to which all candidates ranked on a partial ranking ballot are preferable to all unranked candidates, and according to which the voter has no strict preference among any unranked candidates. The third model takes a "anchor-and-adjust" point of view, according to which the 'default' rating of a movie is ⋆⋆⋆, unless the viewer has given the movie an explicit rating himself. All three models assume that a rater prefers movie x to movie y whenever she gives x a higher rating than y.

Figure 1 shows the Hasse diagram of an example where a hypothetical person gave ratings to all five movies, say, A: ⋆⋆⋆⋆⋆, B: ⋆⋆⋆⋆, C: ⋆⋆⋆⋆, D: ⋆ ⋆, and E: ⋆. Under all three models, this translates into the binary relation

$$R = \{(A, B), (A, C), (A, D), (A, E), (B, D), (B, E), (C, D), (C, E), (D, E)\}$$

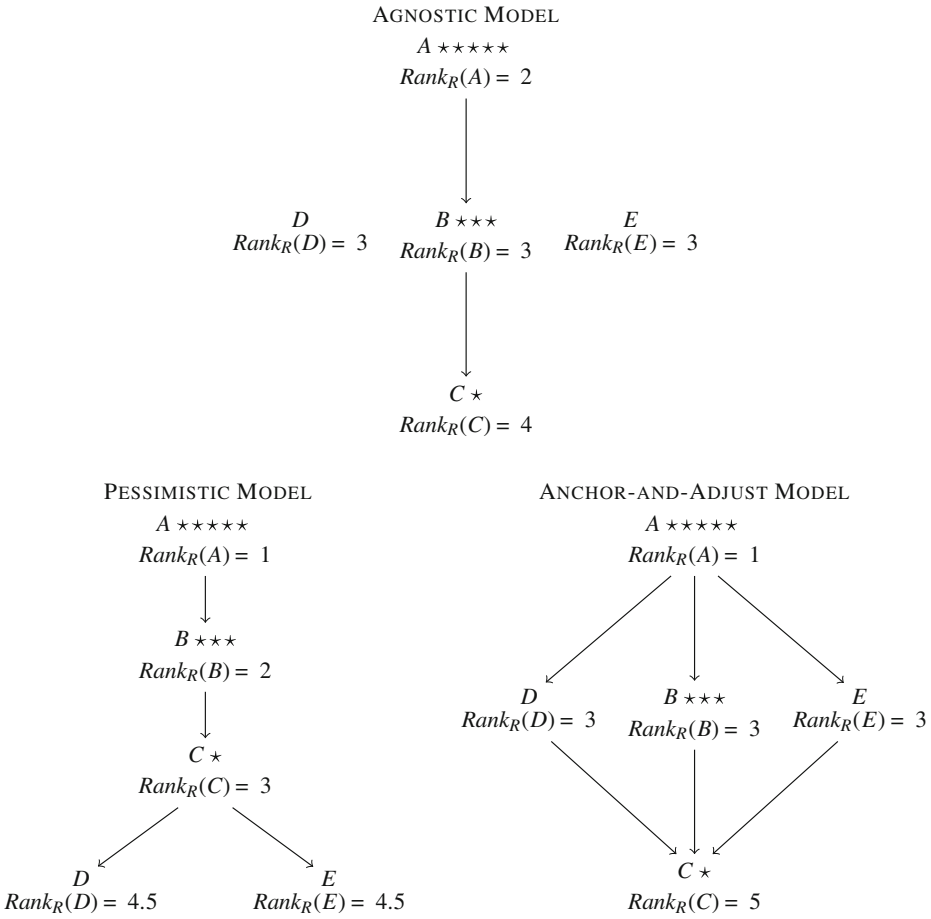

Fig. 2 Hasse diagrams for three models of the binary preference relation of a hypothetical viewer who has rated only some but not all movies in a set. The figure shows an arrow from a preferred movie to a less liked movie, with arrows implied by transitivity omitted

depicted by the Hasse diagram in Fig. 1. The figure also shows the generalized rank of each option in that preference relation: $Rank_R(A) = 1$ because A is strictly preferred to all other options. $Rank_R(B) = Rank_R(C) = 2.5$, whereas D and E have generalized ranks 4 and 5, respectively.

