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

$$\text{Challenger payoff} =$$

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

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

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* $S \geq 1050.1$.

1 Attack and challenger evaluation rates in equilibrium

In the previous section, we consider what u is necessary in order to incentivize the challenger to evaluate each submission. In this section, we instead consider

challengers that are willing to adopt a mixed strategy, of evaluating some of the submissions, and we consider what kind of equilibria we can expect.

Suppose that the value of placing a malicious entry on the list for an attacker Eve is V. Further, suppose that the challenger takes the strategy of randomly drawing y percent of all submissions and evaluating them to determine whether to challenge or not. An attacker that attempts to make a submission that does not belong on the list can have the following outcomes based on whether the behavior of the challenger:

- Challenger evaluates whether the attacker's submission belongs on the list, correctly concludes that it does not
  - probability yp
  - payoff to attacker $-S_J - S_C$
- Challenger evaluates the attacker's submission and incorrectly comes to conclusion that it belongs on the list, doesn't challenge
  - probability $y(1-p)$
  - payoff to attacker V
- Challenger does not evaluate attacker's submission
  - probability $1-y$
  - payoff to attacker V

Then the attacker's payoff function is

$$\text{Attacker payoff} = yp(-S_J - S_C) + (1-yp)V = yp(-S_J - S_C - V) + V.$$

Again, we have

$$\begin{aligned}\text{Challenger payoff} &= pu(S_C - e) - (1-p)ue - p(1-u)e + (1-p)(1-u)(-S_J - S'_C - e) \\ &= puS_C + (1-p)(1-u)(-S_J - S'_C) - e.\end{aligned}$$

We consider whether, for any values of S_J, S_C, S'_C, p, and e, the attacker of the challenger have a dominant strategy.

Eve has a dominant strategy to always attack or never attack if her payoff function is always non-negative or always non-positive respectively for all values of $y \in [0,1]$. At $y = 0$, her payoff is $V > 0$. Hence her only possible dominant strategy is to always attack. As her payoff function is linear in y, the function always taking non-negative values for $y \in [0,1]$ is equivalent to it taking a positive value at $y = 1$, i.e.:

$$p(-S_J - S_C - V) + V > 0 \Leftrightarrow S_C \leq \frac{(1-p)V - pS_J}{p}.$$

Similarly, the challenger has a dominant strategy to always evaluate or never evaluate if his payoff function is always non-negative or always non-positive

respectively for all values of $u \in [0,1]$. At $u = 0$, his payoff is $(1-p)(-S_J - S'_C) - e < 0$. Hence the only possible dominant strategy is to never evaluate. Then as

$$\begin{aligned}\text{Challenger payoff} &= pu(S_C - e) - (1-p)ue - p(1-u)e + (1-p)(1-u)(-S_J - S'_C - e) \\ &= puS_C + (1-p)(1-u)(-S_J - S'_C) - e \\ &= u\left(pS_C + (1-p)(S_J + S'_C)\right) + (S_J + S'_C) - e,\end{aligned}$$

which again is linear, the payoff function will be non-positive for all values of $u \in [0,1]$ if and only if it takes a non-positive value at $u = 1$. Namely, the challenger has a dominant strategy to not evaluate if and only if

$$pS_C - e \leq 0 \Leftrightarrow S_C \leq \frac{e}{S_C}.$$

Thus, if

$$S_C \geq \max\left\{\frac{(1-p)V - pS_J}{p}, \frac{e}{p}\right\},$$

neither side has a dominant strategy.

Then, in equilibrium, in order for both parties to be willing to randomize their strategies, we have:

$$yp(-S_J - S_C - V) + V = 0 \Leftrightarrow y = \frac{V}{p(S_C + S_J + V)},$$

and

$$puS_C + (1-p)(1-u)(-S_J - S'_C) - e = 0 \Leftrightarrow u = \frac{(1-p)(S_J + S'_C) + e}{pS_C + (1-p)(S_J + S'_C)}.$$

Remark 1. *Note some submitters may make submissions to the list that they believe belong but that ultimately would be rejected by the jurors. In this model, we have conflated such people with the attacker Eve. Whatever percentage of submissions this represents, one would expect the "true attackers" to adjust to that to reattain the equilibrium. If there are enough honest but wrong submitters to represent a value of u that already exceeds the equilibrium value, the challenger would be incentivized to take a pure strategy of always evaluating.*

Note that the requirement that $S_C \geq \frac{(1-p)V - pS_J}{p}$ implies that (assuming parameters are not tuned in a way that depends on honest but wrong submitters as in Remark 1) one must take the total deposit for the submitter

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

Then, we can think of the percentage of the list that will consist of hostile submissions that do not get challenged as

$$u(1-yp) = \begin{cases} \frac{(1-p)(S_J+S'_C)+e}{pS_C+(1-p)(S_J+S'_C)} \cdot \left(\frac{S_C+S_J}{S_C+S_J+V}\right) & : S_C \geq \max\left\{\frac{(1-p)V}{p} - S_J, \frac{e}{p}\right\} \\ 1 & : \frac{(1-p)V}{p} - S_J \leq S_C < \frac{e}{p} \\ 1-p & : \frac{e}{p} \leq S_C < \frac{(1-p)V}{p} - S_J \\ 1 & : S_C < \min\left\{\frac{(1-p)V}{p} - S_J, \frac{e}{p}\right\} \end{cases}$$

Note that the presence of the $\frac{e}{p} \leq S_C < \frac{(1-p)V}{p} - S_J$, $u(1-yp) = 1-p$ case is something of an artifact ot the assumption that we only have one challenger. Then we are limited by the probability that that challenger tries to evaluate a case and reaches an incorrect conclusion, even if the challenger is always incentivized to participate.

