DECOMPOSITION BEHAVIOR IN AGGREGATED DATA SETS

SARAH BERUBE AND KARL-DIETER CRISMAN

Abstract. One kind of aggregation of nonparametric data sets involves combining two (or more) sets with the same structure, but with no interweaving of values. For instance, one could have prices of the same product at different stores on different days, but before and after a significant sales tax hike. In this paper, looking at the case of three 'candidates', we give a precise computation of how decomposition of the underlying voting data of these sets is related to the decomposition of the original set. Among other things, this proves it is possible to obtain data sets with arbitrarily large Basic components relative to their size. A second result gives congruence and other criteria for possible values of the sizes of the various components, and we exhibit pure Basic data sets (discovered using the computer program Sage) which exemplify those criteria.

1. Introduction

Aggregation of data is a profound source of paradox in statistics; this is particularly true in nonparametric samples. Perhaps one of the most famous real-life examples of these is [6], which involved the Yule-Simpson paradox; indeed, "Simpson's paradox" is often referred to in statistical literacy, popular, and a wide variety of other contexts (e.g., [27], [14], [13], even the unusual connection in [25]!). Attempts are also often made to unravel this in various ways (e.g., [9], [24], [15]). However, Arrow's seminal theorem ([1]) naturally implies *any* procedure concerned with ranking should manifest similar paradoxes, and one might want to explain or study these as well.

In particular, in the last two decades, tools from the mathematics of voting have been used to begin to unravel some of them, as well as to the discover their depth. In [18],[10], and [12], Haunsperger and Saari investigated the universality of paradoxes in nonparametric methods. More recently, aggregation paradoxes have been investigated using the same tools. In fact, Haunsperger ([11]) shows that nearly all data sets are inconsistent under aggregation for Kruskal-Wallis ([16]), even if the matrices of ranks for the sets are identical; Bargagliotti ([4]) extends this to several other tests.

In fact, it turns out that only in the highly unrealistic case that the data is equivalent to a data set with all observations yielding the same ranking can one avoid replication paradoxes (not to mention more general aggregation problems)! On the positive side, more recent work of Bargagliotti and Greenwell ([2]) indicate that the *statistical* significance of Haunsperger-style paradoxes is negligible, though nothing is known about less strict (and perhaps more realistic) aggregations.

In any case, one can approach the problem from the other direction. Many nonparametric procedures[1] are equivalent to ones which recover their results by first creating a voting profile, to which a standard points or pairwise procedure is

[1]Including Kruskal-Wallis, Mann-Whitney-Wilcoxon, Bhapkar V; see for instance [4] or [10].

applied. Thus, it is reasonable to look at the underlying profiles instead of the data sets per se, to see for which of them paradoxes might be minimized, since they cannot be avoided. In the context of voting theory, Saari (for instance, see [19, 20, 21, 22, 23]) has used symmetry-respecting decompositions of the profile vector space to determine what components of a voting profile may be viewed as 'causing' various types of paradoxes (see also Orrison et al. in [8] for a representation-theoretic view of this).

This philosophy is applied in Bargagliotti and Saari ([3]) to provide new proofs of several of Haunsperger's results, as well as to give explicit criteria for recognizing when a given data set will avoid various paradoxes for different types of procedures (in terms of the decompositions of the underlying voting profiles of the data sets). For instance, there is a criterion given for recognizing when Kruskal-Wallis appears to indicate that the null hypothesis is confirmed, but most other tests yield various mutually incompatible results (again, naively doing this before looking at the actual test statistic).

The component of any decomposition which yields the fewest paradoxes is called the Basic component[2] of the underlying profile. Even though real-life data sets will have all sorts of decompositions, it is still useful to understand as much as we can about the situation with the greatest consistency, so we raise the following natural questions:

Questions. Regarding the Basic component:

- How does it behave under aggregation, or at least under replication?
- How close can we come to a data set with no other components?
- How might one recognize such a data set?

In this paper, to begin answering these questions, we look at a special form of replication where k data sets are aggregated, all of which have the same matrix of ranks *and* which have no overlap of their underlying data. For a picture, one may think of this as *stacking* matrices of ranks.

For the first question, as it turns out, one can completely calculate how *all* components change under such a replication, and we do so below for the case of $m = 3$ choices. We then answer the second question completely; starting with an appropriate data set with a large Basic component, by stacking it we can approach an arbitrary relative size of Basic components. In fact, we can also exhibit pure Basic data sets, discovered by computer search. Finally, we give several necessary - but by no means sufficient! - criteria for the third question, as corollaries of more general facts about the sizes of the components of the decomposition of the profile of any data set.

2. Definitions

For convenience, we treat the items being compared by the test in the same way as in the voting literature, i.e. $A, B, C \ldots$ Except where we explicitly say otherwise, the number of items is given by $m = 3$.

A *data set* is the raw data itself, usually organized in the form of a matrix, so:

A	B	C
14.5	15.6	16.7
14.3	11.2	13.4

[2]Or, for $n \geq 4$, the Borda component.