The models differ in how they deal with the many missing ratings. Figure 2 illustrates how the three models assign binary preference relations to viewers who did not rate all five movies in a set. Imagine that a rater gives, say, A: ⋆⋆⋆⋆, B: ⋆⋆⋆, C: ⋆, and does not rate D and E. According to the *Agnostic* model, this person prefers movies they gave more stars to movies they gave fewer stars and has no other strict preferences. This model yields a strict partial order, here, the binary relation

$$\{(A, B), (B, C), (A, C)\}.$$

From the *Pessimistic* model's view point, while this person prefers movies with more stars to movies with fewer stars, the key difference to the first model is that unrated movies are treated as though they had zero stars. This yields a strict weak order where the unrated movies are tied at the bottom of the strict weak order, here

$$\{(A, B), (A, C), (A, E), (A, F), (B, C), (B, E), (B, F), (C, E), (C, F)\}.$$

The third model anchors all movies at a default ⋆⋆⋆-rating and then adjusts the ratings of those movies that the viewer has indeed rated. Beyond that assumption, the *Anchor-and-Adjust* model then assumes that this person prefers movies with more stars to movies with fewer stars and has no other strict preferences, here

$$\{(A, B), (A, C), (A, D), (A, E), (B, C), (D, C), (E, C)\}.$$

Note that the three models are mutually irreconcilable in their assumptions about unrated movies and the strict preference relationships between rated and unrated movies.

4 Results

Table 4 summarizes our inferences made about the social orders under the five consensus methods, using the three models, for the three Netflix movie sets. The top panel shows our results for Movie Set 1, the center panel shows the results for Movie Set 2, and the bottom panel those for Movie Set 3. The Agnostic model, Pessimistic model, and the Anchor-and-Adjust model are arranged from left to right in each panel.

For each social order, we also provide the replicability, by which we mean the proportion of bootstrap samples (out of 10,000) that led to the same social order as did the original data. For example, under the Pessimistic model, all bootstrap samples yielded the social order EDCBA (ranked from best to worst) by Condorcet and Borda, in Movie Sets 1 and 2, as did Plurality in Movie Set 2. In contrast, only 27 % of the 10,000 bootstrap samples replicated the social order marked [CEDBA] under Antiplurality in Movie Set 1. All social orders that we replicated in fewer than 50 % of bootstrapped samples are marked by square brackets [...]. Results with replicability above 95 % are marked in **bold**. For instance, under the Anchor-and-Adjust model interpretation of the data, we have high replicability for all rules in Movie Set 2. Under the Agnostic model interpretation, we have low replicability in most cases. Plurality Runoff only yields a winner, not a social order. Candidates listed in set brackets are tied. For example, under the Agnostic model in Movie Set 1, Plurality yields the unique winner C, followed by a tie between B and E, followed by a tie between A and D. This social order is, however, poorly replicable, as it only occurred in 14 % of our 10,000 bootstrapped samples.

As we reviewed in Section 2, behavioral social choice analyses over the past decade share several common features of their findings. As one shifts one's gaze away from random sampling out of highly artificial distributions like the Impartial Culture, towards considering inference about an underlying population from real empirical data, one perceives a landscape that is very different from that painted on the basis of classical analytical results. On the rare occasion where, in past behavioral social choice analyses, a Condorcet paradox could not be ruled out, some pairwise

Table 4 Behavioral social choice inferences for Movie Sets 1, 2, and 3 under three interpretations of numerical ratings

	Movie Set 1					
	Agnostic		Pessimistic		Anchor-and-Adjust	
	Social Order	Repl.	Social Order	Repl.	Social Order	Repl.
Condorcet	**BCEDA**	0.99	**EDCBA**	1	**BECDA**	1
Borda	**BCAED**	1	**EDCBA**	1	BCEDA	0.57
Plurality	[C{B, E}{A, D}]	0.14	**EDCBA**	0.99	**EDBCA**	1
Antiplurality	[C{B, D}EA]	0.11	[CEDBA]	0.27	**ABCDE**	1
Plur. Runoff	C	0.61	**E**	1	**E**	1

	Movie Set 2					
	Agnostic		Pessimistic		Anchor-and-Adjust	
	Social Order	Repl.	Social Order	Repl.	Social Order	Repl.
Condorcet	[C {cycle} A]	0.19	**EDCBA**	1	**ECDBA**	1
Borda	[CDBAE]	0.26	**EDCBA**	1	**ECDBA**	1
Plurality	E{A, B, C, D}	0.63	**EDCBA**	1	**ECDBA**	1
Antiplurality	{B, C, D, E}A	0.86	[ECDBA]	0.03	**ACBDE**	0.99
Plur. Runoff	E	0.63	**E**	1	**E**	1

