Supplementary material: Appendix

Jack R. Williams[1], Samuel Baltz[2], and Charles Stewart III[3]

[1] *Democracy Works. main@jackryanwilliams.com*
[2] *Massachusetts Institute of Technology. sbaltz@umich.edu*
[3] *Massachusetts Institute of Technology. cstewart@mit.edu*

1 Identifiability of IRV versus other electoral systems

The list of rankings submitted in IRV is often accompanied by Cast Vote Records (CVRs) which show how a specific (anonymous) person voted across their entire ballot. So why are we concerned about IRV identifiability, when instead someone may just be able to signal their vote choice by voting according to a certain pattern across their whole ballot (say, selecting candidate A in the first office on the ballot, candidate B in the second office, and so on).

We argue that IRV is especially vulnerable for two reasons. First, the permutations possible in an IRV election are explosively larger than in a single-vote CVR. Second, the information in IRV will often be free, while on a CVR it may not be. We provide a stylized illustration of both arguments.

The core numerical difference between the number of possible sequences in IRV and the number of votes across offices in a CVR is that, in the CVR, candidates are constrained such that they can only appear in specific locations, and only one candidate per contest can be included in the ranking. Consider a ballot that contains an IRV election between 10 candidates. In the main text we will show that the number of possible sequences, depending on the method by which votes are counted and reported, could be on the order of 10^8.

Compare this with the simple case of 5 contests on a ballot, which are each contested by 2 candidates, and in each contest voters can only vote for one option. In office one, the voter can select either candidate A or B, in the second office they can select C or D, and so on. By concatenating the choices they make across offices, we can form a string like $ACEGI$. The possible number of such strings is 32, whereas if these 10 candidates were contesting an IRV election with $L = 10$, more than one hundred million strings could be formed. Of course, the CVR becomes more identifiable if each election has more candidates or a higher magnitude, but heroic assumptions are needed for a CVR of non-ranked elections to approach the identifiability of a reasonably large IRV election.

There is also a crucial structural difference between the scheme we outline for identifying IRV ballots, and anything that can be done on a CVR: in our IRV scheme, the information is encoded without any potential electoral downside. Suppose a vote-buyer purchases a vote from a voter in a 10-candidate IRV race, in support of candidate A, who the voter is paid to rank first. Then, the vote-buyer assigns the voter a specific sequence in which to rank the other nine candidates. If A wins, none of the voter's rankings after A are counted, so the information was placed on the ballot without any effect on the vote count. If A loses, then the scheme has failed. Even if the voter and vote-buyer have preferences between the candidates other than A, if 10 candidates are ranked, for most plausible distributions of vote counts it is unlikely that the last several rankings will be used; this information is included in the ballot, but not relevant to practical vote-counting.

In the version of the scheme in which a sequence is signalled by casting certain votes across single-vote races on CVR, however, some of those elections may be close elections with small electorates (say, a competitive school board race). In this case, the voters involved in the scheme have a serious risk of being pivotal in an unrelated contest. In the CVR scheme, the voter is asked to affect real vote counts for unrelated races, whereas in the IRV scheme, much of the information used is unlikely to affect real vote counts.

As we note in the text, the problem is much worse for systems like Approval Voting or Borda

Political Analysis (2024)

DOI: 10.1017/pan.xxxx.xx

Corresponding author
Samuel Baltz.
sbaltz@umich.edu

Edited by
John Doe

© The Author(s) 2024. Published by Cambridge University Press on behalf of the Society for Political Methodology.

Count. There, our scheme would actually require the vote buyer to increase the vote total of the people who are competing against their preferred candidate. We therefore consider the scheme particularly ill-suited for those systems. Single-Transferable Voting might provide more fruitful ground for our scheme, especially if the scheme were enacted to support a slate of candidates rather than just one, but transfer ballots in particular would complicate any analysis of how feasible the scheme is exactly. Taken together, all these reasons motivate focusing on IRV specifically for the first analysis of this type of scheme, knowing that its extension to other electoral systems is theoretically plausible but severely nontrivial.

2 Reporting methods in real IRV contests

Election	Method	Estimate	Source
Alameda County, 2021	Partial list	$<$ No Blanks	Alameda County (2022)
Alaska, 2022	All rankings	Any Blank	Alaska Division of Elections (2022)
Australia House, 2022	Partial list	$<$ No Blanks	Australian EC (2022)
Maine, 2022	All rankings	Any Blank	Maine Elections Bureau (2022)
New York City, 2021	CVR	$>$ Any Blank	New York City (2021)
Papua New Guinea, 2022	# rankings	None	EC Papua New Guinea (2022)
San Francisco, 2022	CVR	$>$ Any Blank	City of San Francisco (2022)

Table 1. EC stands for Electoral Commission. The method column is the most granular type of reporting we found: All rankings for just a full list of the rankings that were cast in the IRV election, Cast Vote Record (CVR) for entire ballots including the ranked choices in the IRV contest, a partial list for incomplete lists of IRV orderings, or just the candidates' total numbers of points. The "ruleset" is the appropriate abstention rule for estimating how identifiable that type of reporting is.