2 Multiple challengers

We now take the model that we have K challengers, each that can obtain the correct judgment of a submission with probability p by exerting effort e. The probability of any two challengers correctly assessing a given submission is assumed to be independent.

When multiple challengers come to the conclusion that they want to challenge the same submission, whoever is first does so, while the others wasted their effort e but keep their deposit. Depending on the application, we expect that the amount of time required to assess and challenge a submission will vary enough (though we do not include this in how we model e) that it will not generally be worthwhile to pay high gas fees to have one's challenge included before others. So in our simplified model, when there are multiple challengers, they each have an equal chance of their challenge being selected.

Then the payoff for a challenger who decides to evaluate a given case is:

$$\text{Challenger payoff} = upS_C\frac{1}{1+X} + (1-u)(1-p)(-S_J - S'_C)\frac{1}{1+Z} - e,$$

where X is distributed as Binomial$(K-1, yp)$ and Z is distributed as Binomial$(K-1, y(1-p))$.

Then

$$E[\text{Challenger payoff}] = upS_C\frac{1-(1-py)^K}{Kpy} + (1-u)(1-p)(-S_J - S'_C)\frac{1-(1-(1-p)y)^K}{K(1-p)y} - e$$

$$= \frac{uS_C}{Ky}(1-(1-py)^K) + \frac{(1-u)(-S_J-S'_C)}{Ky}(1-(1-(1-p)y)^K) - e.$$

Meanwhile, the attacker payoff for making a hostile submission is given by:

$$E[\text{Attacker payoff}] = \left[1-(1-yp)^K\right](-S_J - S_C) + (1-yp)^K V$$

$$= \left[1 - (1 - yp)^K\right](-S_J - S_C - V) + V.$$

The attacker has no dominant strategy that she should employ regardless of the strategy of the challengers: indeed, if $y = 0$ the attacker's payoff is $V > 0$, so the only possible dominant strategy would be to always attack. However, if $y = 1$, the attacker's payoff is $\left[1 - (1 - p)^K\right] + (-S_J - S_C - V) + V$, which approaches $-S_J - S_C < 0$ for sufficiently large K.

Similarly, if $u = 0$, the challenger's payoff is given by

$$\frac{(-S_J - S'_C)}{Ky}(1 - (1 - (1 - p)y)^K) - e < 0.$$

Hence the only possible dominant strategy for the challenger is to never evaluate cases. However, if $u = 1$ the challenger has a payoff of

$$pS_C \frac{1 - (1 - py)^K}{Kpy} - e.$$

If

$$S_C \geq \frac{e}{p},$$

this will result in positive payouts for some choices of y and K (specifically, when $K = 1$). Hence, under this assumption, neither the attacker nor the challenger has a dominant strategy.

So, in equilibrium,

$$(1 - yp)^K = \frac{S_J + S_C}{S_J + S_C + V} \Rightarrow y = \frac{1}{p}\left(1 - \sqrt[K]{\frac{S_J + S_C}{S_J + S_C + V}}\right)$$

and

$$u = \frac{Kye + (S_J + S'_C)(1 - (1 - (1 - p)y)^K)}{S_C(1 - (1 - py)^K) + (S_J + S'_C)(1 - (1 - (1 - p)y)^K)}.$$

Then the percentage ultimate entries to the list that is composed of hostile submissions that went unchallenged is:

$$u(1 - yp)^K$$

$$= \frac{Kye + (S_J + S'_C)(1 - (1 - (1 - p)y)^K)}{S_C(1 - (1 - py)^K) + (S_J + S'_C)(1 - (1 - (1 - p)y)^K)} \cdot \frac{S_J + S_C}{S_J + S_C + V}$$

$$= \frac{Kye + (S_J + S'_C)(1 - (1 - (1 - p)y)^K)}{S_C\left(\frac{V}{S_J + S_C + V}\right) + (S_J + S'_C)(1 - (1 - (1 - p)y)^K)} \cdot \frac{S_J + S_C}{S_J + S_C + V}.$$

Example 2. *Again, we take $e = 10$, $p = .8$, $S_J = 70$. Suppose we have $K = 10$. Now instead of assuming a fixed value of u, we find that in equilibrium $y = .2365$ and $u = .0825$. This means that the percentage of the list that consisting of hostile submissions that went unchallenged is* .0101.