	Movie Set 3					
	Agnostic		Pessimistic		Anchor-and-Adjust	
	Social Order	Repl.	Social Order	Repl.	Social Order	Repl.
Condorcet	**CBEAD**	1	[CBEAD]	0.42	CEBDA	0.77
Borda	**BCEAD**	1	[BCEAD]	0.2	CEBDA	0.78
Plurality	[B{A, E}DC]	0.03	CDBAE	0.93	CBEDA	0.78
Antiplurality	[{A,E}B{C,D}]	0.02	BAEDC	0.39	EBCAD	0.53
Plur. Runoff	B	0.59	**C**	1	**C**	1

"Repl." stands for *bootstrapped replicability*

margins were narrow enough that even slight deviations from the observed ballot counts eliminated the paradox. In other words, the Concorcet paradox has been rare and when it could not be ruled out, it had very poor replicability. Our findings here are compatible with that pattern of findings. However, we have a bit of an exception in that this appears to be the first time that we find somewhat (19 %) replicable evidence for a Condorcet cycle. This cycle is located in the middle of the social order for Movie Set 2 under the Agnostic model analysis and does not affect the existence of a Condorcet winner and of a Condorcet loser. In all other Movie Set 1 & 2 analyses, we have a strict linear order by Condorcet with high or perfect replicability. In Movie Set 3, despite the large numbers of ratings, we are confronted with low replicability for Condorcet under two models, i.e., there are narrow margins that can be flipped fairly easily in the bootstrap.

Despite the centuries-old and ongoing debate about the relative merits of Condorcet and Borda, the empirical evidence has suggested over and again that the two rules frequently led to the same social order. Table 4 shows separately computed inferences for Condorcet and Borda, but we can already see that in all cases where we find social orders with high or near perfect replicability, they are also identical. However, there are many cases (many more than in the prior literature we have

cited) in which Condorcet or Borda or both are inferred with low or dismal replicability. In other words, we have many cases where we cannot make solid inferences from the data. This is particularly true for Movie Set 3.

Table 4 highlights the two key messages we hope to convey:

1. Inference: The social orders we computed from these data vary dramatically in how confidently we can make inferences about them from the same set of data if we treat these data as uncertain and incomplete reflections of the population's preferences. An individual preference enters the Plurality tally only when one choice option is preferred to all other options and, hence, when one choice option has generalized rank one. Preferences enter the Antiplurality tally only if they have a choice option to which all other options are strictly preferred, hence if one option has generalized rank five. For the Agnostic model, where we very rarely have a single best or single worst movie for a given rater, Plurality, Antiplurality, and Plurality Runoff depend on the few raters in Table 1 who identify a movie with generalized rank one or five. For Movie Set 1, this leads to low replicability of Plurality, Antiplurality, and Plurality Runoff. Interestingly, in Movie Set 2, because only candidate E is ever at generalized rank 1 (by one rater), and only candidate A is ever at generalized rank 5 (by two raters), the replicability for Plurality and Antiplurality in Movie Set 2 is, in fact, not very low, because more than half of bootstrapped samples include one or more such data points. However, Plurality fails to yield more than a winner and Antiplurality fails to yield more than a loser because there is not enough information in the ratings of Table 2 to yield more consensus information. Hence, there is also not enough information in the bootstrap samples to yield more consensus information. Plurality, Antiplurality, and Plurality Runoff in Movie Set 2 hinge completely on including the two or three informative raters of Table 2 in the tally. If we were to drop the three raters in Table 2, then those three consensus methods would completely collapse. This highlights the importance of considering an inference perspective that takes into account how much information is really contained in a given set of human data and how sensitive our conclusions are to minor or major distortions in those data. When using the Agnostic model, our ability to draw inferences for Plurality, Antiplurality and Plurality Runoff is very limited.
2. Model Dependence: Now, one might think that the easy way out of this problem is to simply add additional information to the data. This is where model dependence comes into play. We know from the inference discussion above that we often have very little confidence that we are able to extract the 'correct' social order for some of the procedures. Hence, to the extent that we gain confidence through imputation of additional information, this confidence may be gained at the cost of additional model dependence, that is, conclusions could very much hinge on the methods by which we might impute additional information. Imputing values for unrated movies can quickly 'take over' in that there can be more imputed ratings than real ratings in the data being aggregated: The hypothetical data may overwhelm the real data and create a false sense of confidence in what the social outcomes are.