Table 1 reports the largest amount of data that we have been able to find through a straightforward perusal of public sources. This is a conservative table in the sense that, with a more concerted effort (such as contacting the governments responsible with a specific request for a richer type of data), it may be possible to uncover more data from each of these elections, or more data may become available. We have restricted our attention to recent elections for government office in democracies with robust public election result reporting.

In Australia as well as Alameda County, California, the files which provide distributions of preferences provide enough information that a version of our scheme may be theoretically feasible, but reveal less information than any of the cases that we examine in the paper. Specifically, those files show the flow of preferences as candidates were eliminated. From those files, one could infer how many people ranked a given non-eliminated candidate immediately after each eliminated candidate. However, we have not found a way to infer from public election data whether a) a previously eliminated candidate was ranked in between some eliminated candidate and the candidate the votes flow to, or b) any rankings that follow the final candidates in the election. This is a truncated version of the problem that we study, but how much less information it provides will depend on the distribution of ballots and the candidate elimination order, so it is an open question under what conditions this practice provides enough information to pose a risk in small electorates. In the case of Australia, we confirmed that there is no obvious way to infer this from public data with a phone call to the Australian electoral commission, and we are grateful to Campbell Sharman for providing helpful context.

Note that in Alaska and Maine alike, where the full list of rankings is called a "Cast Vote Record", the main data outputs that contain this information actually just provide the full rankings of candi-

dates is the IRV race alone, without tying them to voters' choices in any other election.[1] Because the reporting of IRV rankings includes undervotes in exactly the positions in which voters listed them, both elections are exactly examples of the Any Blanks ruleset.

Alaska explicitly motivates its choice to provide CVRs for the IRV race with the explanation that they are "providing the CVR to increase transparency and provide voters the data necessary to confirm the results".

We have not found an example of a governmental election using the No Blanks ruleset. This is as expected, since that ruleset was intended to provide a theoretical lower bound on the number of available sequences.

3 Identifying the most scientifically conservative assumption

We suspect that we are not the first to maximize the estimator of the number of unique types in a sample, but we have been unable to find a fully worked through derivation of the values of the population proportions p_i which maximize observed types for a fixed sample size m anywhere in the existing literature. We therefore produce an original proof ourselves below, although we think it is likely that this result is not new, and maybe indeed be quite old. We are very grateful to Kevin E. Acevedo Jetter for his collaboration in this derivation, which also benefited from notes by Iain Osgood.

$$\max[E(S_m)] = \max[S - \sum_{i=1}^{S}(1-p_i)^m] \text{ such that } \sum_{i=1}^{S} p_i = 1$$

Since all terms are nonnegative,

$$\max[E(S_m)] = S - \min[\sum_{i=1}^{S}(1-p_i)^m] \text{ such that } \sum_{i=1}^{S} p_i = 1$$

This is a classic Lagrangian, of the form

$$\mathcal{L} = \sum_{i=1}^{S}(1-p_i)^m + \lambda\big(\sum_{i=1}^{S} p_i - 1\big)$$

Then, for each p_i,

$$\mathcal{L}_{p_i} = -m(1-p_i)^{m-1} + \lambda$$

Setting the partial derivative to 0,

$$\lambda = m(1-p_i)^{m-1}$$

$$p_i = 1 + \sqrt[m-1]{\frac{\lambda}{m}}$$

So the number of distinct ballots in the sample is maximized exactly when the p_i are all equal. What value do they have? Substituting this identity into the constraint, and by the fact that the p_i are all equal,

1. It is always possible that we have overlooked some data output, or one can be obtained through special request or will subsequently be produced by the governments, which in these specific cases connects the ballots in those files to ballots cast in other contests so that individual voters' choices in an IRV contest can be connected through another means to their choices in other contests. If that is the case, we contend that the usefulness of Alaska and Maine as illustrative examples should not be undermined. We are discussing the main way that these states report IRV results to the public, and because we do not believe that vote buying took place in these specific elections, we are only using them as illustrative examples to discuss a *type* of data reporting.

$$p_i = \frac{1}{S} \quad \forall i$$

Because the constraint function is linear, the constraint qualification is satisfied, this is the unique critical point. To establish that this critical point is a minimum, we assume $S \geq 1$ and $m \geq 2$ (in our application these respectively mean there is at least one candidate and at least two voters), and it will be sufficient to show that the final $S-1$ leading principle minors of the bordered Hessian are negative (Sundaram 1996, §5.3). The Bordered Hessian $\mathbf{B}$ has the form

$$\mathbf{B} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & \mathcal{L}_{p_1 p_1} & \mathcal{L}_{p_1 p_2} & \cdots & \mathcal{L}_{p_1 p_S} \\ 1 & \mathcal{L}_{p_2 p_1} & \mathcal{L}_{p_2 p_2} & \cdots & \mathcal{L}_{p_2 p_S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathcal{L}_{p_S p_1} & \mathcal{L}_{p_S p_2} & \cdots & \mathcal{L}_{p_S p_S} \end{bmatrix}$$

