Draft: Fees in Kleros

1 Juror rewards

We consider a model similar to [4].

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 $p > 1/2$.)

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
- a total of S_J is awarded to the coherent jurors in this arbitration round

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.

In [4], there are two strategies considered 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 $X \sim Binom(M-1,p)$. 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 [2] - 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.)

Remark 1. *Note that*

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

Hence, the bound on S_J becomes

$$S_J \geq \max\left\{ \begin{array}{lcl} \frac{eM}{(1-(1-p)^M)\left(1-\frac{t}{p}\right)} - dM & : & d \leq e\left(\frac{1}{1-\frac{t}{p}} - 1\right) \\ \frac{(d+e)M}{1-(1-p)^M} - dM & : & d > e\left(\frac{1}{1-\frac{t}{p}} - 1\right) \end{array} \right\}.$$

If one wanted to find the minimum S_J for which this bound can hold over the choice of the parameters that we can set, that would mean minimizing

$$g(d) = \max\left\{ \begin{array}{lcl} \frac{eM}{(1-(1-p)^M)\left(1-\frac{t}{p}\right)} - dM & : & d \leq e\left(\frac{1}{1-\frac{t}{p}} - 1\right) \\ \frac{(d+e)M}{1-(1-p)^M} - dM & : & d > e\left(\frac{1}{1-\frac{t}{p}} - 1\right) \end{array} \right\}.$$

However,

$$g'(d) = \max\left\{ \begin{array}{lcl} -M & : & d \leq e\left(\frac{1}{1-\frac{t}{p}} - 1\right) \\ \frac{M}{1-(1-p)^M} - M & : & d > e\left(\frac{1}{1-\frac{t}{p}} - 1\right) \end{array} \right\}.$$

As $\frac{1}{1-(1-p)^M} - 1 > 0$ for $p \in (1/2, 1)$, this minimum is achieved at

$$d = e\left(\frac{1}{1-\frac{t}{p}} - 1\right).$$

However, it will not necessarily be the case that one wants to minimize S_J; rather would-be jurors will have some supply curve for the price at which they are willing to provide arbitration services and one expects an equilibrium to depend on the demand curve of those who would submit to the list, as a function of the price they are willing to pay (in terms of capital lock-up costs if nothing else).

However, this discussion should be thought of as giving a minimum level of fees S_J that can be set in terms of e, t, and p.

In practice, one would probably want to be conservative in using this minimum S_J, as if different jurors have different values of e, a choice of S_J that prices some of them out of the honest strategy could result in the value of t changing, potentially leading to a vicious cycle where more and more jurors abandon the honest strategy.

2 Deposits for Kleros curated lists

We consider a model where, a submitter can submit an entry to a curated list, which can be challenged by a challenger.

In order to make a submission to a curated list, the submitter must place a deposit of deposit of $S = S_J + S_C$. Here, if the submitter loses, S_J is distributed among the coherent first-round jurors as in Section 1 and S_C is given to the challenger

In this section, we will consider requirements on the choice of S_C. Specifically, one must 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

$$\begin{aligned} S = S_J + S_C &\geq S_J + \frac{e + (1-p)(1-u)}{u+p-1} S_J \\ &= S_J \left(1 + \frac{e + (1-p)(1-u)}{u+p-1}\right) \end{aligned}$$
$$\geq \left(1 + \frac{e + (1-p)(1-u)}{u+p-1}\right) \cdot \max\left\{\frac{eM}{(1-(1-p)^M)\left(1-\frac{t}{p}\right)} - dM, \frac{(d+e)M}{1-(1-p)^M} - dM\right\}.$$

Note that this is a deposit; so the true cost of a submission to a submitter should also depend on the probability that 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 $S \geq 1050.1$.*

2.1 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 or 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 2. *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 2) 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.

2.2 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 $u = .67223$. This means that the percentage of the list that consisting of hostile submissions that went unchallenged is $.0825$.*

Remark 3. *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.*

2.2.1 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:

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

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

- to be simplified

Remark 4. *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.*

Remark 5. *Note, one might want to fix a given security level $u(1-yp)^K$, and then adjust the parameters S_J, S_C, and S'_C to minimize the deposit that must be paid by the submitter S_J+S_C. Note by the structure of Equation 1, any increase in S_J that is accompanied by an equal decrease in S_C increases $u(1-yp)^K$ hence weakening the security level.*

This is as

- *y, and indeed any term where S_J and S_C appear as $\frac{S_J+S_C}{S_J+S_C+V}$ is unchanged*
- *one can write*

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

  , and as $p>1/2$, $\frac{1-p}{p}<1$.

Hence, an increase in S_J accompanied by an equal decrease in S_C increases the numerator and decreases the denominator of this bound.

3 Appeal structure and fees

We analyze fees that should be paid in Kleros between the parties to (binary) disputes. Typically, we will have two parties which we denote by Alice and Bob. Unless indicated otherwise, we consider the situation where Alice has one the most recent ruling of the dispute.