However, for nonparametric statistics, the point is to ignore the actual data, and focus only on the relative information involved. We will only look at data sets with equal numbers of entries in each column, so having n rows; we call the *matrix of ranks* the matrix obtained by substituting the numbers $1, 2, \cdots, 3n$ for the actual data points. In this example, we get

A	B	C
4	5	6
3	1	2

or, for convenience, $\begin{pmatrix} 4 & 5 & 6 \\ 3 & 1 & 2 \end{pmatrix}$

To connect this to voting, we create a *profile*, with accompanying *profile vector*, to encapsulate voting data. Namely, we look at all possible triplets of ranks (one rank for each item) and, for each of these triplets, return the ranking of the items corresponding to that. In this example, we can see that (4 1 2) would correspond to $A \succ C \succ B$, while (4 1 6) gives $C \succ A \succ B$, and so on. The collection of all such rankings is the profile, and the profile vector is the vector obtained by placing these in the now-canonical order

$$A \succ B \succ C, A \succ C \succ B, C \succ A \succ B, C \succ B \succ A, B \succ C \succ A, B \succ A \succ C\,,$$

yielding $(0, 2, 2, 2, 0, 2)$ for our example.

As with any situation involving vectors, one looks for meaningful decompositions of the vector space[3] involved. In Saari and Bargagliotti, the standard decomposition from [20] is used; we refer the interested reader there for proofs of the assertions.

- The Basic components, $B_A = (1, 1, 0, -1, -1, 0)$, $B_B = (0, -1, -1, 0, 1, 1)$, and $B_C = (-1, 0, 1, 1, 0, -1)$, are the components which are unaffected by pairwise or tallying-procedure paradoxes. Note that these are *not* mutually orthogonal, and indeed $B_A + B_B + B_C$ is the zero vector.
- The Reversal components $R_A = (1, 1, -2, 1, 1, -2)$, $R_B = (-2, 1, 1, -2, 1, 1)$, and $R_C = (1, -2, 1, 1, -2, 1)$ yield perfect ties for all pairwise comparisons, but give different results for different tallying procedures. Note that they have the same algebraic structure as the Basic profiles.
- The Condorcet component $C = (1, -1, 1, -1, 1, -1)$ yields a perfect tie for all tallying procedures (such as plurality or the Borda Count), but give a paradoxical cycle $A \succ B \succ C \succ A$ for pairwise votes.
- The Kernel component $K = (1, 1, 1, 1, 1, 1)$ measures the number of voters (more specifically, one-sixth the number of voters).

Note that all but the Kernel component are actually *profile differentials* which sum to zero total voters (and hence are orthogonal to the Kernel); this does not affect our results, but rather is a direct result of the symmetry involved. Likewise, each sub-vector space is orthogonal to the others, so this is a true decomposition, and one can check they are simple Σ_3-modules (see, for instance, [8]), so this is an irreducible decomposition with respect to switching the names of the candidates. We easily obtain the decomposition of a profile vector by (left-)multiplying it by the inverse of the matrix whose rows are the vectors above (excepting B_C and R_C, as otherwise they are not linearly independent). In this case, we obtain coefficients

[3]Here, we are nowhere near a vector space yet, but considering the examples as part of $\mathbb{Q}^6$ or even $\mathbb{R}^6$ is standard for voting theory, and all that is important for us is that the vectors coming from data sets live in that vector space and inherit its symmetry - not the numbers themselves.

as follows:

$$\begin{pmatrix} 4 & 5 & 6 \\ 3 & 1 & 2 \end{pmatrix} \Rightarrow (0, 2, 2, 2, 0, 2) \Rightarrow (-1/3, -2/3, -1/3, 0, -2/3, 4/3)$$

Using the basis above, we can write $\frac{1}{3}(-B_A - 2B_B - R_A - 2C + 4K)$, which corresponds to a Kruskal-Wallis outcome of $C \succ A \succ B$, since the B_C component would be the only positive one (using the formula above, for instance). Other procedures could be influenced by the other components; for example, the voting procedure corresponding to Bhapkar's V test ([5]) (plurality) would yield $C \succ A, B$ but would not distinguish between A and B.

Finally, we recall Haunsperger's aggregation definitions (see [11]). For a given statistical procedure whose outcome is ranking (possibly with ties) of the candidates, and for all matrices of ranks:

- The procedure is *consistent under aggregation* if any aggregate of k sets of data, all of which yield a given ordering of the candidates, also yields the same ordering.
- The procedure is *consistent under replication* if any aggregate of k sets of data, all of which have the same matrix of ranks, yields the same ordering as any individual data set.

Our concern is with a specific form of replication, which we call *stacking*, which is the aggregate of k data sets, all of which have the same matrix of ranks, and which in addition do not have any overlap between the numerical ranges of their data. Data like this may arise in situations where one expects no change in observations relative to one another, but significant change as a whole. Examples might include prices before and after some large external event (like a tax increase or wholesale price change), timings of several algorithms before and after a hardware upgrade, or animal populations before and after a conservation effort.