Because the Lagrangian is

$$\mathcal{L} = \sum_{i=1}^{S} (1 - p_i)^m + \lambda \big(\sum_{i=1}^{S} p_i = 1 \big)$$

all cross-partial derivatives equal zero, so $\mathbf{B}$ simplifies to

$$\mathbf{B} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & \mathcal{L}^2_{p_1} & 0 & \cdots & 0 \\ 1 & 0 & \mathcal{L}^2_{p_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & \mathcal{L}^2_{p_S} \end{bmatrix}$$

By a standard theorem (Nicholson 1995, p. 113), $\det(\mathbf{B})$ is equal to the determinant of any matrix obtained by adding a multiple of one of the rows of $\mathbf{B}$ to another row of $\mathbf{B}$. So,

$$det(\mathbf{B}) = \begin{vmatrix} -\frac{1}{\mathcal{L}^2_{p_1}} - \frac{1}{\mathcal{L}^2_{p_2}} - \cdots - \frac{1}{\mathcal{L}^2_{p_S}} & 0 & 0 & \cdots & 0 \\ 1 & \mathcal{L}^2_{p_1} & 0 & \cdots & 0 \\ 1 & 0 & \mathcal{L}^2_{p_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & \mathcal{L}^2_{p_S} \end{vmatrix}$$

This is a matrix of the form

$$\mathbf{B} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{X} & \mathbf{C} \end{bmatrix}$$

Where $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{C}$ is a $m \times m$ matrix, so by another standard theorem (p. 117), $\det(\mathbf{B}) = \det(\mathbf{A}) \cdot \det(\mathbf{C})$. Since $n = 1$,

$$\det \mathbf{A} = -\frac{1}{\mathcal{L}^2_{p_1}} - \frac{1}{\mathcal{L}^2_{p_2}} - \cdots - \frac{1}{\mathcal{L}^2_{p_S}}$$

And because $\mathbf{C}$ is a diagonal matrix, $\det(\mathbf{C})$ is the product of the entries on its principal diagonal, so

$$\det(\mathbf{C}) = \prod_{i=1}^{S} \mathcal{L}_{p_i}^2$$

For a generic $\mathcal{L}_{p_i}^2$,

$$\frac{\partial^2 \mathcal{L}}{\partial i^2} = (m^2 - m)(1 - p_i)^{m-2}$$

Since $m \geq 2$,

$$\frac{\partial^2 \mathcal{L}}{\partial i^2} > 0$$

and therefore

$$\prod_{i=1}^{S} \mathcal{L}_{p_i}^2 > 0$$

and

$$-\frac{1}{\mathcal{L}_{p_1}^2} - \frac{1}{\mathcal{L}_{p_2}^2} - \cdots - \frac{1}{\mathcal{L}_{p_S}^2} < 0$$

so that

$$\det(\mathbf{A}) \cdot \det(\mathbf{C}) < 0$$

$$\det(\mathbf{B}) < 0$$

We next use the determinant of the full matrix to check the final $S - 1$ leading principle minors of $\mathbf{B}$. Consider the kth leading principal minor, where k is any natural number from 2 up to S (since we need only check the sign of the final $S - 1$ leading principal minors). By the same theorem we invoked to compute the determinant of a diagonal matrix, the kth leading principal minor is

$$\frac{\det(\mathbf{B})}{\mathcal{L}_{p_{k+1}}^2 \cdot \mathcal{L}_{p_{k+2}}^2 \cdots \cdot \mathcal{L}_{p_S}^2}$$

The denominator is the product of individually positive numbers, so the whole fraction is negative for any k. The condition for the critical point to be a local minimum is that the sign of the leading principal minors must match the sign of $(-1)^m$, which is satisifed. So, the minimizer of the population proportions is $p_i = \frac{1}{S} \quad \forall i$

□

We have shown that setting $p_i = \frac{1}{S}$ for all i maximizes the number of distinct types we expect to appear in a sample of generic size. By a nearly identical argument we could show that the same critical point maximizes the sample coverage estimator.

4 Proportions cast under varying abstention rules

Here we supplement the Blanks Last table in the article. Table 2 shows how many complete sequences of candidates can be constructed for contests that include between 2 and 10 candidates using the No Blanks ruleset, and ballot lengths that also range from 2 to 10.

Ballot length \ Cands	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
2	2	6	12	20	30	42	56	72	90
3		6	24	60	120	210	336	504	720
4			24	120	360	840	1,680	3,024	5,040
5				120	720	2,520	6,720	15,120	30,240
6					720	5,040	20,160	60,480	151,200
7						5,040	40,320	181,440	604,800
8							40,320	362,880	1,814,400
9								362,880	3,628,800
10									3,628,880

Table 2. The number of sequences that can be formed when no rankings can be skipped.

Table 2 shows that even under the harsh restriction that voters must complete every ranking, many thousands of distinct sequences can be formed even in contests with a fairly modest number of candidates and reasonably short ballots. If just eight candidates are contesting an election in which voters may rank seven of them, 40,320 distinct sequences can be cast. If there are 10 candidates with a ballot length of eight or more, the number of distinct sequences rises into the millions.

How does this compare to the situation in which any rankings can be blank? Table 3 shows how the number of rankings grows in the number of candidates and in the ballot length under the Any Blanks ruleset.

