Fuller Disclosure than Intended

Joe Otten
Email: joe@datator.co.uk

1 Introduction

The full disclosure of preferences in the case of an STV election carries one danger of abuse. That is the potential for a unique preference list to identify a particular voter. Suppose there are 10 candidates in an election. Then there are $10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 = 3\ 628\ 800$ possible complete preference lists as well as a number of incomplete lists. In an electorate of a few tens or hundreds of thousands, it is obvious that the vast majority of the possible preference lists will not be used.

Of the preference lists that are used, they will generally follow some sort of pattern, such as the candidates of one party, followed by the candidates of another party, etc. It will therefore be fairly easy to create a large number of different preference lists that favour a particular candidate (with first preferences), and are most unlikely to be used by any voter.

2 The problem

The full disclosure of preference data facilitates the following fraud: The fraudster bribes or coerces a large number of voters to vote according to an exact preference list that is provided, and is different for each voter. The preference lists provided will be different unlikely sequences, such as the preferred candidate followed by alternate liberals and fascists or conservatives and communists.

Disclosure of the full preference data will then disclose, with a high probability, the voting behaviour of the bribed voters. There may be some false positives, but there will be no false negatives — i.e. if a preference list is missing then it is certain that a bribed voter welched.

3 The solution

One solution has been proposed — that of anonymising the preference data in a similar way to how census data is anonymised. Changes are made to the individual records in such a way as to minimise changes that result to any statistical aggregates an analyst might be interested in. The problem with this is that the statistical analysis of preference data is in such infancy that it is not clear what aggregates should be preserved, or how they might be preserved.

My preferred solution is that prior to disclosure, preference lists should be aggregated by censoring lower preferences until there are at least, say, 3 instances of every preference list to be published. So for example, if there are 10 votes of ABCDEFG then that fact can be published. If there is 1 vote of BCDEFGA, 1 of BCDEFAG and 1 of BCDEGAF then the fact that there were 3 votes of BCDExxx would be published. This would mean that no single individual's vote would be identifiably disclosed.