In terms of the matrices of ranks, this means the aggregate matrix looks literally like a stack of k matrices with values

$$\overbrace{1, 2, \ldots, 3n}^{1}\ \overbrace{3n+1, 3n+2, \ldots, 6n}^{2}\ \ldots \overbrace{3(k-1)n+1, 3(k-1)n+2, \ldots 3kn}^{k}$$

Stacking our original example with $k = 3$ yields the following matrix of ranks:

$$\left(\begin{array}{ccc} 16 & 17 & 18 \\ 15 & 13 & 14 \\ \hline 10 & 11 & 12 \\ 9 & 7 & 8 \\ \hline 4 & 5 & 6 \\ 3 & 1 & 2 \end{array}\right)$$

Each part of the matrix corresponding to the original matrix of ranks we will call a *stanza*, and we will typically delineate the stanzas, as we have done here.

3. Decomposing Stacks of Ranks

Our first result answers the first question from the introduction - how do the components behave under stacking? We have a complete characterization.

Theorem 1. *If we stack an $n \times 3$ matrix of ranks k times, each Basic component is multiplied by k^2, each Reversal component is multiplied by k, the Condorcet component is multiplied by k^2, and the Kernel component is multiplied by k^3.*

Since the Kernel component is ignored by *all* our nonparametric procedures (because it yields a complete tie for all the underlying voting procedures), the implication of the theorem is that as long as you start with a Condorcet component smaller than the Basic components, stacking is a good way to find data sets with very large Basic components (and hence great regularity in outcome with respect to a variety of procedures).

Recall that we call each part of the kn-row matrix of ranks that corresponds to the original matrix of ranks a *stanza* of the new matrix of ranks. For a general $p \times 3$ matrix of ranks, each of the p^3 contributions to the corresponding voting profile may be called a *triplet*. This observation makes the statement about the kernel trivial to prove. Namely, if we have n rows, there are n^3 triplets, and so the Kernel (by definition) is $n^3/6$; but do the same thing for nk rows and we get a kernel of size $k^3(n^3/6)$.

The proof of the rest of the theorem will proceed by separating out two different types of triplets, then combining their results.

Lemma 2. *All triplets that are formed from elements taken from three different stanzas add only kernel components to the resulting profile decomposition.*

(In fact, for $m > 3$ 'candidates', all m-tuplets formed from elements taken from m different stanzas add only kernel components.)

Lemma 3. *For a stacking with $k = 2$, each Basic component is quadrupled, each Reversal component is doubled, and the Condorcet component is quadrupled.*

Proof of Theorem 1. By stacking the data set k times, we have k stanzas in this new data set. If we were to just consider the decomposition of each individual stanza, we have k times the original components. For instance, for $k = 2$, we would have twice each of the original components. So the actual content of Lemma 3 is that there is no additional Reversal obtained from stacking with $k = 2$, but that we obtain *additional* Basic and Condorcet to that expected - twice as much, in fact.

Now, to the k expected, we must add the number of each component that is formed from interactions between stanzas. By Lemma 2, we know that any triplets that are formed by taking rankings from three different stanzas add only kernel components to the resulting profile decomposition, so we only need to be concerned with rankings that come from two different stanzas.

From Lemma 3, we know that any set of two stanzas in this data set will produce twice the original Basic and Condorcet components from their interactions. In this k-stanza data set, there are $\binom{k}{2}$ ways to choose a subset of two of the stanzas. Therefore, interactions between stanzas give us

$$2\binom{k}{2} = 2\frac{k!}{2!(k-2)!} = \frac{k!}{(k-2)!} = k(k-1) = k^2 - k$$

additional components (Basic or Condorcet, respectively). Adding these to the components obtained from the decomposition of individual stanzas, this gives $k^2 - k + k = k^2$, as desired.

On the other hand, Lemmas 2 and 3 both argue that we obtain no additional Reversal components beyond those from k separate stanzas; but this is exactly the statement of the theorem. □

4. Seeking Pure Basic Data Sets

In some sense it is a misnomer to talk about pure Basic data sets, since any voting profile associated to a data set has a Kernel component as delineated above. Nonetheless, we call any voting profile with only Kernel and Basic non-vanishing components *pure Basic* (and likewise for pure Reversal, pure Condorcet). Indeed, in the theory of voting, it turns out that pure Basic profiles may be viewed as providing the least amount of paradox (see for instance [8, 22, 23]), a particularly nice subspace. Hence it is reasonable to search for profiles with as large a (relative) Basic portion as possible, or even pure Basic, in the related context of nonparametric data sets. This is particularly interesting since the space of profiles arising from such sets is much smaller than the space of all possible profiles.

Our first result in this regard is the simple statement that stacking can yield matrices of ranks with as large a Basic component as one desires. First, recall that the B_A and B_B basis vectors (as well as the corresponding Reversal basis vectors) are *not* orthogonal, so one cannot simply use the coefficients in the decomposition by themselves. However, we *can* calculate the length of the total vector in 2-space as $\sqrt{B_A^2 + B_B^2 - B_A B_B}$[4], where we consider B_A to have unit length. For numerical results, one might object at this point that the Reversal and Condorcet basis vectors are not the same length, but we will only be making qualitative statements.