Ballot length \ Cands	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
2	6	12	20	30	42	56	72	90	110
3		33	72	135	228	357	528	747	1,020
4			208	500	1,044	1,960	3,392	5,508	8,500
5				1,545	4,050	9,275	19,080	36,045	63,590
6					13,326	37,632	93,228	207,774	424,050
7						130,921	394,352	1,047,375	2,501,800
8							1,441,728	4,596,552	12,975,560
9								17,572,113	58,941,090
10									234,662,230

Table 3. The number of sequences that can be formed when any number of rankings can be left blank, so long as at least one candidate is ranked.

When any rankings can be left blank, the number of sequences reaches the thousands (about the size of many American precincts) so long as there are five candidates and all of them can be ranked, or if there are at least six candidates and at least four can be ranked.

Figure 1 shows the number of possible rankings under each of the three rules when the number

of candidates is fixed at 10 but the number of candidates that can be ranked varies from 2 to 10.

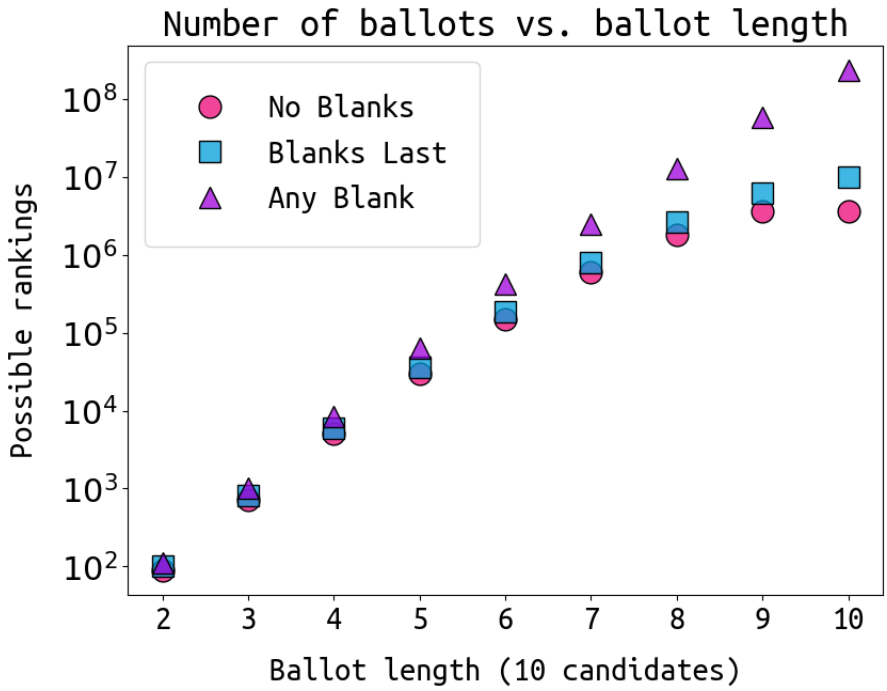


Figure 1. How the possible number of ballot sequences varies in the case of a 10-candidate contest as the number of candidates who can be ranked on a ballot varies.

Similarly, Table 4 holds the number of candidates fixed at 10, and shows how the expected number of ballots that will not be cast varies as the number of those 10 candidates that can be ranked varies.

Population \ Ranks	3	4	5	6	7	8	9	10
10^1	810	5,850	36,090	187,290	792,090	2,606,490	6,235,290	9,864,090
10^2	726	5,761	36,000	187,200	792,000	2,606,400	6,235,200	9,864,000
10^3	242	4,941	35,114	186,303	791,101	2,605,500	6,234,300	9,863,100
10^4	0	1,063	27,365	177,562	782,163	2,596,519	6,225,308	9,854,105
10^5	0	0	2,262	109,816	698,155	2,508,394	6,136,098	9,764,605
10^6	0	0	0	899	224,129	1,775,979	5,311,368	8,913,118
10^7	0	0	0	0	3	56,216	1,254,146	3,579,147
10^8	0	0	0	0	0	0	1	390

Table 4. The expected number of sequences that will *not* be cast in an election with a certain ballot length, fixing the number of candidates at 10, and in an electorate of a given population, using the Blanks Last vote-counting method. We take the maximally conservative assumption that all sequences are equally likely to be cast.

5 Estimating collision probabilities

Here we explore the probability that a bought vote collides with a legitimate vote, using the size and rules of Oakland as an example, and then we consider three options for the hypothetical vote

buyer seeking to secure the estimated 1,500 votes that they need.

Since we assume that a voter has equal probability of casting any particular sequence, a sequence generated by the vote buyer in this election has probability $\frac{54,684}{63,590}$ of being present already in the cast votes without any vote-buying activity. A reasonable approximation, when the population and the number of sequences are large, is to assume that the probabilities that each new sequence collides with a legitimate vote are independent of each other. Then whether or not each sequence matches a legitimate vote is a Bernoulli trial, and the expected number of ballots that will be precisely identifiable is the sum of the probabilities that each ballot is identifiable. In that case, the vote-buyer should expect that about 14% of the sequences that they assign to voters will be cast by *only* that voter. This means that, if they purchase 1,500 votes, they should only expect to be able to confidently identify about 210 votes.