Let x be the value that must be paid in arbitration fees for the next round. A first observation is that each of Alice and Bob must pay at least x in appeal fees if the appeal fees are to be refunded to the winning party. Moreover, any scheme where an insurer pays arbitration fees for an honest parties in which successful insurers are paid out of the fees returned to the winner and (are not paid out of the value at stake in the dispute), the winning party must in fact receive more than what she put in for arbitration fees.

Denote:

- s_A - the additional "stake" Alice must pay in case of an appeal to not lose the case, namely Alice must pay a total of $x + s_A$
- s_B - the "stake" Bob must pay in case of appeal to not lose the case, namely Bob must pay a total of $x + s_B$.

3.0.1 Structure of proposed insurance mechanism

Remember that in order to appeal, Bob must pay $x + s_B$ and then Alice must pay $x+s_A$ to not forfeit the appeal. We detail a sort of crowd-sourced insurance mechanism that can cover these fees:

- The party (Alice or Bob) might be given the opportunity to pay these fees directly/themselves; if a party refuses to do so and/or a fixed period of time elapses where she hasn't, then the fees can be "adopted/crowd-funded" as follows:
  - Any user can pay some percentage of the required fees x. Denote $s_r = \frac{\text{contribution of } \mathcal{USR}_r}{x+s_{\text{party}}}$.
  - If less than $x + s_{\text{party}}$ is raised, everyone is refunded and that party's fees are not paid. (In Bob's case the dispute is not appealed; if Bob's fees are paid and Alice's are not, Alice loses the dispute.)
  - Once $x + s_{\text{party}}$ is raised, the contract stops accepting additional contributions.
  - If the other party ultimately wins the dispute (possibly after additional appeals), the adopters/crowdfunders lose their contributions.
  - If Bob wins the dispute as a result of the fees for his side being paid and Alice's fees not being paid in some round (i.e. if Bob wins in a round without a call to Kleros), then whoever pays Bob's fees in the final round receives z percent of the total losing side's stake across all appeal rounds, where z is some parameter to be chosen. This is divided proportionally in the event that these fees were crowdfunded.
  - If the crowdfunded side of the dispute ultimately wins, each contributor $\mathcal{USR}_r$ receives back the contribution she paid towards $x+s_{\text{party}}$ and a corresponding portion of the losing party's stake given by $s_{\text{losing party}} \cdot s_r$ (or $(1-z) * s_{\text{losing party}} \cdot s_r$ in the event that the dispute ended in an appeal round which did not include a call to Kleros).

In Section 3.1 we will consider what are appropriate values for s_A and s_B. In Sections 3.2-3.4, we will consider how several edge cases should be handled.

3.1 Analysis assuming rational insurers whose total resources match that of any attack

Suppose that rational insurers exist who are willing to fund appeal fees when they estimate

$$E[\text{return on insurance}] > 0.$$

Note that if multiple parties contribute to the fees, the return will be divided proportionally, but this does not change whether this expected value is positive or negative, so without loss of generality we assume that there exists a single insurer with a given prior belief on Alice and Bob's respective winning chances, Isaac.

Remark 6. *In practice, if there are actually multiple insurers who are willing to pay a party's entire arbitration fee and are capable of making this decision quickly after the window to pay the fee has opened (for example, because they have already analyzed the case in advance), then it is possible that the insurers will get into a gas war among each other to have their insurance transaction included. This is not likely to be a worse problem than, for example, multiple parties rushing to challenge an image in the Doge pilot, so we will consider this effect to be negligible for the current work. However, as a sort of instance of the honest unity problem we may study these dynamics in future work.*

Suppose that rational actors (Isaac) evaluate Alice's chances of eventually winning at p_A. Then

$$E[\text{return on insurance}] = p_A(s_B)+(1-p_A)(-x-s_A) \geq 0 \Leftrightarrow p_A \geq \frac{x+s_A}{x+s_A+s_B}.$$

Denote this threshold by

$$t_A = \frac{x+s_A}{x+s_A+s_B}. \tag{2}$$

Then Isaac will only finance Alice's appeal if he estimates that $p_A \geq t_A$.