To be precise, let's apply the theorem to the Basic component. Stacking any matrix of ranks k times will cause these components to grow as k^2, making the size of this component grow as

$$\frac{\sqrt{(k^2 B_A)^2 + (k^2 B_B)^2 - (k^2 B_A)(k^2 B_B)}}{\sqrt{B_A^2 + B_B^2 - B_A B_B}} = k^2 ;$$

similar reasoning shows that the Reversal component only grows as k, and the Condorcet component obviously grows as k^2. Hence the Basic and Condorcet pieces will drown out the Reversal as k gets large, and if there is no Condorcet component, the Basic will be an arbitrarily large part of the total (since $\lim_{k\to\infty} \frac{kR}{k^2 B} = 0$ for any R, B).

However, one might want more. Our second result is purely computational - that, contrary to the suggestion in [3], it turns out there do exist pure Basic matrices.

Let $|XYZ|$ be the number of voters in a voting profile which prefer $X \succ Y \succ Z$ [5]. Implicit in [3] are the second halves of the following two propositions, the first halves of which appear there.

Proposition 4. *A profile contains no reversal terms if and only if* $|ABC| + |CBA| = |ACB| + |BCA| = |CAB| + |BAC|$. *Hence if such a profile comes from a data set, we must have* n^3 *divisible by 3, and so* $3 \mid n$.

Proposition 5. *A data set has no Condorcet components if and only if* $|ABC| + |CAB| + |BCA| = |ACB| + |CBA| + |BAC|$. *Hence if such a profile comes from a data set, we must have* n^3 *divisible by 2, and so* $2 \mid n$.

[4] This is easy to derive by realizing B_A and B_B may be viewed as unit vectors at $2\pi/3$ radians from each other.

[5] This notation will be used slightly differently, for convenience, in the proofs for the previous section.

In particular, note that combining the two conditions (which is equivalent to being pure Basic) means that the number of observations (rows in the matrix of ranks) must be divisible by six, so we have a first criterion for pure Basic sets.

One can do a direct computation of how many there are for $n = 6$. We implemented this using the Permutation class in the open source mathematics software system Sage (see [26]). See [7] for some more details, especially why we chose to use Sage to implement this computation. Of the total of 2,858,856 total data sets for $n = 6$ with the top row of the matrix of ranks giving $A \succ B \succ C$ and with all columns of the same in decreasing order (which is sort of a normal form for nonparametric data sets), only 1,334 were Basic, or about 4.666% of one percent! Since the number of computations involved for for the next possible case, $n = 12$, is about one and a half billion times as great as for $n = 6$, we are skeptical of the computational feasibility of more brute-force searches. However, see the next section for preliminary results for recognizing and characterizing pure Basic data sets.

5. Characterization of Pure Basics

We are nowhere near a full characterization of pure Basic data sets, not even at the level of the characterizations of pure Condorcet, Reversal, and Kernel voting profiles arising from nonparametric data sets found in [3]. Nonetheless, we present here several interesting and non-trivial necessary conditions for such sets. Along the way, we are able to make some general statements about the possible sizes of *all* the components in a profile coming from the matrix of ranks of a nonparametric data set.

In order to begin characterizing pure basic data sets, we return to Propositions 4 and 5. Let Let $|ABC| = a$, $|ACB| = b$, $|CAB| = c$, $|CBA| = d$, $|BCA| = e$, and $|BAC| = f$. The linear system we obtain from the propositions, which is $a + d = b + e = c + f$ and $a + c + e = b + d + f$, reduces to

$$a = d - 2e + 2f, b = 2d - 3e + 2f, c = 2d - 2e + f\,.$$

By symmetry of the propositions, we immediately know that if we know any three in a row of the cycle $abcdefa\ldots$, we can obtain the entire profile (given that it is a pure Basic profile). This is surprising enough - we have a very easy computation to perform to exclude potential Basic profiles! However, more is true.

Theorem 6. *If any three entries in a pure Basic profile vector are known, or if we know two entries which do not correspond to opposite rankings,*[6] *it is possible to find the remaining entries.*

This is plausible, because of the symmetry. However, the next result is quite surprising, and characterizes Basics in a very different way.

Theorem 7. *If $n = 6\ell$ is the size of the data set and the data set is pure Basic, then all entries in the underlying profile vector are divisible by 3ℓ.*

For instance, all profile entries from a pure Basic data set with six observations are divisible by three. This is the first result we know of along these lines which relies in a fundamental way upon the profile arising from a nonparametric data set.

[6]Such as $A \succ B \succ C$ and $C \succ B \succ A$, which correspond to a and d.