Of course, the independence assumption is severely strained here, with more than $\frac{1}{7}$ of the expected number of unique sequences being assigned. Perhaps a better model is to imagine the vote buyer sequentially generating 1,500 sequences to assign. Each sequence generated that is not already cast by some voter reduces the total number of sequences that remain to be generated, but does not reduce the number of legitimate sequences that might be matched by the next generated sequence; meanwhile, each sequence that matches a legitimate vote reduces the total number of remaining legitimate sequences by one, and also the remaining number of covered sequences by one. This specifies a sum in which the probability of encountering a collision with each sequence decreases proportional to the initial probability of a collision, but depending on the order in which collisions happen to be encountered. So, a vote buyer might be well-advised to model the necessary number of votes using draws from a hypergeometric distribution. However the probabilities are modeled, this poses a serious challenge to the vote buyer: the number of rankings available to voters has been limited, and consequently their ability to buy votes is highly constrained.

So, what are the vote buyer's options? First, they could decide that this is a sufficient number of votes to expect to be confident about, and purchase those 1,500 votes, expecting that even if every voter cooperates and successfully casts the ranking they assigned, only about 210 of those sequences will be uniquely cast by the assigned voter.

A second option is for the vote-buyer to decide that they would like to be able to confidently identify that 1,500 votes were successfully purchased, and to simply generate extra sequences until they expect 1,500 of them to be uniquely cast. In this case, using the Bernoulli approximation, that requires a dramatically larger investment: they must assign about 10,710 sequences. Then they have the delicate challenge of deciding whether to reward only those voters who were lucky enough to cast a unique sequence, or paying all 10,710 voters even though only 1,500 of them are expected to cast a unique sequence.

This dilemma suggests the third approach, which is much more complicated but also much more efficient: the vote buyer could actually compute the number of expected times that each sequence will appear in the population, and then assume that a vote was successfully purchased if that vote appeared above the expected number of times in the election results. Of course, while this approach seems theoretically much better, it has two sharp practical downsides: first that it requires highly accurate expectations about precisely how many people will submit each possible ordering, and second that it admits randomness into whether or not voters are rewarded for colluding.

6 Data from other IRV races

Table 5 presents summary information for 36 IRV races obtained via a library of voting preference data (Mattei and Walsh 2013). The number of candidates and ranks in many of these races suggests that they would have similar hypothetical opportunities for vote-buying.

Election	Candidates	Ranks	Possible Seqs	Ballots Cast	Coverage
Aspen City Council 2009	9	9	986409	2487	<1%
Aspen Mayor 2009	4	4	64	2528	62%
2010 Berkeley City Council (d7)	3	3	15	4189	60%
2006 Burlington Mayoral Election	5	5	325	9788	45%
2009 Burlington Mayoral Election	5	5	325	8980	46%
2009 Minneapolis Parks & Rec	8	3	400	36655	100%
2009 Minneapolis Tax Board	6	3	156	32086	100%
2010 Oakland City Council (d4)	7	3	259	21040	100%
2010 Oakland Mayor	10	3	820	119962	99%
2010 Oakland City Council (AL)	5	3	85	145443	100%
2010 Oakland City Council (d1)	7	3	259	28766	100%
2010 Oakland City Council (d3)	6	3	156	22193	100%
2010 Oakland City Council (d5)	4	3	40	11358	100%
2010 Oakland School Director (d3)	3	3	15	20753	60%
2009 Pierce County Auditor	3	3	15	153721	60%
2008 Pierce County Council (d2)	3	3	15	40031	60%
2008 Pierce County Executive	4	3	40	299664	100%
2008 Pierce County Treasurer	6	3	156	262810	100%
2008 S.F. Board of Sups. (d1)	9	3	585	28998	97%
2011 S.F. District Attorney	5	3	85	184046	100%
2011 S.F. Mayor	16	3	3616	195237	85%
2011 S.F. Sheriff	4	3	40	183611	100%
2012 S.F. Board of Sups. (d5)	8	3	400	35356	100%
2012 S.F. Board of Sups. (d7)	9	3	585	31566	99%
2008 S.F. Board of Sups. (d11)	9	3	585	25083	67%
2008 S.F. Board of Sups. (d3)	9	3	585	27482	97%
2008 S.F. Board of Sups. (d4)	3	3	15	29522	60%
2008 S.F. Board of Sups. (d9)	7	3	259	26799	100%
2010 S.F. Board of Sups. (d10)	21	3	8421	18308	39%
2010 S.F. Board of Sups. (d2)	6	3	156	24180	99%
2010 S.F. Board of Sups. (d6)	14	3	2380	21443	67%
2010 S.F. Board of Sups. (d8)	4	3	40	35029	100%
2010 San Leandro Mayor	6	3	156	22539	54%
2012 San Leandro City Council (d2)	3	3	15	25564	60%
2012 San Leandro City Council (d4)	4	3	40	23359	100%
2007 Takoma Park City Council	3	3	15	204	60%

Table 5. The sample coverage of some real elections. District numbers are identified by "d" and then a number in parentheses, with at-large races denoted AL. Data are from the PrefLib library (Mattei and Walsh 2013). We drop every vote that includes a write-in candidate, and every invalid vote, for example votes that rank the same candidate in multiple positions.

