SUPPLEMENTARY INFORMATION

"Measuring Portfolio Salience Using the Bradley-Terry Model for Paired Comparisons"

Cesar Zucco, Mariana Batista, and Timothy Power

The Bradley-Terry Model

What follows is the textbook derivation of the Bradley-Terry model, which is available from multiple sources. We start by defining $\pi_i j$ as the probability that item i is chosen over item j in a pairwise comparison:

$$\pi_{ij} = \frac{\alpha_i}{\alpha_i + \alpha_j} \quad (1)$$

We then parametrize the probability as function of $\exp(\alpha)$. Other choices of parametrization are possible, and have been employed over the years, but this original Bradley-Terry formulation leads to:

$$\pi_{ij} = \frac{\exp(\alpha_i)}{\exp(\alpha_i) + \exp(\alpha_j)} \quad (2)$$

We then express $\pi_i j$ a log-odds, with a logit transformation

$$logit(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \log\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = \log(\pi_{ij}) - \log(\pi_{ji}) \quad (3)$$

Finally, replacing 2 in 3, and with minimal algebra, we arrive at:

$$\begin{aligned} logit(\pi_{ij}) &= \log\left(\frac{\exp(\alpha_i)}{\exp(\alpha_i) + \exp(\alpha_j)}\right) - \log\left(\frac{\exp(\alpha_j)}{\exp(\alpha_i) + \exp(\alpha_j)}\right) \\ &= \log(\exp(\alpha_i)) - \log\left(\exp(\alpha_j)\right) \quad (4) \end{aligned}$$

If we then define $\lambda_i = \exp(\alpha_i)$, we get

$$logit(\pi_{ij}) = \lambda_i - \lambda_j \quad (5)$$

Which is the logit formulation of the problem as a function of the exponential transformation of the underlying worth parameters.

We employ this simplest BT specification, referred to as the "unstructured" model. Numerous extensions have been proposed. These include allowing for ties Davidson (1970), allowing for rankings of more than two options, accounting for the order of the items in the pairwise contest, incorporating referee-specific covariates (Böckenholt 2001) or referee fixed effects (Cattelan 2012),1 factoring in the "margin of victory" of one item over another (New York Times 1979), incorporating contest-specific variables such as experience of the items/players (Stuart-Fox et al. 2006), or even allowing abilities to be modeled as a function

[1] Referee fixed effects can be estimated only if all respondents rate all items.

of item-specific covariates (Springall 1973). These extensions are unnecessary to us, as we intend simply to retrieve the underlying relative valuations of portfolios. Moreover, keeping the survey instrument as simple as possible is of utmost importance when surveying legislators.

Wording of Question in Surveys

The wording varied in very marginally between the expert and elite surveys. In the expert survey the question was worded as follows: "Imagine a hypothetical situation in which a future President of the Republic is sounding out a political party about taking on a cabinet-level position in the government. For each one of the pairs below, select the position you think a typical federal legislator would prefer."

In the elite Survey, the wording was: "Imagine a hypothetical situation in which a future President of the Republic is sounding out your party about taking on a cabinet-level position in the government. For each one of the pairs below, select the position you think your party would prefer."

Robustness to deviations from perfect survey randomization

One limitation of our elite survey is that because it was partially conducted using pen and paper, we could not fully implement the randomization of contests being evaluated. Although we circulated versions of the survey that included different sets of pairwise comparisons of ministries, fewer contests were presented more frequently than in the fully randomized electronic version of the survey. There were 221 different contests in the full legislator dataset. In the electronic subset of the data, 179 contests happened only once, 21 twice, and only three appear three or four times. In the manual version, there were a total of only 19 different contests,2 which were repeated at least 12 times.

This deviation from randomization is less of a problem that it might seem at first. Each additional contest adds little to the estimated worth for two reasons. First, what matters in the model is the probability of victory in a certain contest, and not the number of victories, so increasing the N has limited impact. Moreover, any changes in probabilities that occur with an increase in N are typically in the tails of the distribution, and these tend to matter less because the model implies a logit link function (as defined in Equation 2). Second, recall the way in which the model handles transitivity: the value of "defeating" another item depends on the whole set of interactions between all items. Therefore, additional matchups between the same items have progressively less impact on the estimates of α.

To test these intuitions, we simulated the effect of eliminating the excess repeated contests that were generated by the manual survey. We did this by implementing a form of non-parametric bootstrapping in which we sampled only four of each of the 19 high-frequency contests that were included in the manual survey.3 We appended the drawn contests to all the contests from the online portion of the legislative survey, and estimated the same BT model as before. We repeated this procedure for 1,000 different samples and then took the average point estimate across all simulations to produce a new set of estimates.

---

[2] Although the four contests in each of the five versions of the paper questionnaire were randomly selected, by pure chance one matchup appeared on two versions.

[3] We chose four to allow for ties, but results are identical with sampling fewer matchups.