Our approach here has been to illustrate the effects of three simple models in our analyses. The Agnostic model did not impute any binary preference information. It captures the idea that viewers cannot possibly view all movies, hence a lack of

Table 5 Movie Set 1, bootstrap replicability using 10,000 bootstrapped samples

Row and Column: Same Unique Winner

	Agnostic Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*1*				
Borda	1	*1*			
Plurality	0.11	0.11	*0.82*		
Antiplurality	0	0	0.28	*0.39*	
Plur. Runoff	0.11	0.11	0.82	0.28	*0.82*

	Pessimistic Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*1*				
Borda	1	*1*			
Plurality	1	1	*1*		
Antiplurality	0.12	0.12	0.12	*0.83*	
Plur. Runoff	1	1	1	0.12	*1*

	Anchor-and-Adjust Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*1*				
Borda	1	*1*			
Plurality	0	0	*1*		
Antiplurality	0	0	0	*1*	
Plur. Runoff	0	0	1	0	*1*

Row and Column: Same Unique Loser

	Agnostic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*1*			
Borda	0.003	*1*		
Plurality	0	0	*0*	
Antiplurality	0.86	0.003	0	*0.94*

	Pessimistic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*1*			
Borda	1	*1*		
Plurality	1	1	*1*	
Antiplurality	1	1	1	*1*

	Anchor-and-Adjust Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*1*			
Borda	1	*1*		
Plurality	1	1	*1*	
Antiplurality	0	0	0	*1*

Table 5 (continued)

Unique Row Winner & Column Loser Exist and Match				
	Agnostic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0	0.003
Borda	0		0	0.003
Plurality	0	0		0.003
Antiplurality	0	0	0	
Plur. Runoff	0	0	0	0.003

	Pessimistic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0	0
Borda	0		0	0
Plurality	0	0		0
Antiplurality	0	0	0	
Plur. Runoff	0	0	0	0

	Anchor-and-Adjust Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0	0
Borda	0		0	0
Plurality	0	0		1
Antiplurality	1	1	1	
Plur. Runoff	0	0	0	1

Replicability of agreement (off-diagonal) and replicability of existence (*diagonal*) of an unambiguous winner (generalized rank 1) in the upper panel, and an unambiguous unique loser (generalized rank 5) in the center panel. The bottom panel gives the replicability of agreement between a unique row winner and a unique column loser

a rating may not tell us anything about the counterfactual whether they 'would' prefer a given unrated movie to one they have already rated. The model captures this intuition formally with the explicit assumption that there is no strict preference involving unrated options. This model led some voting rules to have almost no valid input because very few raters gave enough information for the Agnostic model to yield rankings, or at least single most or single least preferred choice options, from individual decision makers. The Pessimistic model can be thought of as imputing information where none was given, because it assumes that the users are indifferent between all unrated movies and strictly prefer all their rated movies to all their unrated movies. This would make sense if users did not rate a movie because they did not deem it good enough to watch and rate. But clearly, there can be many other reasons for not rating a movie. The Anchor-and-Adjust model captures the intuitive notion that the default rating of a movie is ⋆⋆⋆ and that actual ratings could be upwards adjustments for movies that the rater enjoyed and downwards adjustments for movies that the rater did not enjoy. A similar, but more elaborate model, which we did not include here, would be to use, say, each rater's median ratings as their individual default rating.

Table 6 Movie Set 2, bootstrap replicability using 10,000 bootstrapped samples

Same Unique Winner

	Agnostic Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*0.66*				
Borda	0.66	*1*			
Plurality	0	0	*0.63*		
Antiplurality	0	0	0	*0*	
Plur. Runoff	0	0	0.63	0	*0.63*

	Pessimistic Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*1*				
Borda	1	*1*			
Plurality	1	1	*1*		
Antiplurality	0.51	0.51	0.51	*0.51*	
Plur. Runoff	1	1	1	0.51	*1*

	Anchor-and-Adjust Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*1*				
Borda	1	*1*			
Plurality	1	1	*1*		
Antiplurality	0	0	0	*1*	
Plur. Runoff	1	1	1	0	*1*

Same Unique Loser

	Agnostic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*0.51*			
Borda	0	*1*		
Plurality	0	0	*0*	
Antiplurality	0.46	0	0	*0.86*

	Pessimistic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*1*			
Borda	1	*1*		
Plurality	1	1	*1*	
Antiplurality	0.49	0.49	0.49	*0.80*

	Anchor-and-Adjust Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*1*			
Borda	1	*1*		
Plurality	1	1	*1*	
Antiplurality	0	0	0	*1*

Table 6 (continued)

Unique Row Winner & Column Loser Exist and Match

	Agnostic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0	0
Borda	0		0	0
Plurality	0	0.63		0
Antiplurality	0	0	0	
Plur. Runoff	0	0.63	0	0

	Pessimistic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0	0
Borda	0		0	0
Plurality	0	0		0
Antiplurality	0	0	0	
Plur. Runoff	0	0	0	0

	Anchor-and-Adjust Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0	1
Borda	0		0	1
Plurality	0	0		1
Antiplurality	1	1	1	
Plur. Runoff	0	0	0	1

Replicability of agreement (off-diagonal) and replicability of existence (*diagonal*) of an unambiguous winner (generalized rank 1) in the upper panel, and an unambiguous unique loser (generalized rank 5) in the center panel. The bottom panel gives the replicability of agreement between a unique row winner and a unique column loser

Table 4 shows that the social orders differ substantially within a movie set, depending on the behavioral modeling assumptions that entered the analysis. The three models are simple cases of a potentially large set of conceivable descriptive models one may develop. We used these to illustrate how such models can impact both the conclusions and the replicability of the conclusions one draws.