7 Proportion cast of Any Blanks ballots

The following corresponds to Figure 3 in the text, but under the Any Blanks ruleset instead of the Blanks Last ruleset.

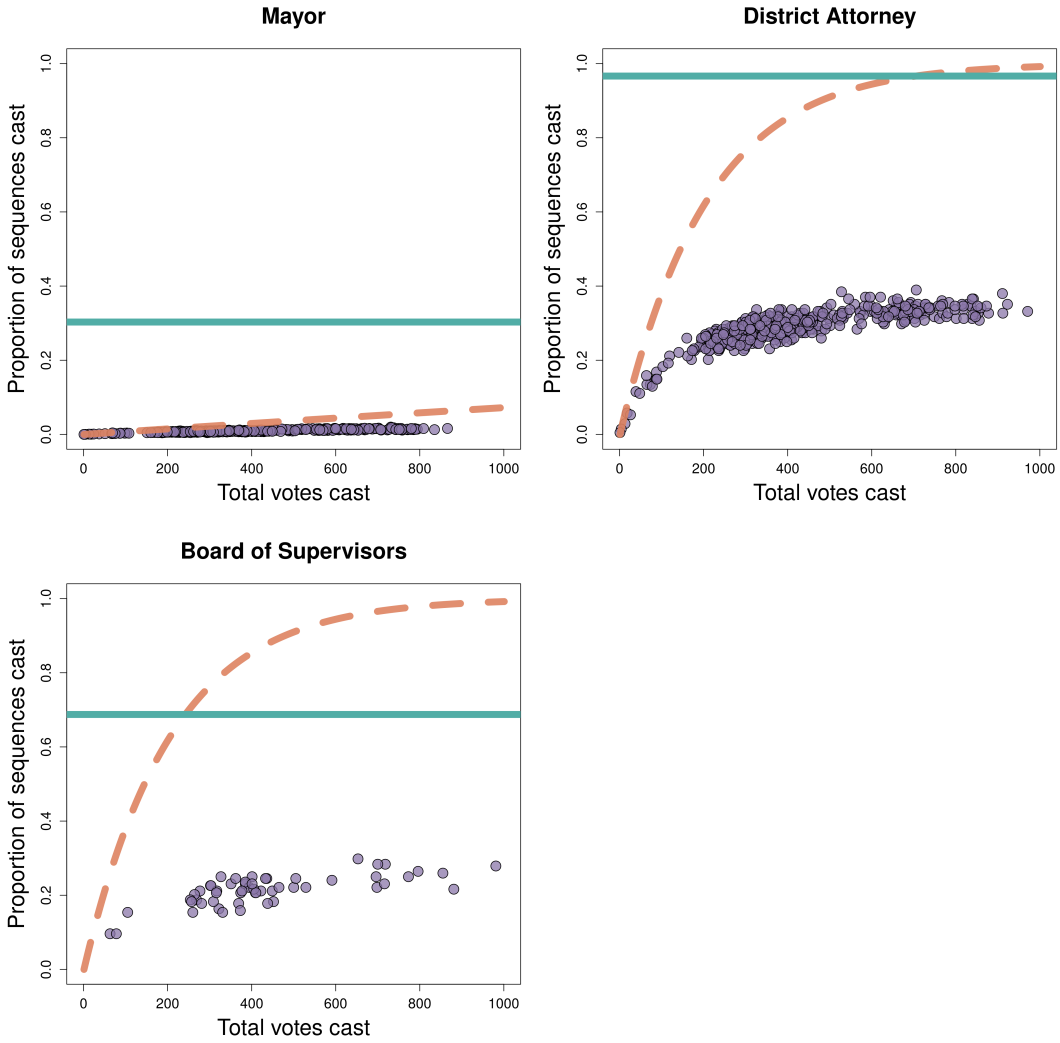


Figure 2. The proportion of all possible sequences cast in three San Francisco IRV races, using the Any Blanks ruleset. Each dot represents the proportion cast in a precinct, with the total number of ballots cast in that precinct on the x-axis. The solid horizontal line is the proportion of possible ballots cast across the whole election, and the dashed curve is the expected number of ballots cast in a precinct of a given size under our conservative assumption that $p = \frac{1}{S}$.

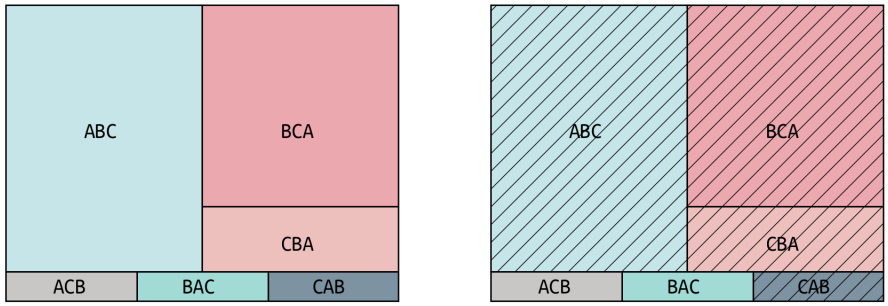
8 Connection between number of types and "sample coverage"