Similarly, rational actors would only finance Bob's appeal if

$$p_B(s_A) + (1-p_B)(-x-s_B) \geq 0 \Leftrightarrow p_B \geq t_B = \frac{x+s_B}{x+s_A+s_B}.$$

Then

$$t_B - t_A = \frac{s_B - s_A}{x+s_A+s_B}. \tag{3}$$

Remark 7. *As Alice and Bob are playing a negative sum game (the payment to the jurors is consuming part of their appeal fees), it is impossible to calibrate the stakes s_A and s_B such that perfectly rational insurers that evaluate Alice and Bob's winning chances in appeal both at 50% will fund the appeals of each. Indeed, we will see that there is some range of estimations of p_A in which is profitable to finance the appeal fees of neither Alice nor Bob. In practice as the evaluations of Alice and Bob's chances will vary in the population of insurers it is possible that they might both have their fees funded.*

We consider the consequences of a few possiblities for s_A and s_B:

3.1.1 Possibility 1: $s_A = s_B$

So note that if $s_A = s_B$, $t_A = t_B$ but

$$t_A = t_B = \frac{x + s_A}{x + 2s_A}.$$

So for example,

- if $s_A = 0$, $t_A = 1$ and insurers are never incentivized to fund appeals,
- if $s_A = x$, $t_A = 2/3$ (in particular if both Alice's and Bob's chances are estimated by Isaac to between $1/3$ and $2/3$, he is not incentivized to fund either of them), and
- as $s_A \to \infty$ for fixed x, the threshold $t_A = t_B$ tends to $1/2$.

In this setting, it may often be the case that an honest party Alice that won the previous round would not have a high enough expected return for insurers to have an incentive to fund her appeal.

3.1.2 Possibility 2: $t_A = 1/2$ (recommended)

Instead, suppose we want $t_A = 1/2$. Then, rearranging Equation 2 we must have

$$s_B = x + s_A.$$

Plugging this into Equation 3, we have

$$t_B = \frac{x}{2x + 2s_A} + 1/2.$$

Then again s_A is a parameter that could be tuned by the governance process. To illustrate several choices:

- if $s_A = 0$, then $s_B = x$ and $t_B = 1$. In this case, insurers are never incentivized to fund appeals for Bob,
- if $s_A = x$, then $s_B = 2x$ and $t_B = 3/4$, (these may be reasonable choices) and
- as $s_A \to \infty$ for fixed x, then s_B also tends to infinity and t_B tends to $1/2$.

So, in the context of taking $t_A = 1/2$ there is a basic tradeoff here between the size of the arbitration fees $x+s_A$ that we are willing to impose on Alice, the party that won the previous round (as well as the even higher fees $x + s_B = 2x + s_A$), versus the threshold probability t_B of an eventual Bob victory that would be required for insurers to be willing to fund his fees. As s_A tends to infinity for fixed x, t_B still tends to $1/2$.

We now address the question of how insurers might estimate p_A and p_B. Consider the related probability π_A - the probability that a randomly selected PNK will correspond to a juror that votes with Alice in an idealized setting where there are no attacks that influence jurors' votes and where this probability does not change from round to round (such as by new information becoming available). Similarly we denote by π_B the probability that a randomly selected PNK will correspond to a juror that votes with Bob under the same conditions.

3.2 Pre-funding fees

To simplify user experience, one can allow users to pre-fund fees for a given side in a future round of arbitration. Note that, typically if one or more insurers pay fees that overpay what is currently necessary, the first payment received should be used, and whatever difference should be refunded to the party that paid it. Hence, essentially, a distinction should be made between two types of fee payment transactions

- refundable overpay - for insurers that only want to pay the current round fees (and would potentially want to evaluate the result of the current round before paying fees in a future round)
- non-refundable overpay - for insurers (and potentially the parties themselves) that want to pre-pay funds for the following rounds that may be required.

3.3 Mid-appeal fee increases

Due to a governance decision, a arbitrator contract may change the required juror fees for a type of case while there is some ongoing dispute of that type.

Note that, as the amounts of stake considered in section 3.1 depend on the amount of arbitration fees for the corresponding rounds, then ideally the amount of stake would adjust accordingly to any adjustments in the arbitration fees. (Depending on the code complexity of doing such an adjustment, it may be acceptable to leave the values of stake unadjusted, as small changes in arbitration fees should have limited impact on the values of t_A and t_B.)

Depending on when this decision is made relative to a given appeal round, and whether the change in fees is an increase or a decrease, this could have different effects.

If the fee change is made before either party has paid their fees for a subsequent appeal, both of their required contributions can adjust accordingly.

If the governance process has decreased the fees required, then at worst one or both parties to the dispute will have contributed too much in fees and/or stake. Hence, this amount can simply be refunded.

However, imagine the case where a change is made increasing the arbitration costs after the previous round loser has already paid their fees, but before the previous round winner has. In this case, it is possible that, unless the fees for the previous round winner are adjusted, inadequate fees will be paid to cover the arbitration.

There are a range of ways of handling the eventuality based on how the cost of the increase is spread between the two parties.

The previous round winner can be made to pay

1. newAppealCost+oldStake

   - This concentrates all of the hit from the fee increase on the previous round winner, but guarantees that people that crowdsourced funding for the fees of the previous round loser are no worse off than they would have been had their been no fee increase.

2. max(oldAppealCost+oldStake, newAppealCost-(2oldAppealCost+oldStake))

   - This discourage people from trying to game fee increases by having both sides be worse off after a fee increase and minimizes the additional cost to the party that won the previous round. However, there is the risk that people who honestly funded the fees of the party that lost the previous round have their contribution eaten up by the fee increase, even if they wind up on the winning side.

