INPUTS, ALGORITHMS, QUALITY MEASURES: MORE REALISTIC SIMULATION IN SOCIAL CHOICE

MARK C. WILSON

Abstract. Much of my research deals with trying to evaluate the performance of social choice *algorithms* via simulations, which requires appropriate *inputs* and *quality measures*. All three areas offer substantial scope for improvement in the coming years. For concreteness and because of my own limited experience, I focus on the allocation of indivisible goods and on voting, although many of the ideas are more broadly applicable.

Introduction

There are hugely many possible algorithmically defined rules for voting and allocation. Why then do we see so few of them in the literature? This phenomenon is not limited to social choice — in my experience, in most application areas the number of heavily studied and implemented algorithms is fairly small. For example, there are many ways to sort a list, but quicksort, mergesort and heapsort dominate the literature and practice. Of course, in the case of sorting, there is only one really interesting performance criterion, namely optimal (average or worst case) running time, and it is (asymptotically) achieved by all the algorithms listed, while naive sorting algorithms fail to achieve optimality. However in economic design, we are typically faced with multiple competing success criteria, such as strategyproofness, efficiency, fairness, and welfare, which intuitively should lead to more algorithms that are considered viable.

Axiomatic methods are very common in economic design. An axiom is essentially a statement of the form "if the input satisfies property P, then the output satisfies property Q". For example, for the unanimity axiom for voting, P might be "all voters have the same top choice" and Q might be "the common top choice is the winner of the election". A rule satisfying an axiom leaves no room for confusion, or concerns about input distributions, since we are dealing with logic and not probability. The fact remains, however, that there are many axioms, and they usually conflict with each other. This leads often to impossibility results if too many axioms are imposed, occasionally to characterization results if we impose exactly the right number, and frequently to tradeoffs that must be investigated when we impose even fewer. These tradeoffs necessarily use notions of probability, such as how often P implies Q, or optimization, such as how close the outcome is to an outcome in which Q holds.

For a given collection of axioms, it is unnecessarily limiting to consider only the extreme cases, namely the algorithms that satisfy at least one of the axioms. Rules satisfying some axioms often fail to satisfy others, which is not surprising. What is more surprising is how badly an "extremal" algorithm, which satisfies axiom A, fails to satisfy axiom B. For example, Serial Dictatorship is strategyproof and efficient for allocation, but overall performs relatively poorly on fairness and welfare criteria. Similarly, throwing away all items and allocating none is a fair and strategyproof

Date: August 5, 2018.

Thanks to Klaus Nehring for his feedback on an earlier draft.

procedure, but very inefficient. A large number of papers start with one of the extremal algorithms and investigate how badly it performs with respect to other axiomatic properties. We should at least aim to explore the "Pareto frontier of the space of rules" consisting of rules not dominated by any other rule. The rule we need for a given situation may fail to satisfy any of the axioms in general, but may only just fail to satisfy all of them simultaneously, while extremal rules may score 100% on one axiom but close to zero on another. I believe that failure to appreciate this fact is a major reason for the small number of algorithms in the literature.

As an aside, it is important to choose the right criteria and set of algorithms for our analysis. For example, a large number of papers have been written about minimizing the manipulability of voting rules within various classes of rules, for example positional scoring rules. But we already know that dictatorial voting rules minimize manipulability absolutely since they are strategyproof. Since dictatorial rules are considered bad, presumably, because of other axioms relating to fairness and welfare performance, any such studies should surely consider those two criteria and their relative importance. Otherwise, we cannot *a priori* exclude the possibility that a dictatorial rule performs overall worse than the rules in the class under study. Surprisingly, a large number of papers have not taken this into consideration. I published a few papers along these lines before realizing this.

Exploring the Pareto frontier of rules leads us immediately to issues of measurement. If an algorithm does not satisfy a certain axiom, then we need to measure how close it comes to doing so, and there are many more ways to do this than the literature would indicate. A general setup is to have measures $\mu_1, \dots, \mu_k$, one for each axiom, each taking values between 0 and 1, where 1 indicates that the axiom holds for that input. For example, we might measure the fraction of agents who envy another agent under the algorithm's allocation for the given input, or simply have the indicator function which is 1 if the algorithm is strategyproof on the given input and 0 otherwise. Ideally for each input we will represent each algorithm by a k-tuple of values of the measures. An algorithm A_1 is dominated by A_2 if and only if $\mu_i(A_1) \leq \mu_i(A_2)$ for all i and there is some i for which the inequality is strict.