We now shift our attention from the social orders to just the winners and losers under the various consensus methods.

Tables 5, 6 and 7 show the existence of unique winners, unique losers, and the degree of agreement about winners and losers among consensus methods under the two models for the three data sets. For example, the top panel of Table 5 shows, on the diagonal, the existence of a unique winner (a movie with generalized rank 1) in the social order, for each consensus method. Condorcet and Borda yielded a unique winner in all 10,000 bootstrap samples under all three models for Movie Set 1. The Anchor-and-Adjust model yielded unique winners with perfect replicability for every consensus method. In the other models, antiplurality yielded such a unique winner only in some of the bootstrap samples.

The off-diagonal in the top panel shows how often two rules yielded one and the same movie with generalized rank 1, i.e., the same unique winner. The rates of

Table 7 Movie Set 3, bootstrap replicability using 10,000 bootstrapped samples

Same Unique Winner

	Agnostic Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*1*				
Borda	0	*1*			
Plurality	0	0.59	*0.82*		
Antiplurality	0.01	0.03	0.09	*0.63*	
Plur. Runoff	0	0.59	0.82	0.09	*0.82*

	Pessimistic Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*0.71*				
Borda	0.44	*1*			
Plurality	0.54	0.30	*1*		
Antiplurality	0.10	0.41	0	*0.92*	
Plur. Runoff	0.54	0.30	1	0	*1*

	Anchor-and-Adjust Model				
	Condorcet	Borda	Plurality	Antiplurality	Plur. Runoff
Condorcet	*1*				
Borda	1	*1*			
Plurality	1	1	*1*		
Antiplurality	0	0	0	*1*	
Plur. Runoff	1	1	1	0	*1*

Same Unique Loser

	Agnostic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*1*			
Borda	1	*1*		
Plurality	0	0	*0.46*	
Antiplurality	0.37	0.37	0.20	*0.83*

	Pessimistic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*0.82*			
Borda	0.72	*1*		
Plurality	0.01	0.03	*1*	
Antiplurality	0.03	0.01	0	*1*

	Anchor-and-Adjust Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet	*1*			
Borda	0.98	*1*		
Plurality	0.87	0.87	*1*	
Antiplurality	0.13	0.13	0	*1*

Table 7 (continued)

Unique Row Winner & Column Loser Exist and Match				
	Agnostic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0.02	0.54
Borda	0		0.10	0.30
Plurality	0.03	0.01		1
Antiplurality	0.01	0.03	0.07	
Plur. Runoff	0.03	0.01	0	1

	Pessimistic Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0.46	0.38
Borda	0		0	0.08
Plurality	0.02	0.02		0.03
Antiplurality	0	0	0	
Plur. Runoff	0.02	0.02	0	0.03

	Anchor-and-Adjust Model			
	Condorcet	Borda	Plurality	Antiplurality
Condorcet		0	0	0
Borda	0		0	0
Plurality	0	0		0
Antiplurality	0	0	0	
Plur. Runoff	0	0	0	0

Replicability of agreement (off-diagonal) and replicability of existence (*diagonal*) of an unambiguous winner (generalized rank 1) in the upper panel, and an unambiguous unique loser (generalized rank 5) in the center panel. The bottom panel gives the replicability of agreement between a unique row winner and a unique column loser

agreement vary substantially across rules and across models. The panel in the center shows the corresponding results for movies with generalized rank 5 in the social order, i.e., movies that are ranked strictly worse than any other movie, in a given social order. Again the results are highly model dependent. Antiplurality, for which we have hardly any valid ballots, yields essentially useless results. For the other rules, using the Pessimistic and Anchor-and-Adjust model, we consistently have agreement in Movie Set 1 with perfect replicability. Note that this analysis does not apply to Plurality Runoff, which only yields a winner. The bottom panel in Table 5 shows how often we find the situation that is so highly advertised in textbooks on social choice: We search for an option that is the unique best option by one consensus method and yet the unique worst option by another consensus rule. The results are much more model-dependent than they have been in earlier papers. For Movie Set 1 and the Pessimistic model, not once in 10,000 bootstrap iterations did we see a movie have generalized rank one in one rule (row) and generalized rank 5 in another rule (column). The same applies for the other models and voting rules, except for Antiplurality, which completely hinges on whether a model generates many, some, or virtually no ballots with individual preferences that rank some option as unique worst. Table 5 highlights how little we can infer when we do not impute assumptions

about preferences where no information was given by the movie rater, but also how artificially we might inflate our confidence in conclusions drawn from data that have a high imputed component, like the Pessimistic model.