3. (Recommended)

   max(oldAppealCost+oldStake,newAppealCost)

   - Here one minimizes the fee increase required of the previous round winner subject to the constraint that if the previous round loser wins, she is at least guaranteed to get back what she put in but not necessarily win any stake.

Ideally, if a governance change is made very near the time limit of a given fee payment period, that should extend the limit. Otherwise, fee changes that occur very near cut-offs will inevitably cause problems.

3.4 Ties, refusals to arbitrate, and non-decisive outcomes

Due to one or more jurors failing to vote, there is the possibility that an appeal round will end in a tie. Moreover, due to jurors voting “refuse to arbitrate,” there is the possibility that there will be a non-decisive outcome, hence there will be a future round in which there is not a “losing party” and a “winning party.”

In this case, the parties should be required to pay the arbitration fees plus the same stake, namely we are in the symmetrical case of Section 3.1.1. If the recommended fees of $s_A = x$, $s_B = 2x$ as in Section 3.1.2 are used for appeal rounds after decisive rounds, then for appeal rounds after indecisive rounds, $s_A = s_B = x$, $t_A = t_B = 2/3$ seems to be a reasonable choice. Namely both sides are asked to pay what the winning side who have had to pay had the previous round been decisive.

Furthermore, due to the symmetry of this situation, both sides should pay their fees in the same payment period. While there is still the potential that a governance change will raise fees during the (common) payment period, this removes the possibility of a change after one party has paid its fees but before the other has, as in Section 3.3.

If the dispute ends in a non-decisive outcome (tie or refuse to arbitrate), then each insurer that contributed fees receives back what they paid in, minus a portion of the arbitration costs for their round that corresponds to the proportion of the total fees that that insurer contributed (for the two sides combined) for that round.

3.5 Insurers have a perfect knowledge of π_A

If Isaac knows a priori $\pi_A \neq 1/2$ then he knows who would eventually win the dispute with probability arbitrarily close to one if the case was appealed to a sufficiently large juror pool. (Or, more realistically, in the setting where there is a maxAppeal set that is very large and on which it is unviable for an attacker Eve to launch attacks to influence the last appeal round(s), Isaac knows who would win that round with high probability.) Hence, if $\pi_A > 1/2$, Isaac would know that Alice would eventually win the dispute as long as she paid whatever arbitration fees are required of her. Moreover, Isaac would know that funding each of these appeals is profitable as long as $s_A > 0$.

With similar reasoning, if Isaac is merely very confident that π_A is even slightly greater than 1/2, then he will estimate p_A as close to one.

3.5.1 Evaluating a case requires an effort

Suppose we are still in the setting where insurers can obtain a perfect knowledge of π_A, but now in order to do so they must expand an effort of e.

We consider an insurer who takes the strategy choosing random cases that Bob has appealed and then deciding whether to insurer Alice. (So for the moment, we do not consider any attempt to insure individuals who lost the previous round; hence this discussion is focused on resistance to bank attacks rather than on how the appeal system provides resistance to other kinds of attacks.) For the moment, to avoid the honest unity problem, we assume we have a single insurer.

Let N be the total number of cases under consideration in a given period. Suppose that uN of these cases are appealed by Bob that, while losing the previous round, knows $\pi_B > 1/2$ and hence that he will win eventually if he is willing to continue his appeals. On the other hand zN cases are appealed by hostile parties Eve who know that $\pi_B < 1/2$ and hope to win the case due to Alice's fees not be insured.

Let y be the percentage of cases that Isaac decides to evaluate to determine whether to insure them. Then

$$\text{Isaac payoff if decides to evaluate} = \frac{z}{u+z}(s_A - e) + \left(1 - \frac{z}{u+z}\right)(-e) = \frac{z}{u+z}s_A - e,$$

while

$$\text{Isaac payoff if decides not to evaluate} = 0.$$

Similarly, if V is the value at stake in the case that Eve is attempting to win unjustly via a bank attack, we compute

$$\text{Eve's payoff if she decides to appeal} = y(-s_A - x) + (1-y)(V),$$

and

$$\text{Eve's payoff if she decides not to appeal} = 0.$$

We seek to find the equilibrium values of y and z. To illustrate this balance, if insurers evaluate so often that attackers give up because they never win, the result will be that evaluating is unlikely to find cases worth insuring. In equilibrium, each of Isaac and Eve are indifferent between their two strategies, hence:

$$\frac{z}{u+z}s_A = e \Leftrightarrow z = \frac{\frac{e}{s_A}u}{1-\frac{e}{s_A}},$$

and

$$y(-s_A - x) + (1-y)V = 0 \Leftrightarrow y = \frac{V}{s_A + x + V}.$$

The ultimate measure of success of this system is the percent of cases that are misjudged, i.e. where Bob wins despite $\pi_B < 1/2$ (and particularly comparing this rate of success to the average costs in fees/capital lock-up, etc to use the system). In this simple model, (if we assume that all Bobs who have winning cases appeal, limiting our conclusions to the effectiveness of bank attacks on Alice), the number of cases misjudged is

$$z(1-y)N = \frac{\frac{e}{s_A}u}{1-\frac{e}{s_A}} \cdot \frac{V}{s_A + x + V} \cdot N.$$