The new concept we need to prove this is that of a *transposition* or *swap* of two elements (i, j) of a matrix of ranks. This is simply a switch of these ranks between two matrices of ranks. The following shows a $(5, 2)$ transposition:

$$\begin{pmatrix} 6 & 5 & 4 \\ 1 & 3 & 2 \end{pmatrix} \text{ becomes } \begin{pmatrix} 6 & 3 & 5 \\ 1 & 2 & 4 \end{pmatrix}.$$

Note that in the default column ordering, some elements may rise or fall in such a swap. We will concern ourselves in this proof only with neighbor transpositions $(i, i-1)$, which are special because they cannot cause such reordering.

It is well-known that the set of all neighbor transpositions generates the symmetric group, viewed as a permutation group, and likewise the set of all neighbor swaps from a given matrix of ranks will generate *all* possible matrices of ranks for a given shape $n \times 3$. In particular, we can begin with the canonical matrix of ranks

$$\begin{pmatrix} 3n & 2n & n \\ 3n-1 & 2n-1 & n-1 \\ \vdots & \vdots & \vdots \\ 6 & 5 & 4 \\ 3 & 2 & 1 \end{pmatrix}$$

and move from there to any other matrix of ranks using neighbor transpositions. (It is okay that some nontrivial neighbor transpositions will have trivial effect on the matrix of ranks. Note also that this matrix of ranks is a counterexample to this theorem being sufficient as well as necessary.)

This is important because we know the voting profile and decomposition of this data set! They are $(n^3, 0, 0, 0, 0, 0)$ and $\frac{n^3}{6}(B_A - B_C - R_B + C + K)$. In fact, since we know from Proposition 5 that n is even, we can let $n = 2k$ and get decomposition $\frac{4k^3}{3}(B_A - B_C - R_B + C + K)$.

Lemma 8. *Any neighbor transposition $(i, i-1)$ between the columns for candidates Y and Z (respectively) changes the Condorcet component by $\pm\frac{2k}{3}$, the Basic component by $\frac{k}{3}(B_Z - B_Y)$, and the Reversal component by an integer multiple of $\frac{1}{6}(R_Y - R_Z)$.*

(In future work, we aim to use this lemma as a theorem to characterize decompositions directly from the combinatorics of nonparametric data sets.)

Lemma 9. *A sequence of neighbor transpositions which brings the Condorcet component to zero must make the Basic component an integer multiple of k.*

Proof of Theorem 7. Recall that if $n = 6\ell$, then $k = 3\ell$, so that the Basic components are a multiple of 3ℓ. The Kernel also is, as $n^3/6 = (6\ell)(6\ell)(2k)/6 = 3\ell(4k\ell)$, and clearly the Condorcet and Reversal components are, since they are zero! Then using the (integer!) column matrix obtained from the basis to get the voting profile, all entries are still divisible by 3ℓ, and we are done. □

We omit the proof of the following more general theorem, which is similar. Our hope is that thinking of the space of data sets in this way will prove fruitful for future characterization of the data sets themselves, not just their profiles! After all, once one has calculated the voting profile, the difficult task has already been done.

Theorem 10. *If one decomposes a profile coming from a nonparametric data set with n rows:*

- *The Basic components are all multiples of* $n/6$.
- *The Reversal components are either multiples of* $1/3$ *or* $1/6$*, depending on whether* n *is even or odd.*
- *The Condorcet component is either an even or odd multiple of* $n/6$*, depending on whether* n *is even or odd.*
- *The Kernel component is* $n^3/6$.

6. Complements

This paper began with the work of Haunsperger and Bargagliotti about consistency under aggregation. It should now be clear that there is in fact a great deal of consistency, so that at least for this sort of aggregation the paradoxes fail to some extent. Some examples which follow immediately from the theorems above include:

Corollary 11. *The Kruskal-Wallis test is consistent under stacking, as are all tests derived from procedures relying only on pairwise data (such as Mann-Whitney,* [17]*).*

Corollary 12. *All tests derived from points-based voting procedures (such as the V test) are consistent under stacking of data sets with no Reversal component.*

Corollary 13. *Paradoxes due solely to Reversal components (for instance, including most differences between Kruskal-Wallis and the V test) lessen under stacking* k *times (and disappear in the limit as* $k \to \infty$*).*

More important is whether the concept of stacking will help us with other aggregation questions. As noted above, it is somewhat frustrating that nearly all results about decompositions at this point rely upon first calculating the profile associated to a nonparametric matrix of ranks. Ideally, one would be able to say something directly about the statistical data, and not need to go through the proxy of voting profiles, for the statements of results (even if the proofs need the profiles).

Further work in this direction is needed, and, as mentioned earlier, we believe Lemma 8 will be a very useful tool for this. On a related note, any characterization of the subset of general voting profile space that these generate would be immensely helpful.

We would like to thank Don Saari and Anna Bargagliotti for helpful emails and encouragement of this project, which came out of a Gordon College REU. We would also like to thank the Gordon College Faculty Development Committee for the Initiative Grant which made the REU possible, and for Mike Veatch and the queuing theory group at Gordon for providing a good working environment during that time.