Remark 2. *In the model of the token[2] curated list, where entries that have made it onto the list can be later challenged, one can imagine that challengers will patrol the list/long-established entries with a lower y than the candidates to be added due to the lower percentage of hostile submissions. – May approach some limit of long-term percentage of malicious submissions that survive/to think about.*

2.1 Estimates on $u(1-yp)^K$

–May come up with better estimates, not clear how tight these are so far.

Note that

$$Kye = K\frac{e}{p}\left(1 - \sqrt[K]{\frac{S_J+S_C}{S_J+S_C+V}}\right)$$

is a constant multiple of a function of the form $f(K) = K\left(1 - \sqrt[K]{z}\right)$. Functions of this form are monotonically increasing as $f'(K) = 1 + z^{1/K}\left[\ln z^{1/K} - 1\right] \geq 0$ (as a standard argument shows that $x(\ln x - 1)$ takes a global minimum of 0 at $x = 1$). Then

$$f(K) \leq \lim_{K\to\infty} K(1 - \sqrt[K]{z}) = \ln(1/z).$$

Hence

$$Kye \leq \frac{e}{p}\ln\left(\frac{S_J+S_C+V}{S_J+S_C}\right).$$

We will also find upper and lower bounds on $(1-(1-(1-p)y)^K)$. To this end, note that

$$[1-y(1-p)]^K = \left[1 - \left(1 - \sqrt[K]{\frac{S_J+S_C}{S_J+S_C+V}}\right)\cdot\frac{1-p}{p}\right]^K = \left[\left(2-\frac{1}{p}\right) + \left(\frac{1-p}{p}\right)\sqrt[K]{\frac{S_J+S_C}{S_J+S_C+V}}\right]^K$$

is of the form $((1-b)+b\sqrt[K]{c})^K$ for $b, c \in [0,1]$. Here we have used the assumption that $p > 1/2$.

Note that, if $f(K) = ((1-b)+b\sqrt[K]{c})^K$, then

$$f'(K) = ((1-b)+b\sqrt[K]{c})^K\left[\ln((1-b)+b\sqrt[K]{c}) - \frac{b\ln c\cdot\sqrt[K]{c}}{K((1-b)+b\sqrt[K]{c})}\right].$$

Then, if $(1-b)+b\sqrt[K]{c}$, b, $c \in [0,1]$, $f(K)$ is monotonically decreasing where we use

$$\ln((1-b)+b\sqrt[K]{c}) - \frac{b\ln c\cdot\sqrt[K]{c}}{K((1-b)+b\sqrt[K]{c})} \leq 0 \Leftrightarrow ((1-b)+b\sqrt[K]{c})\ln((1-b)+b\sqrt[K]{c}) \leq b\ln(\sqrt[K]{c})\sqrt[K]{c},$$

which holds by the convexity of the function $g(x) = x\ln x$ between $x = \sqrt[K]{c}$ and $x = 1$.

By our discussion on the non-existence of a pure strategy of non-participation for the challenger(s), we can assume $K \geq 1$. Thus,

$$\lim_{K\to\infty}\left[1-\left(1-\sqrt[K]{\frac{S_J+S_C}{S_J+S_C+V}}\right)\cdot\frac{1-p}{p}\right]^K \leq [1-y(1-p)]^K \leq \left[1-\left(1-\frac{S_J+S_C}{S_J+S_C+V}\right)\cdot\frac{1-p}{p}\right]^1.$$

A standard computation shows, for a, b, $c \in (0,1)$:

$$\lim_{K\to\infty}(a+b\sqrt[K]{c})^K = c^b.$$

So

$$\left(\frac{S_J+S_C}{S_J+S_C+V}\right)^{(1/p-1)} \leq [1-y(1-p)]^K \leq 1-\left(\frac{V}{S_J+S_C+V}\frac{1-p}{p}\right).$$

Then, the percentage of the list that we expect to consist of hostile submissions that went unchallenged is:

$$u(1-yp)^K$$

$$\leq \frac{\frac{e}{p}\ln\left(\frac{S_J+S_C+V}{S_J+S_C}\right)+(S_J+S'_C)\left(1-\left(\frac{S_J+S_C}{S_J+S_C+V}\right)^{(1/p-1)}\right)}{S_C\left(\frac{V}{S_J+S_C+V}\right)+(S_J+S'_C)\left(\frac{V}{S_J+S_C+V}\frac{1-p}{p}\right)}\frac{S_J+S_C}{S_J+S_C+V}.$$

- to be simplified

Remark 3. *We consider the realism of our model of challenger behavior. In general, if there are K people participating as challengers, one would not expect them to be all actively online at any given time. Hence the probability that a given challenger loses the cost of his effort while failing to submit his challenge because some other challenger has already done so should reflect a K that is number of challengers actively participating at that given time. However, challengers may choose cases that have already been evaluated, and decided to be not worth challenging, by others. This has the effect that the effective rate of dishonest submissions in the pool from the point of view of challengers is lower than the u that is the rate of dishonest submissions originally submitted. Hence we can think of the real situation as being somewhat intermediately profitable to challengers compared to considering K to be the total number of challengers versus considering it to be the number of challengers online at a given time.*