This is too weak a partial order to impose on our algorithms. For example, no extremal algorithm can be dominated on any input by any other algorithm, since one of the measures has value 1. Thus we may need to give weights to the various measures and consider only the weighted sum (this includes the case of weight zero, where we exclude a measure completely). But even if we do not, we can still search for rules on the frontier that are not dominated. The main problem is that the dominance relation is defined for each input separately, which is too detailed for most purposes. We typically need to relax this by considering a distribution over all inputs. But this introduces statistical notions - how to summarize all this information about distributions? The most obvious measures are the maximum, minimum, mean and median. Worst case comparisons don't tell us much — quicksort is worse than insertion sort in the worst case, but it is still much more used in practice because its expected running time is better. Best case performance is usually perfect for all algorithms. So the mean or median make more sense.

Of course, these statistics depend substantially on the input distribution. What kind of data will our algorithm be faced with? If inputs are chosen uniformly at random from all linear orders of n distinct elements, then quicksort has running time in $\Theta(n \log n)$. However its worst case is quadratic. In the case of allocation rules, the Impartial Culture (where each agent independently chooses a linear preference order as above) is the easiest case. The lack of correlation between agents means that many agents have different first preferences, and even if they coincide on first preference, they have different second preferences. This makes allocation very easy (note that for voting, this is the hardest case, because every order of candidates has approximately equal support and only random variability of order $\sqrt{n}$ prevents a complete tie). Impartial Culture has value as an extreme case and is mathematically tractable, but is well known to be very unrealistic [7].

Having pointed out some issues that I feel have not been adequately addressed in the literature so far, in the following sections I discuss some ideas for improvement in methodology. I assume that some kind of simulation or mathematical analysis will be used in order to compare algorithm performance. I consider the following setup. We generate several input datasets according to various distributional assumptions. Each is analysed separately. For each algorithm in our set, we compute the value of each measure μ_i when running the algorithm on the given input. These are aggregated to find the expected value with respect to the given distribution. These values $\mu_i^* := E[\mu_i]$ are taken as the coordinates in k-space and the algorithms compared using dominance as above. Results will have the form "Serial Dictatorship is dominated by the Boston mechanism with respect to efficiency and envy-freeness for 10 of 12 datasets."

There are other issues I have not considered here which will need to be dealt with. For example, comparing means may be reasonable in some cases, but if the distributions of a given measure for outputs of two different algorithms using the same input distribution overlap substantially and have large variance, we may need a more refined analysis.

Inputs

If we are happy to use simulation rather than proving analytic results, we can use any input distribution we like. How to stress-test our algorithm by choosing interesting and "realistic" data? For example, evaluating and comparing voting rules using only the Impartial Culture, although it has been done often, is not sufficient.

The Mallows model has been used by several authors, based on the idea of there being an underlying true ranking of alternatives ρ and the probability of an agent having the preference π being proportional to $\exp(-cd(\pi, \rho))$, for some $c > 0$, where d is the Kendall tau (swap) distance. This is typically applied with independence between agents.

To get correlation between agents, we can use an urn process in which we think of a new agent as copying a randomly chosen agent. This can lead to the Impartial Anonymous Culture, which is analytically very tractable and less unrealistic than Impartial Culture, but still far from describing reality. Different values of the parameter associated with this Eggenberger-Pólya urn (we may add more than one agent at a time) lead to more general distributions which may be more realistic. I have not seen serious work on fitting such distributions to real data. Of course, obtaining real data on preferences is difficult, increasingly so as has become clear from the failure of recent electoral predictions worldwide.