7. Proofs

Proof of Lemma 2. Given an $n \times 3$ matrix, if we were to replicate this matrix k times and stack the k matrices on top of one another, we have a $kn \times 3$ matrix with k stanzas. If we were to take all elements for each candidate from three separate stanzas, there would be $\binom{k}{3}$ ways to choose the three stanzas.

For any of these groupings, we can choose triplets so that the data point for one candidate comes from the top stanza, the data points for another candidate comes from the middle stanza, and the data point for the last candidate comes from the bottom stanza. There are six possible arrangements of the candidates in this way, corresponding to the six possible rankings of candidates in a standard

profile vector. Further, each such arrangement produces n^3 triplets, so for every arrangement, there are n^3 of each possible ranking of the candidates.

Since each ranking of the candidates receives $\binom{k}{3}n^3$ triplets, this simply adds $\binom{k}{3}\frac{n^3}{6}$ to the Kernel component, and nothing more. Note that the proof of an equivalent statement for $m > 3$ is essentially the same, with m stanzas needed. □

Proof of Lemma 3. Assume we have the voting profile of an arbitrary matrix of ranks given by (a, b, c, d, e, f)[7]. We obtain the decomposition of the profile by multiplying (see [3, 21]) by the inverse matrix to the row space matrix of the components:

$$M = \frac{1}{6}\begin{pmatrix} 2 & 1 & -1 & -2 & -1 & 1 \\ 1 & -1 & -2 & -1 & 1 & 2 \\ 0 & 1 & -1 & 0 & 1 & -1 \\ -1 & 1 & 0 & -1 & 1 & 0 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Therefore, we have each of the components in terms of the entries of the original profile:

$$\begin{array}{lcl} B_A & = & 1/6(2a + b - c - 2d - e + f) \\ B_B & = & 1/6(a - b - 2c - d + e + 2f) \\ R_A & = & 1/6(b - c + e - f) \\ R_B & = & 1/6(-a + b - d + e) \\ C & = & 1/6(a - b + c - d + e - f) \\ K & = & 1/6(a + b + c + d + e + f) \end{array}$$

Now we look at the stacked matrix of ranks, which in this case is a new matrix with two stanzas. When we consider just the profiles of each individual stanza, we end up with twice our original profile, or $(2a, 2b, 2c, 2d, 2e, 2f)$. The rest of the associated voting profile must come from interactions between the two stanzas, then. In order to catalog these interactions, we count the number of triplets coming from each of the six possible situations in which one 'candidate"s rank comes from a different stanza than that of the other two candidates.

In the original profile, there are $a + b + c$ triplets in the original profile where $B \succ C$, and so with n choices for the rank of A, that yields $\frac{a+b+c}{n}$ duplets with $B \succ C$. Now looking at the situation of triplets coming from

$$\left(\begin{array}{c} A \\ \hline B \quad C \end{array}\right),$$[8]

where there are n choices for the rank of A, we see that there are $\frac{a+b+c}{n}n = a+b+c$ ways for $A \succ B \succ C$. We slightly abuse notation here and abbreviate this as $|ABC| = a + b + c$ (from this way of getting a triplet). By symmetry or direct calculation, that means $|ACB| = n(d + e + f)$ from this way of getting triplets.

Doing similar calculations yields the following summary:

$$\left(\begin{array}{c} A \\ \hline B \quad C \end{array}\right) \quad \begin{array}{l} |ABC| = a + e + f \\ |ACB| = b + c + d \end{array} \qquad \left(\begin{array}{c} B \quad C \\ \hline A \end{array}\right) \quad \begin{array}{l} |BCA| = a + e + f \\ |CBA| = b + c + d \end{array}$$

[7]Recall that this is in the order: $A \succ B \succ C, A \succ C \succ B, C \succ A \succ B, C \succ B \succ A, B \succ C \succ A, B \succ A \succ C$.

[8]That is, where A comes from the higher-stacked matrix of ranks, and hence always is ranked above the B and C ranks.

$$\left(\begin{array}{c} B \\ \hline A \quad C \end{array}\right) \quad \begin{array}{l} |BAC| = a+b+f \\ |BCA| = c+d+e \end{array} \qquad \left(\begin{array}{c} A \quad C \\ \hline B \end{array}\right) \quad \begin{array}{l} |ACB| = a+b+f \\ |CAB| = c+d+e \end{array}$$

$$\left(\begin{array}{c} A \quad B \\ \hline C \end{array}\right) \quad \begin{array}{l} |ABC| = a+b+c \\ |BAC| = d+e+f \end{array} \qquad \left(\begin{array}{c} C \\ \hline A \quad B \end{array}\right) \quad \begin{array}{l} |CAB| = a+b+c \\ |CBA| = d+e+f \end{array}$$

This gives us a list of all the triplets formed by interactions between the stanzas, which can be summed up in the profile vector

$$(2a+b+c+e+f, a+2b+c+d+f, a+b+2c+d+e, \\ b+c+2d+e+f, a+c+d+2e+f, a+b+d+e+2f)\,.$$