Limitations and remarks:

- We examine insurer participation only for insuring the winner of the previous round - while relevant to the discussing the effectiveness of bank attacks, this is not satisfactory.
- It might seem that by increasing s_A, one can make the failure rate arbitrarily low. However, this will increase the appeal fees of Bob as well, which in practice will result in more misjudged cases, even if they are not included in this calculation due to the simplicity of the model.
- It is unrealistic that the insurer can obtain perfect knowledge of π_A.
- To avoid the honest unity problem, we assumed there was only one insurer, which is also unrealistic.

- In order to optimize the percentage of misjudged cases, what needs to be tuned is essentially u. Rather, in order to decrease u, one increases the number of jurors in the previous round so that there are fewer honest parties who would have winning cases upon appealing their results. Then by choosing the number of jurors in the previous round appropriately, one can arrive at a percentage of cases misjudged below some acceptable threshold.

3.6 Insurer's estimate of π_A updates after successive appeal rounds

After each round of voting insurers will have learned information about π_A, so in a more realistic model, Isaac's estimate of p_A may change over time. Of course, if there is an ongoing attack, jurors' votes may not be representative of the true value of π_A. However, note that in the absence of an attack that changes the jurors' voting incentives, in a given appeal round where there are N jurors, the number of votes that Alice receives is distributed as $\text{Binomial}(N, \pi_A)$.

A standard way to model a changing prior for a binomial probability π_A is with a beta distribution. Namely, one supposes that the insurer's prior distribution for π_A is given by $\text{Beta}(a, b)$. Then, if Isaac sees votes of (a_1, b_1) that are drawn from $\text{Binomial}(N, \pi_A)$, he updates his prior to $(a + a_1, b + b_1)$. Isaac's expected value for π_A given a prior of $\text{Beta}(a, b)$ is $\frac{a}{a+b}$ and his certainty in this believe scales with the size of $a + b$. Consequently, if $a + b$ is very large, Isaac's estimate of π_A will change little with new evidence.

In this section, we will consider an insurer's willingness to fund appeals based on his expectations that Alice will eventually win, and what this insurer behavior itself implies for Alice's eventual chances. Specifically, here we can analyze how this insurance proposal handles bank attacks - where an attacker Eve with a large budget continually appeals until the opposing party does not fund an appeal.

We assume that the insurers expect that there will be no future attacks that change the voting patterns of the jurors. In future work, we think a reasonable model for handling attacks that change voter incentives (such as bribes, $p + \epsilon$ attacks, pre-revelation attacks, etc) and incorporate them into this analysis will be for Isaac to not consider any votes seen while such an attack is active. In practice, insurers may be likely to increase their estimation of π_A, concluding that Alice likely has a winning argument, if they see that Alice is being attacked.

Note that, when deciding whether to insure a case, insurers must make predictions on whether future insurers will be willing to pay fees in future appeals (or indeed, Isaac will make predictions about his own willingness to pay fees in a future round when presented with updated voting data). Hence, there is some possible range of outcomes where Alice could win each round, but by small enough margins that insurers' priors of π_A degrade to the point where they are no longer willing to finance her appeals.

Faced with this phenomenon, one might expect this system to manifest properties of a gambler's ruin problem - namely that if there is some probability that

Alice obtains a result that is no longer insurable in any given round by chance, if the number of rounds is sufficiently high, eventually such a negative result will occur with high probability by chance. Anticipating future insurers lack of willingness to pay fees, present insurers would become less willing to pay fees in the current round, creating a vicious cycle that would drive down Alice's chances. This could be true even if insurers have a strong prior belief that Alice would win any given round.

Instead, we see that even in the extreme case where there is a bank attacker who is willing to appeal indefinitely, this vicious cycle is avoided, or at least limited. In Theorem 1, we provide a lower bound that insurers can compute on Alice's chances that takes into account both the possibility she may lose a round and the possibility that she obtains mediocre enough results to not be insured in future rounds. Theorem 1 currently assumes that there is some finite (but typically large) maximum number of rounds, avoiding the heart of classical gambler's ruin problems; however, we expect that the growth in the number of jurors per round is likely sufficiently fast to allow convergence to a non-zero lower bound on Alice's chances even if an infinite number of rounds was allowed if one makes some modest improvements to the argument, see Remark 8, which may be considered in future work.

In the following results, we drop the assumption that Alice won the previous dispute and the results apply equally to Bob.