Some analyses require more subtlety. For example, with a coauthor I have recently [6] investigated the performance of electoral systems. Most electoral systems involve geographic districts, and treating them as independent or merging them into a single district are both oversimplifications. We used a coupled urn model with one urn for each district, allowing for imitation both within and across districts. To my knowledge this is the most sophisticated model used in the area so far. Surely there are better ones.

We have been modelling true preferences so far. In order to understand real performance of algorithms, it would be useful to have good models of strategic behavior, so that the agents' expressed preferences can be used as inputs but the real ones used for quality measures. For simple voting rules such as plurality it is easy to model behaviour of voters. Of course, the more complicated the rule the less likely manipulation should be in practice, because of agent fears about other agents' strategic voting and simply because of the complexity of computing a strategic preference.

I hope to see a greater variety of input distributions used in the future literature.

Quality measures

At the most basic level, we can take μ to be the indicator for any axiom, so that μ_* is the probability that an input leads to an outcome satisfying the axiom. This has been widely used and is very simple.

More sophisticated measures may consider the number of agents directly affected by failure of the axiom to hold. For example, the fraction of ordered pairs (i, j) of agents for which i does not envy j's allocation (that is, prefers it to the allocation given to i) measures partial envy-freeness. The fraction of agents for whom truthtelling is a dominant strategy is a measure of resistance to individual manipulation.

The next level involves measurements based on preference intensity. For example, if I have my 2nd choice and you have my 1st choice, I may envy you less than if I have my 3rd choice and you have my 1st choice. As another example, a single manipulator may find it harder to change its expressed preference if that is very different from its true preference (for example, because the latter might be partially known to other agents, and preferences may not be given secretly, or the agent may require a bigger bribe in order to submit a vote far from its sincere one).

Measures based on cost of forming coalitions are important. For example, the communication overhead of organizing a coalition to manipulate a voting rule, or trade among themselves to restore an efficient outcome, may grow rapidly with coalition size.

Implicit in the measures so far is an idea of distance from the outcome to an axiomatically perfect one. For example, in the manipulation context above, a cost function as in the Mallows model may be appropriate. There are many other possible metrics, however. As another example, consider the directed graph G formed by agents, in which each agent i points to the agent holding i's favorite item. Gale's Top Trading Cycle (TTC) algorithm reallocates items according to cycles formed in G, passing to items ranked 2nd, 3rd, etc and deleting agents and items as it goes. The algorithm terminates precisely when the digraph has only trivial cycles (each agent points to itself), and that occurs if and only if the allocation is efficient. A sophisticated measure of efficiency, then, might use a metric that measures the distance from G to the nearest acyclic digraph. Such a metric might take into account the number and length of cycles in the original digraph.

So far all the measures have involved finding the minimum distance (or cost) to achieving axiomatic perfection for a given input. We can also consider probabilistic measures, which may be more realistic. In practice, it may be hard to find a minimum manipulating coalition, and the prevalence of manipulators should be considered. A voting rule that can be manipulated for a given input by 5 agents, never by 4 or fewer, and by very few coalitions of size 6, may be less manipulable in practice than one for which all coaltiions of size 6 can manipulate but no smaller coalition can. A general idea is to consider the simple game on the set of agents, with a winning coalition being a subset containing a manipulating coalition. An index such as the Coleman index (or generalizations discussed in [5]) gives a measure of how likely a randomly chosen coalition is to be winning.

A new fairness idea (to my knowledge) is having low *order bias*. Almost all deterministic allocation algorithms require a fixed order on agents. For example, Serial Dictatorship has this built into the definition, while the Boston mechanism requires ties to be broken. A completely fair algorithm would be not only fair *ex ante* (by the usual randomization of the agent order) but also *ex post* — my welfare should not suffer just because I am chosen to submit my preferences last. Serial Dictatorship is clearly extremely unfair — the expected rank of the object chosen by the last agent is much higher than 1, the rank achieved by the first player. Boston similarly is unfair to later players. Randomization during the algorithm (after player order has been chosen) might deal with this problem, but if we insist on deterministic algorithms, what can we do (see "Algorithms" below)?