Here we establish why we use the phrase "sample coverage" for the proportion of rankings represented in the election. "Sample coverage" is the proportion of individuals of any type that is represented in the sample. Let us illustrate this idea in our application. Consider a 3-way contest between candidates A, B, and C under the No Blanks ruleset. The possible sequences are:

$$\{ABC, ACB, BAC, BCA, CAB, CBA\}$$

Imagine a poll is taken before the election and we find that 45% of people expect to be ABC voters, 35% are BCA voters, 11% are CBA voters, and the remaining 9% are split equally across

the other three options. Then the election occurs, and among all of the ballots cast, we observe four distinct rankings: at least one voter cast each of ABC, BCA, CBA, and CAB. Then the coverage of this sample is $0.45 + 0.35 + 0.11 + 0.03 = 0.94$, or 94%.[2] Figure 3 illustrates the example graphically.



(a) Population types (b) Sampled types

Figure 3. The size of the rectangle indicates the share of voters of each type in the population of eligible voters. On the right, the shaded regions represent sequences that were cast in the election. The proportion of the area that is shaded is the coverage.

We will show that, under the conservative assumption that each voter has equal probability of casting each possible ranking, the expected sample coverage is the same as the expected share of the types that are represented in the sample. From Good (1953), the expected number of types represented in a sample is

$$E[S_m] = S - \sum_{i=1}^{S}(1-p_i)^m$$

The proportion of all of the types represented in that sample is therefore

$$\frac{E[S_m]}{S} = 1 - \frac{1}{S}\sum_{i=1}^{S}(1-p_i)^m$$

We have identified that the most scientifically conservative assumption we can make is that that $p_i = p_j \ \forall i, j$. With that assumption,

$$\frac{E[S_m]}{S} = 1 - (1-p)^m$$

Now note that expected coverage, per Chao and Jost (2012, p. 2535), is

$$E[C_m] = 1 - \sum_{i=1}^{S} p_i(1-p_i)^m$$

Under our assumption,

$$E[C_m] = 1 - Sp(1-p)^m$$

Since we assume $p = \frac{1}{S}$,

2. This is closely adapted from an example in Chao and Jost (2012, p. 2534).

$$E[C_m] = 1 - (1-p)^m$$

$$E[C_m] = \frac{E[S_m]}{S}$$

□

9 Simulated effectiveness of buying sequences

We simulate bought votes by assigning unique sequences to be bought, and then counting the overlap at two hypothetical reporting levels: precinct-by-precinct using CVRs, and city-wide. One small complication is that San Francisco's CVRs include two ballot styles: one type includes the board of supervisors race for District 5 and the other type does not. We analyze these separately, as they correspond to different possible sequences.

In the simulations we fix the first rank of one of the races for a particular candidate and generate all possible sequences with that candidate first. We take a sample of these sequences as a proportion of all ballots cast in a race, and increase the number of sequences sampled until we have simulated as many bought votes as there were real votes cast in the election. We repeat the samples 500 times, and report the average collision rate across simulations. Figure 4 shows that the rate of collisions, when votes are bought for any candidate, remain well under 1%: fewer than 1% of the bought votes collide with legitimate votes.