Proposition 1. *Let $z \in [0,1]$. If $\pi_A > z$, then the probability that Alice receives a proportion of votes inferior to z in the ith appeal round is at most*

$$\frac{(1-z)\pi_A}{\left[2^{i+2}-1\right]\left(\pi_A - z\right)^2}.$$

Specifically, if $\pi_A > 1/2$, then the probability that Alice loses the ith appeal round is at most

$$\frac{\pi_A}{2(2^{i+2}-1)\left(\pi_A - \frac{1}{2}\right)^2}.$$

Proof. Note that there are $2^{i+2}-1$ total votes in the ith appeal round.

By [1][p.151], if X is distributed as Binomial(N,π), then if $r \le N\pi$ then

$$\text{Prob}(X \le r) \le \frac{(N-r)\pi}{(N\pi - r)^2}.$$

Denote by X the number of votes for Alice in the ith round. As we assume that $\pi_A > z$, in our case this inequality translates into

$$\begin{aligned}\text{Prob}\left(X \le z \cdot (2^{i+2}-1)\right) &\le \frac{\left(2^{i+2}-1-z(2^{i+2}-1)\right)\pi_A}{\left((2^{i+2}-1)\pi_A - z(2^{i+2}-1)\right)^2}\\ &= \frac{(1-z)\pi_A}{\left[2^{i+2}-1\right]\left(\pi_A - z\right)^2}.\end{aligned}$$

□

Proposition 2. *Let $\delta > 0$. Suppose we have a sequence $z_i \in [0,1]$ such that $|z_{i-1} - z_i| > \left(\sqrt{\frac{1+\delta}{2}}\right)^i$ for all $i = R+1, \ldots, L$. Then*

$$\prod_{i=R+1}^{L} \left(1 - \frac{(1-z_i)z_{i-1}}{\left[2^{i+2}-1\right](z_i - z_{i-1})^2}\right) \geq \left(1 - \frac{1}{2(1+\delta)^R}\right)^{\frac{1+\delta}{\delta}(1-(1+\delta)^{R-L})}.$$

This represents a lower bound on the probability that Alice obtains at least a proportion of z_i of the votes in the ith round for every round after the Rth, assuming her fees are always paid.

Proof. In general

$$\prod_{i=N}^{L} \left(1 - \frac{c}{(1+\delta)^i}\right) \geq \left(1 - \frac{c}{(1+\delta)^N}\right)^{\frac{1+\delta}{\delta}(1-(1+\delta)^{N-L})},$$

when $N \geq \frac{\log_2(c)}{\log_2(1+\delta)}$.

To see this note that

$$\log_2\left(\prod_{i=N}^{L}\left(1 - \frac{c}{(1+\delta)^i}\right)\right) = \sum_{i=N}^{L} \log_2\left(1 - \frac{c}{(1+\delta)^i}\right) \geq \sum_{i=N}^{L} (1+\delta)^{N-i} \log_2(1 - c(1+\delta)^{-N})$$

where we have used the fact that $(1+\delta)^i \geq (1+\delta)^N$ for all the i we consider so

$$(1+\delta)^{-N} \log_2(1 - c(1+\delta)^{-i}) \geq (1+\delta)^{-i} \log_2(1 - c(1+\delta)^{-N}).$$

Note

$$\sum_{i=N}^{L} (1+\delta)^{N-i} = \frac{1+\delta}{\delta}\left(1 - (1+\delta)^{N-L}\right).$$

Then

$$\log_2\left(\prod_{i=N}^{L}\left(1 - \frac{c}{(1+\delta)^i}\right)\right) \geq \frac{1+\delta}{\delta}\left(1 - (1+\delta)^{N-L}\right) \log_2(1 - c(1+\delta)^{-N})$$

establishing the claim.

Note, by our assumption on $|z_{i-1} - z_i|$, we have

$$2^i (z_{i-1} - z_i)^2 > (1+\delta)^i.$$

Furthermore, as $z_i \in [0,1]$ for all i, we have that $(1 - z_i)z_{i-1} \leq 1$. Then substituting $c = 1/2$, we get

$$\prod_{i=R+1}^{L} \left(1 - \frac{(1-z_i)z_{i-1}}{\left[2^{i+2}-1\right](z_i - z_{i-1})^2}\right) \geq \left(1 - \frac{1}{2(1+\delta)^R}\right)^{\frac{1+\delta}{\delta}(1-(1+\delta)^{R-L})}.$$

Note that $R \geq \frac{\log_2(1/2)}{\log_2(1+\delta)} = \frac{-1}{\log_2(1+\delta)}$. □

Suppose at the end of the Rth round, Isaac has a prior of $Beta(a_R, b_R)$. This prior may have been acquired through a combination of Isaac observing the votes in preceding voting rounds, as well as whatever analysis he makes of the case himself. Denote by a_i^* the number of votes that Alice receives in the ith round and b_i^* the number of votes that Bob receives. Then take

$$a_i = a_R + \sum_{j=R+1}^{i} a_j^* \text{ and } b_i = b_R + \sum_{j=R+1}^{i} b_j^*,$$

which gives the prior that Isaac will have after the ith round.