As we move to Movie Sets 2 and 3, reported in Tables 6 and 7, we find a similar picture: The winners, losers, and the relationship between consensus methods are highly dependent on the modeling assumptions that entered the analysis. Likewise, the bootstrap-based replicability highly depends on the modeling assumptions.

Like in previous empirical studies, we find that voting paradoxes do not appear to loom nearly as large as they are made to appear in the axiomatic and sampling literature. We do not find strong evidence that the best of one rule is the worst of another rule in any analysis that actually treats many raters as providing valid ballots. To understand how large the potential disagreements among voting methods really loom requires that we tackle inference and model dependence. Unfortunately, the behavioral analyses in Tables 4–7 produce much more dramatic and sobering findings than did previous empirical studies on political survey and election ballot data. Because the Netflix data, while being extensive, are so extraordinarily sparse, the challenges associated with inference and model dependence appear to be strongly amplified in these data. We also amplify that contemporary research may need to shift focus away from classical problems of voting paradoxes to more pressing challenges. Consistent with earlier behavioral social choice papers, the threat of no Condorcet winner and/or the threat of dramatic disagreements among competing consensus methods continue to be dwarfed by the much more real treat of inaccurate inference about social preferences as well as the threat of their strong dependence on modeling assumptions.

5 Conclusions and future directions

How can behavioral social choice interface with computational social choice? Imagine a sensitive computer system protected by elaborate cryptography. The security of this system might be called into question if an adversary learned that a high-level user exclusively employed family birthdays and pet names as passwords, unless the cryptographic protection somehow specifically planned for such structured behavior. Similarly, behavioral insights could have extensive implications to computational social choice, because the computational properties of consensus methods could be affected profoundly by behavioral regularities in voter behavior.

Specifically, we hope that scholars in the computational social choice community will continue to investigate how computational considerations in social choice are affected by the two main points we highlight in this paper:

1. Behaviorally accurate evaluation of social choice outcomes depends on effective **inference** from incomplete and possibly noisy or biased data.
2. Some social choice considerations can be profoundly **dependent on modeling assumptions** about the nature of individual preferences and how they are expressed in the ballots, ratings, or rankings that are being aggregated.

We believe that those concerns are almost self-evident, especially in the analyses we have reported. With data as incomplete and sparse as the Netflix data, accurate modeling and reliable inference pose both undeniable and formidable challenges.

Yet, the classical social choice literature has paid almost no attention to these concerns. Our analysis has shown that treating preferences as strict linear orders or strict weak orders may require researchers to impute vast amounts of information not provided by the voters or raters. The resulting conclusions about the consensus processes then often rest on computations that used more hypothetical than real data. As social choice scholars, we do not wish to emulate the drunkard who lost his keys in a dark parking lot and proceeded to search for them under a street light because it was brighter there. Experts in recommender systems have recently started to tackle similar challenges [22]. On the other hand, when making as few assumptions about individual preferences as possible, as we attempted in the Agnostic model, we may not even be able to draw inferences at all for some consensus methods because of data sparsity.

Behavioral social choice has put inference and model dependence at the forefront of its research paradigm, and hence, may provide some helpful guidance to scholars interested in behaviorally adequate computational social choice. Future developments in computational social choice may take into account that strategic interaction, manipulability, and computational complexity may be intertwined in complicated ways with inference and model dependence at various levels. Realistically, both individuals and collectivities who want to compute strategic choices and/or manipulate a consensus process need to account for inference and model dependence issues in their respective computations.

Acknowledgements Regenwetter and Popova acknowledge funding under National Science Foundation grant No. SES-08-20009 (to Regenwetter, PI) and grant No. CCF-1216016 (to Regenwetter, PI). Mattei acknowledges support by the National Science Foundation under Grant No. IIS-1107011 (to Judy Goldsmith, PI) and CCF-1049360 (to Judy Goldsmith, PI). Much of this work was carried out while Mattei was still a graduate student at the University of Kentucky and we acknowledge their support. NICTA is funded by the Australian Government through the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. We thank Sergey Popov for help and advice with programming and Netflix for releasing such valuable data. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the authors' universities.