After adding the profile $(2a, 2b, 2c, 2d, 2e, 2f)$ from the number of triplets from each individual stanza, we get the new total voter profile:

$$(4a+b+c+e+f, a+4b+c+d+f, a+b+4c+d+e, \\ b+c+4d+e+f, a+c+d+4e+f, a+b+d+e+4f)\,.$$

Multiplying by the decomposition matrix gives the components of the $(k = 2)$ stacked matrix of ranks:

$$\begin{array}{rcl} B_A & = & 4/6(2a+b-c-2d-e+f) \\ B_B & = & 4/6(a-b-2c-d+e+2f) \\ R_A & = & 2/6(b-c+e-f) \\ R_B & = & 2/6(-a+b-d+e) \\ C & = & 4/6(a-b+c-d+e-f) \\ K & = & 8/6(a+b+c+d+e+f) \end{array}$$

Clearly, each Basic component and the Condorcet component was multiplied by four, and each Reversal component was multiplied by two, as desired[9]. □

Proof of Proposition 4. Let a to f be as usual. The third and fourth rows of M are what yield the reversal component, which gives

$$R_A = (1/6)(b+e-(c+f)) \text{ and } R_B = (1/6)(b+e-(a+d))\,.$$

So the vanishing of reversal is precisely equivalent to the conditions obtained by setting these expressions simultaneously to zero, which is

$$b+e = c+f \text{ and } b+e = a+d$$

which is precisely the statement of the proposition. □

Proof of Proposition 5. The proof here is identical except for using the fifth row of M to get the Condorcet component, and is left to the reader. □

Proof of Theorem 6. Suppose we know two in a row of the cycle, with a third not adjacent (without loss of generality, a, d, and e). Then $a = d - 2e + 2f$ will give us f, and by the first case we have all of them. Now suppose we know three of a Condorcet cycle, such as a, c, and e. We could appeal to general linear algebra, but it is possible to actually row-reduce directly from the given equations to obtain that $b = \frac{1}{3}(2a+2c-e)$, $d = \frac{1}{3}(-a+2c+2e)$, and $f = \frac{1}{3}(2a-c+e)$.

Up to this point we have not used that our profile comes from a data set, but now we need that there are n^3 total 'voters', so that $a+b+c+d+e+f = n^3$.

[9]The Kernel was multiplied by eight, which can lead to an alternate, but less insightful, proof of that part of Theorem 1.

Now suppose we know only two entries of the profile. If we know two in a row, such as a and b, we can use Proposition 4 to see that $d = n^3/3 - a$, and use the previous cases. If we know two, such as a and c, with one intervening, we can use Proposition 5 to see that $e = n^3/2 - a - c$ and use the previous case. However, it does turn out there are pure Basic profiles coming from data sets which have the same a and d, but differ in their other entries, so this theorem is as sharp as possible. □

Proof of Lemma 8. Because of symmetry, without loss of generality we can assume that XYZ is ABC.

Now we consider which 'votes' change between two data sets under such a transformation. Clearly none that do not involve the the ranks i and $i-1$ directly change. Further, any which has only *one* of them involved also does not change, as the relative position of i and $i-1$ is the same compared to all other ranks in the matrix. That means that only the votes with $B \succ C$ coming from $i > i-1$ will change. Suppose that there are a such votes with $A \succ B \succ C$, and b such votes with $B \succ C \succ A$ (where clearly $a+b=n$). The the voting profile differential involved is $(-a, a, 0, b, -b, 0)$, and its decomposition is, as desired,

$$\left(-\frac{1}{6}(a+b), -\frac{1}{3}(a+b), \frac{1}{6}(a-b), \frac{1}{3}(a-b), -\frac{1}{3}(a+b), 0\right) = \frac{k}{3}(B_C - B_B) + \frac{a-b}{6}(R_B - R_C) - \frac{2k}{3}C\,.$$

□

Proof of Lemma 9. Since the Condorcet component of the unanimous matrix of ranks is $4k^3/3$, to bring this to zero we will need a (net) total of $(4k^3/3)/(2k/3) = 2k^2$ transpositions where Y, Z is one of the pairs A, B, B, C, or C, A. That means there are p_{YZ} neighbor transpositions of each type, with

$$p_{AB} + p_{BC} + p_{CA} = 2k^2 + r \text{ and } p_{BA} + p_{AC} + p_{CB} = r$$

for some integer $r \geq 0$. Given Lemma 8, starting component $\frac{4k^3}{3}(B_A - B_C)$, the fact that $B_B = -B_A - B_C$, and the identities for p_{XY}, that means the total Basic component is

$$\frac{4k^3}{3}(B_A - B_C) + \frac{k}{3}\Big[p_{AB}(B_B - B_A) + p_{BC}(B_C - B_B) + p_{CA}(B_A - B_C) + p_{BA}(B_A - B_B) + p_{AC}(B_C - B_A) + p_{CB}(B_B - B_C)\Big] =$$