In general there are so many possible measures that we ought to try to justify the ones we use. Most of the literature I have seen simply uses a naive measure or copies the choice of a previous author in the field. A bigger intellectual contribution can be made by giving axiomatic properties that we wish a measure to hold. Spurred on by a perceptive referee, my coauthors and I did this for manipulability of voting rules [5], and now I try to apply this general advice in all my research. Every time we want to measure "partial X", where X is an axiom, we should think about the axiomatic foundations of our measure. I hope to see much better justification of measures in the future literature.

Algorithms

In addition to the design criteria for algorithms such as computational efficiency and solution quality, simplicity of explanation to the public is also a consideration if we want our algorithm to be adopted. There are limits to this — Single Transferable Vote is unlikely to be completely understood by the public yet has been used in Australia (which, perhaps not coincidentally, has compulsory voting) for decades. Nevertheless, a coherent "story" is often helpful.

Inspired by a Christmas party game, with a coauthor I recently introduced the Yankee Swap algorithm for allocation of indivisible goods [4]. Again spurred by a perceptive reviewer complaining of how arbitrary this seemed, we generalized this to a family of 8 algorithms, 4 of which can be profitably followed by TTC to yield 12 algorithms in total. Two of these are equivalent to our old friends Serial Dictatorship and Boston. These algorithms are all derived from the Gale-Shapley algorithm for two-sided matching by assigning fictitious preferences for items over agents, allowing these to change during the algorithm, and adopting a queue or stack discipline for agents to propose to items. The point is that a single standard algorithm for an allocation problem gave rise to 12 algorithms for a closely related problem in a coherent way. All of these algorithms have an easy interpretation in terms of a party game involving stealing gifts. Some of them appear to make good tradeoffs although none satisfies any of the standard axioms, and very likely none would have been found by concentrating only on the standard algorithms. It is still too early to tell whether these algorithms will attract attention. One of them has good welfare properties, and another has extremely low order bias even though it is deterministic.

In voting theory, there may be many more good algorithms to be found. To mention one among several recently introduced rules, Zwicker and coauthors have introduced the *mediancenter* rule [1] which is still little explored. The general technique of distance rationalization has been used to construct new rules with given axiomatic properties, based on given notions of consensus and distance between profiles [2, 3]. I am sure that appropriate choices of consensus and distance can yield some new interesting rules, which may have good overall performance even if they fail to be extremal in the sense discussed above.

Conclusion

I hope to have convinced the reader that with high probability, there are many interesting new algorithms still to be found, and researchers should be on the lookout for them, sometimes hidden in plain sight in ordinary life. This should lead in many situations to better rules being used than those currently under consideration. Much more sophisticated and better motivated measures of performance, and more thorough analysis with respect to more realistic input distributions, will be needed if these algorithms are to be fairly compared with the standard ones in the literature.

References

[1] Davide P. Cervone, Ronghua Dai, Daniel Gnoutcheff, Grant Lanterman, Andrew Mackenzie, Ari Morse, Nikhil Srivastava, and William S. Zwicker. "Voting with rubber bands, weights, and strings". *Mathematical Social Sciences* 64. (2012), 11–27.

[2] Edith Elkind, Piotr Faliszewski, and Arkadii Slinko. "Distance rationalization of voting rules". *Social Choice and Welfare* 45. (2015), 345–377.

[3] Benjamin Hadjibeyli and Mark C Wilson. "Distance Rationalization of Anonymous and Homogeneous Voting Rules". *Social Choice and Welfare* (to appear, 2018+).

[4] Jacky Lo and Mark C Wilson. "New algorithms for matching problems". *arXiv:1703.04225* (2017), 19pp.

[5] Geoffrey Pritchard, Reyhaneh Reyhani, and Mark C. Wilson. "Power measures derived from the sequential query process". *Math. Social Sci.* 65. (2013), 174–180.

[6] Geoffrey Pritchard and Mark C Wilson. "Multi-district preference modelling". *10.31235/osf.io/xpb8w* (2018).

[7] Michel Regenwetter, Bernard Grofman, A. A. J. Marley, and Ilia Tsetlin. *Behavioral social choice*. Probabilistic models, statistical inference, and applications. Cambridge: Cambridge University Press, 2006, pp. xvi+240.