Heuristics on submitter deposit size for curated lists using Kleros

Response to: https://medium.com/wireline/when-can-token-curated-registries-actually-work-%C2%B9-2ad908653aaf

Building on a heuristic used in that post: for a price of e in effort, an observer can obtain knowledge of the "correct" ruling (for us the ruling that a large Kleros jury would eventually choose) as follows - with probability p the user obtains the right "correct" ruling, and with probability $1-p$ they are convinced that the opposite ruling is correct. (For this to be useful one needs $p > 1/2$.) These notes essentially follow similar ideas to those of the Medium post - with the same ultimate goal of looking at the size of required submitter deposits - in the context of Kleros, which due to juror deposits, etc is somewhat different/more complicated.

Fix the following notation:

- p is the probability that you correctly learn whether a submission belongs on the list or not after making an effort of cost e
- t is the larger of the percentage of entries rejected from the list and entries accepted to the list, (i.e. t is the percentage taken by the dominant outcome, $t \geq 1/2$).
- we are in an initial Kleros round with M jurors
- jurors place a deposit of d with incoherent jurors losing their deposits to coherent jurors
- submitters place a deposit of $S = S_J + S_C$, where if the submitter loses, S_J is distributed among the coherent first-round jurors and S_C is given to the challenger

So a juror that votes with the majority receives

$$\frac{S_J + d \cdot \#\text{ incoherent jurors}}{\#\text{ coherent jurors}},$$

while an incoherent juror loses d.

The Medium post considers two strategies for jurors

- making the effort to try to vote honestly and

- voting randomly with 50% probability for each choice.

In fact, better than the completely random strategy, a lazy juror can vote with the dominant outcome - if most entries are rejected from the list, always vote to reject; if most entries are accepted to the list, always vote to accept - and be right with probability $t > 1/2$. Also in Kleros, as a juror can lose a deposit by voting incoherently, it is no longer the case that a juror is necessarily incentivized to participate. Hence, here we consider the following three strategies:

- "honest effort stragey": making the effort to try to vote honestly and
- "lazy strategy": always vote the most frequently occurring outcome - will be right with probability t
- "opt-out strategy": un-stake PNK, do not participate

Parameter choices need to be made so that the first strategy of making an honest effort is an equilibrium (note that the honest effort strategy cannot be dominant as if everyone takes the lazy strategy of always voting for the most common answer that strategy is an equilibrium). Note that this is only possible if $p > t$.

Suppose all of the other $M-1$ participants make the honest effort at cost e. Then the number of coherent votes X from among these jurors is distributed as $X\ Binom(M-1, p)$. So

$$E[\text{honest effort}] = p \cdot E\left[\frac{S_j + d(M-X-1)}{X+1}\right] - (1-p)d - e$$

$$= p \cdot (S_j + dM) E\left[\frac{1}{X+1}\right] - pd - (1-p)d - e$$

$$= p(S_J + dM) \cdot \frac{1}{Mp}(1-(1-p)^M) - d - e,$$

(where we have used the standard calculation of $E\left[\frac{1}{X+1}\right]$ when X is binomial)

$$= \frac{S_J + dM}{M}(1-(1-p)^M) - d - e.$$

On the other hand,

$$E[\text{lazy strategy}] = t \cdot E\left[\frac{S_j + d(M-X-1)}{X+1}\right] - (1-t)d$$

$$= \frac{(S_J + dM)t}{Mp}(1-(1-p)^M) - d.$$

Of course,

$$E[\text{opt-out}] = 0.$$

So for the honest effort strategy to give the highest expected value, we need

$$\frac{S_J + dM}{M}(1 - (1-p)^M)\left(1 - \frac{t}{p}\right) - e \geq 0,$$

and

$$\frac{S_J + dM}{M}(1 - (1-p)^M) - d - e \geq 0.$$

So if we want to set S_j in terms of the other parameters, we should choose

$$S_J \geq \max\left\{\frac{eM}{(1-(1-p)^M)\left(1-\frac{t}{p}\right)} - dM, \frac{(d+e)M}{1-(1-p)^M} - dM\right\}.$$

(One can think of e, p, t, and possibly M as structural constants that cannot be easily tuned. However, one expects the value of d PNK in ETH to depend on the choice of S_J as in the computations in: https://medium.com/kleros/why-kleros-needs-a-native-token-5c6c6e39cdfe - one could use such ideas to remove d from this inequality at the expense of introducing variables such as the prevailing interest rate, the number of disputes that are being arbitrated with Kleros, and the number of staked tokens.)

However, one must also choose S_C to be large enough to encourage challenges. The Medium article considers situations where there is a single challenger to avoid the honest unity problem, so we will make the same assumption here for the moment.

Suppose a challenger must pay a deposit of $S' = S_J + S'_C$ and has the same possibility of determining the true result of an eventual dispute with probability p at cost e. Suppose that a proportion of u of all submissions do not belong on the list.

Then the possible outcomes of any time the challenger reviews an item submitted for the list are:

- Correctly identifies that a submission does not belong on the list
  - probability pu
  - payoff $S_C - e$
- Incorrectly comes to conclusion that a submission that doesn't belong on the list belongs there, doesn't challenge
  - probability $(1-p)u$
  - payoff $-e$
- Correctly identifies that a submission does belongs on the list, doesn't challenge
  - probability $p(1-u)$
  - payoff $-e$

- Incorrectly comes to conclusion that a submission doesn't belong on the list when it does, challenges and loses
  - probability $(1-p)(1-u)$
  - payoff $-S_J - S'_C - e$

Note a consequence of this - if $S_C = S'_C$, namely if the submitter and the challenger place the same deposit - then for the challengers to be incentivized, one must have

$$(1-p)(1-u) < pu \Leftrightarrow u + p > 1.$$

Otherwise, the challenger will lose more money from the false positives of incorrectly flagging submissions that belong on the list than she will gain by correctly challenging false submissions.

For the moment, for simplicity, we consider the $S_C = S'_C$ case and assume $u + p > 1$. Then the challenger is incentivized to participate if

$$pu(S_C - e) - (1-p)ue - p(1-u)e + (1 - p(1-u)(-S_J - S_C - e) \geq 0$$

$$\Leftrightarrow S_C \geq \frac{e + (1-p)(1-u)S_J}{u + p - 1}.$$

Then one should have

$$S = S_J + S_C \geq S_J + \frac{e + (1-p)(1-u)}{u + p - 1} S_J$$

$$= S_J \left(1 + \frac{e + (1-p)(1-u)}{u + p - 1}\right)$$

$$\geq \left(1 + \frac{e + (1-p)(1-u)}{u + p - 1}\right) \cdot \max\left\{\frac{eM}{(1-(1-p)^M)\left(1 - \frac{t}{p}\right)} - dM, \frac{(d+e)M}{1-(1-p)^M} - dM\right\}.$$

Note that this is a deposit; so the true cost of a submission to a submitter should also depend on the probability that probability that a submission made in good faith is rejected - see the notes on the appeal fees/future work.

Example 1. *Suppose* $e = 10$, $p = .8$, $t = .6$, $u = .3$, $M = 7$, $d = 50$. *Then one will need* $S_J \geq 70.0$ *and* $S \geq 1050.1$.