References

1. Arrow, K.J.: Social Choice and Individual Values. Wiley, New York (1963)
2. Arrow, K.J., Sen, A.K., Suzumura, K. (eds.): Handbook of Social Choice and Welfare, vol. 1. North-Holland, Amsterdam (2002)
3. Bennett, J., Lanning, S.: The Netflix Prize. In: Proceedings of the KDD Cup and Workshop (2007). www.netflixprize.com
4. Black, D.: On the rationale of group decision-making. J. Polit. Econ. **56**(1), 23–34 (1948)
5. Chamberlin, J.R., Cohen, J.L., Coombs, C.H.: Social choice observed: five presidential elections of the American Psychological Association. J. Polit. **46**(2), 479–502 (1984)
6. Condorcet, M.: Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix (Essai on the Application of the Probabilistic Analysis of Majority Vote Decisions). Imprimerie Royale, Paris (1785)
7. Conitzer, V., Walsh, T., Xia, L.: Dominating manipulations in voting with partial information. In: Proceedings of the 25th American Association of Artificial Intelligence Conference (AAAI 2011), pp. 638–643 (2011)
8. Erdélyi, G., Fernau, H., Goldsmith, J., Mattei, N., Raible, D., Rothe, J.: The complexity of probabilistic lobbying. In: Proceedings of the 1st International Conference on Algorithmic Decision Theory (ADT 2009), pp. 86–97 (2009)

9. Faliszewski, P., Hemaspaandra, E., Hemaspaandra, L., Rothe, J.: The shield that never was: Societies with single-peaked preferences are more open to manipulation and control. Inf. Comput. **209**(2), 89–107 (2011)
10. Felsenthal, D.S., Maoz, Z., A, R.: An empirical evaluation of six voting procedures: do they really make any difference? Br. J. Polit. Sci. **23**, 1–27 (1993)
11. Gehrlein, W.V.: Concorcet's paradox. Theory Decis. **15**, 161—197 (1983)
12. Gehrlein, W.V.: Condorcet efficiency of constant scoring rules for large electorates. Econ. Lett. **19**, 13—15 (1985)
13. Gehrlein, W.V.: Condorcet efficiency of simple voting rules for large electorates. Econ. Lett. **40**, 61—66 (1992)
14. Gehrlein, W.V.: Condorcet's paradox and the likelihood of its occurrence: different perspectives on balanced preferences. Theory Decis. **52**(2), 171–199 (2002)
15. Gehrlein, W.V., Fishburn, P.C.: Concorcet's paradox and anonymous preference profiles. Public Choice **26**, 1—18 (1976)
16. Gehrlein, W.V., Fishburn, P.C.: The probability of the paradox of voting: a computable solution. J. Econ. Theory **13**, 14–25 (1976)
17. Gehrlein, W.V., Fishburn, P.C.: Coincidence probabilities for simple majority and positional voting rules. Soc. Sci. Res. **7**(3), 272–283 (1978)
18. Gehrlein, W.V., Lepelley, D.: The probability that all weighted scoring rules elect the same winner. Econ. Lett. **66**, 191–197 (2000)
19. Gibbard, A.: Manipulation of voting schemes: a general result. Econometrica **41**(4), 587–601 (1973)
20. Hazon, N., Aumann, Y., Kraus, S., Wooldridge, M.: On the evaluation of election outcomes under uncertainty. Artif. Intell. **189**, 1–18 (2012)
21. Mackie, G.: Democracy Defended. Cambridge University Press, New York (2003)
22. Marlin, B., Zemel, R.: Collaborative prediction and ranking with non-random missing data. In: Proceedings of the 3rd ACM Conference on Recommender Systems (2009)
23. Niemi, R.: The occurrence of the paradox of voting in university elections. Public Choice **8**(1), 91–100 (1970)
24. Nurmi, H.: Voting procedures: a summary analysis. Br. J. Polit. Sci. **13**(2), 181–208 (1983)
25. Regenwetter, M.: Perspectives on preference aggregation. Perspect. Psychol. Sci. **4**, 403–407 (2009)
26. Regenwetter, M., Grofman, B.: Approval voting, Borda winners and Condorcet winners: Evidence from seven elections. Manage. Sci. **44**, 520–533 (1998)
27. Regenwetter, M., Grofman, B.: Choosing subsets: a size-independent probabilistic model and the quest for a social welfare ordering. Soc. Choice Welf. **15**, 423–443 (1998)
28. Regenwetter, M., Rykhlevskaia, E.: On the (numerical) ranking associated with any finite binary relation. J. Math. Psychol. **48**, 239–246 (2004)
29. Regenwetter, M., Rykhlevskaia, E.: A general concept of scoring rules: general definitions, statistical inference, and empirical illustrations. Soc. Choice Welf. **29**, 211–228 (2007)
30. Regenwetter, M., Tsetlin, I.: Approval voting and positional voting methods: inference, relationship, examples. Soc. Choice Welf. **22**, 539–566 (2004)
31. Regenwetter, M., Adams, J., Grofman, B.: On the (sample) Condorcet efficiency of majority rule: an alternative view of majority cycles and social homogeneity. Theory Decis. **53**, 153–186 (2002)
32. Regenwetter, M., Grofman, B., Marley, A.A.J.: On the model dependence of majority preferences reconstructed from ballot or survey data. Math. Soc. Sci. (Special issue on random utility theory and probabilistic measurement theory) **43**, 453–468 (2002)
33. Regenwetter, M., Marley, A.A.J., Grofman, B.: A general concept of majority rule. Math. Soc. Sci. (Special issue on random utility theory and probabilistic measurement theory) **43**, 407–430 (2002)
34. Regenwetter, M., Marley, A.A.J., Grofman, B.: General concepts of value restriction and preference majority. Soc. Choice Welf. **21**, 149–173 (2003)
35. Regenwetter, M., Grofman, B., Marley, A.A.J., Tsetlin, I.M.: Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications. Cambridge University Press, New York (2006)
36. Regenwetter, M., Ho, M.H., Tsetlin, I.: Sophisticated approval voting, ignorance priors, and plurality heuristics: a behavioral social choice analysis in a Thurstonian framework. Psychol. Rev. **114**, 994–1014 (2007)