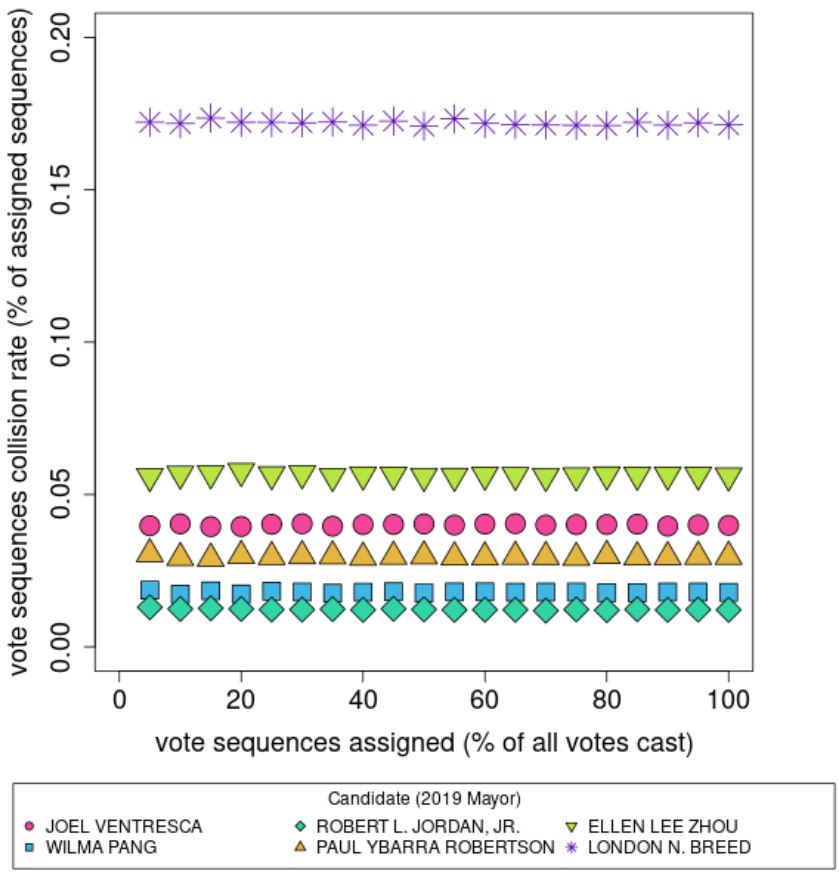
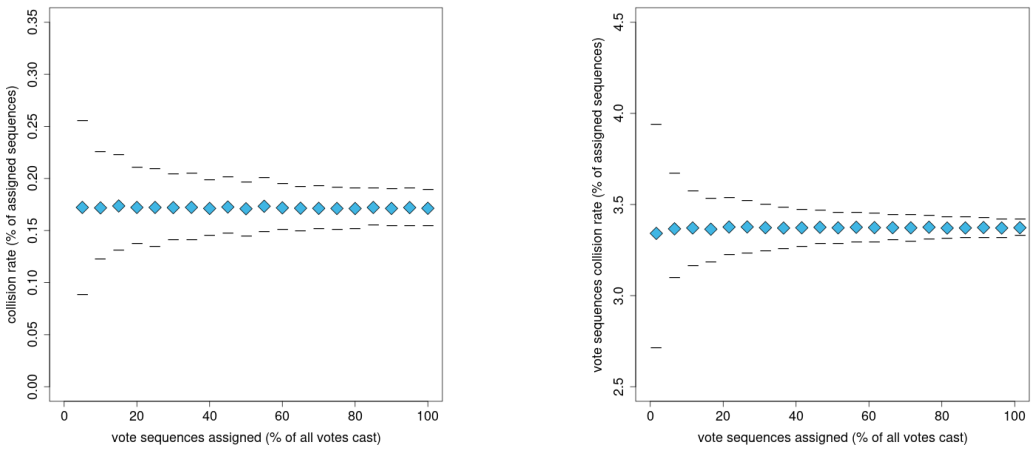


Figure 4. Ballot-type-level (without the Board of Supervisors race) simulations for the 2019 San Francisco Mayoral race. 500 simulations were run where a varying proportion of the votes cast in the election are bought, up to the point where there are as many bought votes as real votes. The figure shows the average number of collisions as a share of the sampled sequences.

Figure 5a shows that as ballot sequences are assigned until the number of bought votes equals the number of real votes, the simulations for a candidate will converge to the sample coverage for the candidate. The degree that this is a consequence of the ballot identifiers available on the CVRs can be tested by taking the combinations of sequences alone and running the same simulations citywide. Figure 5b shows the number of collisions a vote-buyer could expect only knowing the ballot level combinations, without the type of ballot, is still less than 4.5%.



(a) Ballot-type level

(b) City-wide

Figure 5. Simulations (without the Board of Supervisors race) of the collision rate a hypothetical vote buyer would encounter if they bought votes for the winning candidate in the 2019 San Francisco mayoral race. 500 simulations were run at varying percentages of votes cast, until the number of bought votes equals the number of real votes. Reported is the average number of collisions in the sampled vote sequences, along with the 95 percent interval for the simulations.

References

Alameda County. 2022. *General Election (Certified Final Results) - November 08, 2022.* https://www.acgov.org/rovresults/248/.

Alaska Division of Elections. 2022. *2022 General Election.* https://www.elections.alaska.gov/election-results/e/?id=22genr.

Australian EC. 2022. *House of representatives downloads.* https://results.aec.gov.au/27966/Website/HouseDownloadsMenu-27966-Csv.htm.

Chao, A., and L. Jost. 2012. "Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size." *Ecology* 93, no. 12 (December): 2533–2547. ISSN: 0012-9658. https://doi.org/10.1890/11-1952.1.

City of San Francisco. 2022. *November 8, 2022 Final Election Results - Detailed Reports.* https://sfelections.sfgov.org/november-8-2022-election-results-detailed-reports.

EC Papua New Guinea. 2022. *2022 National Elections - Preliminary Results.* https://results.pngec.gov.pg/preliminarydeclarationsummary.html.

Good, I. J. 1953. "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika* 40, nos. 3/4 (December): 237. ISSN: 00063444. https://doi.org/10.2307/2333344.

Maine Elections Bureau. 2022. *November 8, 2022 - General Election.* https://www.maine.gov/sos/cec/elec/results/2022/2022GeneralElectionRankedChoiceOffices.html.

Mattei, N., and T. Walsh. 2013. "PrefLib: A Library for Preferences http://www.preflib.org." In *Algorithmic Decision Theory,* edited by P. Perny, M. Pirlot, and A. Tsoukiàs, 259–270. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. ISBN: 978-3-642-41575-3. https://doi.org/10.1007/978-3-642-41575-3_20.

New York City. 2021. *Election Results Summary 2021.* https://www.vote.nyc/page/election-results-summary-2021.

Nicholson, W. K. 1995. *Linear Algebra with Applications.* 3rd ed. PWS Publishing Company.

Sundaram, R. K. 1996. *A First Course in Optimization Theory.* Cambridge University Press.