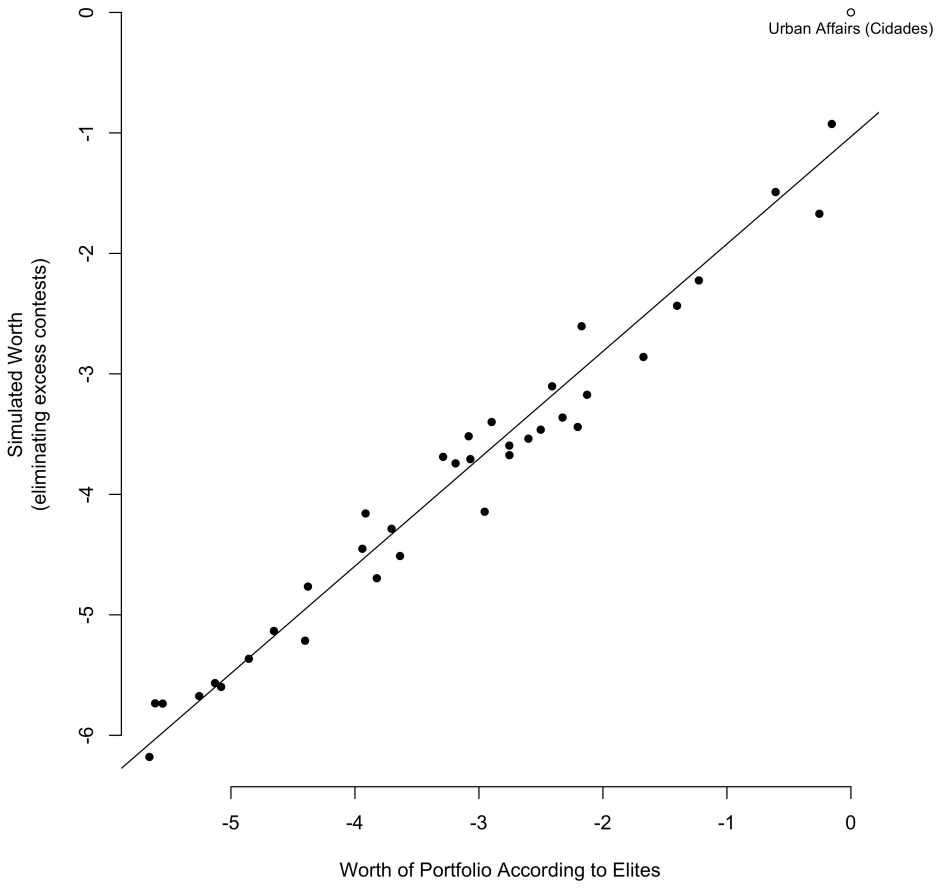

Figure A1: Comparison of Full Sample and Simulated Reduced Sample Estimates

The correlation between the worth of portfolios derived from these 1,000 simulations and our original estimates using the legislator survey (reported in Table 1) is a whopping 0.98, which means that the impact of the excess contests is trivial. Moreover, if we do the same simulations drawing fewer than four contests, the correlation coefficients with the original estimates change only marginally; they are never lower than 0.94, even if we draw only one occurrence of each matchup. And, if anything, as Figure 2 shows, Urban Affairs (Ministério das Cidades) would be even more valuable than what we estimated with the full dataset, which implies that the surprise valuation of this portfolio was not driven by the randomization process. Two other significant outliers are Institutional Relations and Science and Technology; the latter, as we observed above, was already known to be one of the noisier estimates in the set. The bottom line here is that the number of repetitions of each contest matters relatively little to the stability of the estimates.

Sources of Observable Attributes of Ministries Used in Table 2

Appointees is raw number of discretionary DAS positions (levels 1-6) allocated to the ministry, obtained from the Ministry of Planning; Policy Influence is the number of legislative bills signed by the ministry, obtained from the Library of the Presidency; Total Budget and Investment Budget are the value in Brazilian Reais of the authorized total and investment budget of the ministry, obtained from the National Secretary of the Treasury.

Cabinet Values and Confidence Intervals for Weights in Figure 2

In Figure 2, in the main body of the paper, we present coalescence and formateur advantage indicators for 1995 through 2016 as well as "confidence intervals" about these raw (unweighted) figures.

Our survey only rated ministries posts that existed between 2014-2016, so extending the weights to the earlier period required some transformations and assumptions. The period covered by our survey corresponds to the one with the largest number of most cabinet posts. As such, most ministries that existed before combined functions that were later disaggregated. In order to combine the weights of "smaller" ministries into the larger ones, we assumed that the value of each ministry corresponds to a fixed value of 10, which is equal to all ministries, plus a variable component. Applying this logic to the values of our rated ministries (between 10 and 100), allowed us to compute equivalent values for preexisting positions.

The confidence intervals in the Figure were designed to allow us to infer whether the weighted figures that our produce would fall within the range that could be produced by randomly generated weights. We computed these confidence intervals as follows: For each month we take the observed legislative seat shares and cabinet distribution for each party. We then randomly draw 1000 weights for each cabinet position. In the non-parametric version of the confidence intervals, these weights are drawn with replacement from the actual set of BT-weights for the ministries that existed in each month. In the parametric version, weights are drawn from a uniform distribution with the same limits of our sets of weights (the two versions generate almost identical results). For each set of weights that was drawn we compute our indicators of interest (coalescence and formateur advantage), and then take the 5% and 95% percentile across all draws as the bounds of the indicator.

References

Davidson, Roger R. 1970. "On extending the Bradley-Terry model to accommodate ties in paired comparison experiments." *Journal of the American Statistical Association* 65: 317–328.

Cattelan, Manuela. 2012. "Models for paired comparison data: A review with emphasis on dependent data." *Statistical Science* 27(3): 412–433.

New York Times. 1979. "Times' computer ranks football Top 20." October 19, p. A25.

Stuart-Fox, Devi M., David Firth, Adnan Moussalli, and Martin J Whiting. 2006. "Multiple signals in chameleon contests: designing and analysing animal contests as a tournament." *Animal Behaviour* 71(6): 1263–1271.

Springall, A. 1973. "Response surface fitting using a generalization of the Bradley-Terry paired comparison model." *Applied Statistics* 22(1): 59–68.