Let L denote the total, maximum number of rounds.

Proposition 3. *Let $\epsilon > 0$. Suppose $x \in \mathbb{R}$ is such that*

$$Prob\left(N(0, \pi_A(1-\pi_A)) \leq x\right) < \frac{\epsilon}{2},$$

where $N(0, \pi_A(1-\pi_A))$ is a normal variable with mean 0 and variance $\pi_A(1-\pi_A)$. Take

$$z_R = \frac{x}{\sqrt{a_R + b_R}} + \frac{a_R}{a_R + b_R}.$$

Then choose a sequence z_i such that

$$z_{i+1} \leq \frac{x}{\sqrt{a_i + b_i}} + \frac{a_i}{a_i + b_i} \tag{4}$$

for $i = R, \ldots, L-1$. Particularly, if

$$\frac{a_j^*}{a_j^* + b_j^*} \geq z_j$$

for $j = R, R+1, ..., i$, then it is sufficient that

$$z_{i+1} \leq \frac{x}{\sqrt{a_R + b_R + (2^{i+3} - i) - (2^{R+3} - R)}} + \frac{a_R + \sum_{j=R+1}^{i}(2^{j+2} - 1)z_j}{a_R + b_R + (2^{i+3} - i) - (2^{R+3} - R)}.$$

Then, there exists $R_0 \in \mathbb{N}$ such that if $i \geq R_0$,

$$Prob\left(X_{a_i, b_i} \leq z_{i+1} \middle| \frac{a_j^*}{a_j^* + b_j^*} \geq z_j, \forall j = R, R+1, ..., i\right) < \epsilon.$$

Note, that without knowing π_A, an insurer is of course incapable of choosing x and verifying if i (particularly $i = R$) is greater than R_0 as described above. We will see that this result nonetheless gives us insight in Heuristic 1.

Proof. Using a standard application of the delta method, for $\pi_A \in (0,1)$ with $\lim_{i\to\infty} \frac{a_i}{a_i + b_i} = \pi_A$, we have

$$\sqrt{a_i + b_i}\left(X_{a_i, b_i} - \frac{a_i}{a_i + b_i}\right) \to N(0, \pi_A(1-\pi_A))$$

in distribution as $i \to \infty$, where X_{a_i,b_i} is a Beta distribution with parameters a_i and b_i, and $N(0, \pi_A(1-\pi_A))$ is a normal distribution with mean 0 and variance $\pi_A(1-\pi_A)$. Hence, there exists R_0 such that if $i \geq R_0$,

$$\left| \text{Prob}\left(\sqrt{a_i + b_i} \left(X_{a_i,b_i} - \frac{a_i}{a_i + b_i} \right) \leq x \right) - \text{Prob}\left(N(0, \pi_A(1-\pi_A)) \leq x \right) \right| < \frac{\epsilon}{2}.$$

Take some $i \geq R+1$. Note that the number of votes from the $R+1$st round to the ith round inclusive is $\left(2^{i+3} - i\right) - \left(2^{R+3} - R\right)$. Hence, for any given i,

$$a_i + b_i = a_R + b_R + \left(2^{i+3} - i\right) - \left(2^{R+3} - R\right).$$

Suppose

$$\frac{a_j^*}{a_j^* + b_j^*} \geq z_j$$

for all $j = R, R{+}1, ..., i$. If z_j is chosen according to 4 for each $j = R, R{+}1, \ldots, i$, then

$$a_i \geq a_R + \sum_{j=R+1}^{i} (2^{j+2} - 1) z_j$$

and

$$b_i \leq b_R + 2^{i+3} - i - 2^{R+3} + R - \sum_{j=R+1}^{i} (2^{j+2} - 1) z_j.$$

We note by our assumptions on the choice of z_i that

$$\sqrt{a_i + b_i} \left(z_{i+1} - \frac{a_i}{a_i + b_i} \right) \leq x = \sqrt{a_R + b_R} \left(z_R - \frac{a_R}{a_R + b_R} \right)$$

for all $i \geq R$.