37. Regenwetter, M., Kim, A., Kantor, A., Ho, M.H.: The unexpected empirical consensus among consensus methods. Psychol. Sci. **18**, 559–656 (2007)
38. Regenwetter, M., Grofman, B., Popova, A., Messner, W., Davis-Stober, C.P., Cavagnaro, D.R.: Behavioural social choice: a status report. Philos. Trans. R. Soc. Lond., B Biol. Sci. **364**, 833–843 (2009)
39. Riker, W.H.: Liberalism v. Populism. W. H. Freeman, San Fransisco (1982)
40. Rivest, R.L., Shen, E.: An optimal single-winner preferential voting system based on game theory. In: Proceedings of the 3rd International Workshop on Computational Social Choice (COMSOC 2010), pp. 399–410 (2010)
41. Roberts, F.S.: Measurement Theory. Addison-Wesley, London (1979)
42. Roberts, F.: Applications of the theory of meaningfulness to psychology. J. Math. Psychol. **29**, 311–332 (1985)
43. Roberts, F.S.: Meaningless statements. In: Graham, R.L., Kratochvil, J., Nesetril, J., Roberts, F.S. (eds.) The Future of Discrete Mathematics. American Mathematical Society, Providence (1998)
44. Roberts, F., Rosenbaum, Z.: Scale type, meaningfulness and the possible psychophysical laws. Math. Soc. Sci. **12**, 77–95 (1986)
45. Saari, D.: Geometry of Voting. Springer, New York (1994)
46. Satterthwaite, M.: Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. J. Econ. Theory **10**(2), 187–216 (1975)
47. Sen, A.K.: A possibility theorem on majority decisions. Econometrica **34**(2), 491–499 (1966)
48. Shepsle, K., Bonchek, M.: Analyzing Politics. Norton, New York (1997)
49. Shoham, Y., Leyton-Brown, K.: Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press, New York (2009)
50. Tideman, N.: Collective Decisions and Voting: The Potential for Public Choice. Ashgate Publishing, Aldershot (2006)
51. Tideman, N., Plassmann, F.: Modeling the outcomes of vote-casting in actual elections. In: Felsenthal, D., Machover, M. (eds.) Electoral Systems: Paradoxes, Assumptions, and Procedures. Springer, New York (2012)
52. Tsetlin, I., Regenwetter, M.: On the probability of correct or incorrect majority preference relations. Soc. Choice Welf. **20**, 283–306 (2003)
53. Tsetlin, I., Regenwetter, M., Grofman, B.: The impartial culture maximizes the probability of majority cycles. Soc. Choice Welf. **21**, 387–398 (2003)
54. Walsh, T.: An empirical study of the manipulability of single transferable voting. In: Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010), pp. 257–262 (2010)
55. Xia, L., Conitzer, V.: Determining possible and necessary winners under common voting rules given partial orders. J. Artif. Intell. Res. **41**(2), 25–67 (2011)