$$\left[\frac{4k^3}{3} + \frac{k}{3}(-2p_{AB} + p_{BC} + p_{CA} + 2p_{BA} - p_{AC} - p_{CB})\right] B_A + \left[\frac{4k^3}{3} + \frac{k}{3}(-p_{AB} + 2p_{BC} - p_{CA} + p_{BA} + p_{AC} - 2p_{CB})\right] B_C =$$

$$\left[\frac{4k^3}{3} + \frac{k}{3}(-3p_{AB} + 2k^2 + 3p_{BA})\right] B_A + \left[-\frac{4k^3}{3} + \frac{k}{3}(3p_{BC} - 2k^2 - 3p_{CB})\right] B_C =$$

$$k\left[(2k^2 + p_{BA} - p_{AB})B_A + (2k^2 + p_{BC} - p_{CB})B_C\right]\,,$$

which is indeed a multiple of k. □

References

[1] Kenneth J. Arrow. *Social Choice and Individual Values*. Cowles Commission Monograph No. 12. John Wiley & Sons Inc., New York, N. Y., 1951.

[2] A. Bargagliotti and R. Greenwell. Statistical significance of ranking paradoxes. *Preprint*, 2009.

[3] A. Bargagliotti and D. Saari. Symmetry of nonparametric statistical tests on three samples. *Preprint*, 2007.

[4] Anna E. Bargagliotti. Aggregation and decision making using ranked data. *Mathematical Social Sciences*, 58(3):354 – 366, 2009.

[5] V. P. Bhapkar. A nonparametric test for the problem of several samples. *Ann. Math. Statist.*, 32:1108–1117, 1961.

[6] P.J. Bickel, E.A. Hammel, and J.W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–404, 1975.

[7] Karl-Dieter Crisman. Undergraduate research in the mathematics of voting and choice using Sage (in 'AMS Special Session on SAGE and Mathematical Research Using Open Source Software', communicated by David Joyner). *ACM Commun. Comput. Algebra*, 43(2), 2009.

[8] Zajj Daugherty, Alexander K. Eustis, Gregory Minton, and Michael E. Orrison. Voting, the symmetric group, and representation theory. *The American Mathematical Monthly*, 116(8), 2009.

[9] Jianhua Guo, Zhi Geng, and Ningzhong Shi. On collapsibilities of Yule's measure. *Sci. China Ser. A*, 44(7):829–836, 2001.

[10] Deanna B. Haunsperger. Dictionaries of paradoxes for statistical tests on k samples. *J. Amer. Statist. Assoc.*, 87(417):149–155, 1992.

[11] Deanna B. Haunsperger. Aggregated statistical rankings are arbitrary. *Soc. Choice Welf.*, 20(2):261–272, 2003.

[12] Deanna B. Haunsperger and Donald G. Saari. The lack of consistency for statistical decision procedures. *Am. Statist.*, 45(417):252–255, 1991.

[13] Julian Havil. *Impossible?* Princeton University Press, Princeton, NJ, 2008.

[14] S. Julious and M. Mullee. Confounding and Simpson's paradox. *British Medical Journal*, 309:1480–1481, 1994.

[15] Matthias P. Kläy and David J. Foulis. Maximum likelihood estimation on generalized sample spaces: an alternative resolution of Simpson's paradox. *Found. Phys.*, 20(7):777–799, 1990.

[16] William H. Kruskal. A nonparametric test for the several sample problem. *Ann. Math. Statistics*, 23:525–540, 1952.

[17] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics*, 18:50–60, 1947.

[18] Donald G. Saari. The source of some paradoxes from social choice and probability. *J. Econom. Theory*, 41(1):1–22, 1987.

[19] Donald G. Saari. *Geometry of voting*, volume 3 of *Studies in Economic Theory*. Springer-Verlag, Berlin, 1994.

[20] Donald G. Saari. *Basic geometry of voting*. Springer-Verlag, Berlin, 1995.

[21] Donald G. Saari. Explaining all three-alternative voting outcomes. *J. Econom. Theory*, 87(2):313–355, 1999.

[22] Donald G. Saari. Mathematical structure of voting paradoxes. I. Pairwise votes. *Econom. Theory*, 15(1):1–53, 2000.

[23] Donald G. Saari. Mathematical structure of voting paradoxes. II. Positional voting. *Econom. Theory*, 15(1):55–102, 2000.

[24] Myra L. Samuels. Simpson's paradox and related phenomena. *J. Amer. Statist. Assoc.*, 88(421):81–88, 1993.

[25] Rasa Šleževičienė-Steuding and Jörn Steuding. Simpson's paradox in the Farey sequence. *Integers*, 6:A4, 9 pp. (electronic), 2006.

[26] W. A. Stein, M. Hansen, et al. *Sage Mathematics Software (Version 4.0)*. The Sage Development Team, 2009. `http://www.sagemath.org`.

[27] J. Terwilliger and M. Schield. Frequency of Simpson's Paradox in NAEP data, 2004. Conference of the American Education Research Association (AERA, San Diego).

Gordon College, 255 Grapevine Road, Wenham, MA 01984