Then, if $i \geq R_0$,

$$\begin{aligned} \text{Prob}\left(X_{a_i,b_i} \leq z_{i+1} \right) &= \text{Prob}\left(\sqrt{a_i + b_i} \left(X_{a_i,b_i} - \frac{a_i}{a_i + b_i} \right) \leq \sqrt{a_i + b_i} \left(z_{i+1} - \frac{a_i}{a_i + b_i} \right) \right) \\ &\leq \text{Prob}\left(\sqrt{a_i + b_i} \left(X_{a_i,b_i} - \frac{a_i}{a_i + b_i} \right) \leq x \right) \\ &< \text{Prob}\left(N(0, \pi_A(1-\pi_A)) \leq x \right) + \frac{\epsilon}{2} \leq \epsilon \end{aligned}$$

□

Theorem 1. *Suppose that π_A is fixed from round to round and there exist insurers that*

- *are rational and attempt to maximize their expected value (without regards to risk aversion)*

- *collectively have sufficient resources to match any attacker*
- *after the Rth round, estimate the value of π_A as following a Beta(a_R, b_R) distribution and expect insurers in following rounds to estimate the value of π_A as following a Beta(a_i, b_i). distribution*

Further, suppose that there exists a decreasing sequence $z_i \in (1/2, 1]$ for $i = R+1, \ldots, L$ such that

$$\prod_{i=R+1}^{L}\left(1-\frac{(1-z_i)z_{i-1}}{\left[2^{i+2}-1\right]\left(z_i-z_{i-1}\right)^2}\right)\cdot\int_{z_{i-1}}^{1}\frac{y^{a_R-1+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}\cdot(1-y)^{b_R+2^{i+3}-i-2^{R+3}+R-2-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}}{B\left(a_R+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j, b_R+2^{i+3}-i-2^{R+3}+R-1-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j\right)}dy \geq t_A.$$

Then Isaac will insure Alice in the Rth round.

Proof. Denote by E_R the event that Alice receives at least a proportion of z_i of the votes in the ith round for each round from $R+1$ onwards.

We first note that

$$\text{Prob}(E_R) = \text{Prob}\left(\frac{a_i^*}{a_i^*+b_i^*} \geq z_i \colon \forall i > R\right)$$

$$\geq \prod_{i=R+1}^{L}\text{Prob}\left(\frac{a_i^*}{a_i^*+b_i^*} \geq z_i \middle| \frac{a_j^*}{a_j^*+b_j^*} \geq z_j : j = R+1, ..., i-1\right)$$

$$\geq \prod_{i=R+1}^{L}\text{Prob}\left(\frac{a_i^*}{a_i^*+b_i^*} \geq z_i \middle| \pi_A \geq z_{i-1}\right)\cdot\text{Prob}\left(\pi_A \geq z_{i-1} \middle| \frac{a_j^*}{a_j^*+b_j^*} \geq z_j : j = R+1, ..., i-1\right)$$

$$\geq \prod_{i=R+1}^{L}\text{Prob}\left(\frac{a_i^*}{a_i^*+b_i^*} \geq z_i \middle| \pi_A \geq z_{i-1}\right)\cdot\int_{z_{i-1}}^{1}\frac{y^{a_R-1+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}\cdot(1-y)^{b_R+2^{i+3}-i-2^{R+3}+R-2-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}}{B\left(a_R+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j, b_R+2^{i+3}-i-2^{R+3}+R-1-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j\right)}dy$$

$$\geq \prod_{i=R+1}^{L}\left(1-\frac{(1-z_i)z_{i-1}}{\left[2^{i+2}-1\right]\left(z_i-z_{i-1}\right)^2}\right)\cdot\int_{z_{i-1}}^{1}\frac{y^{a_R-1+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}\cdot(1-y)^{b_R+2^{i+3}-i-2^{R+3}+R-2-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}}{B\left(a_R+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j, b_R+2^{i+3}-i-2^{R+3}+R-1-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j\right)}dy$$

Here the first inequality comes from taking the most pessimistic result in each round that is consistent with E_R. The second inequality comes from a standard calculation in Bayesian inference (see, for example 12.2.2 of [3]) and the fact that Prob $\left(\frac{a_i^*}{a_i^*+b_i^*} \geq z_i \middle| \pi_A\right)$ is monotonic in π_A. The third inequality comes the expression of probability constraints on π_A in terms of the beta distribution and the fourth inequality comes from Proposition 1, using the fact that z_i is decreasing.

Denote

$$I(R, \{z_i\}) = \prod_{i=R+1}^{L}\text{Prob}\left(\frac{a_i^*}{a_i^*+b_i^*} \geq z_i \middle| \pi_A \geq z_i\right)\cdot\text{Prob}\left(\pi_A \geq z_i \middle| \frac{a_j^*}{a_j^*+b_j^*} \geq z_j : j = R+1, ..., i-1\right).$$

If a rational insurer computes

$$\text{Prob}(E_R) \geq I(R, \{z_i\}) \geq t_A$$

immediately prior to the last possible round, namely when $R = L - 1$, then using the fact that $z_L > 1/2$, she will conclude $p_A \geq t_A$, so she will pay Alice's fees.

We proceed by induction. Suppose that insurers are willing to pay Alice's fees for any round $R \geq R_0$ for which they compute

$$\mathrm{Prob}(E_R) \geq I(R, \{z_i\}) \geq t_A.$$

Then in the $R_0 - 1$st round, an insurer will note that

$$I(R, \{z_i\}) \geq I(R-1, \{z_i\})$$

for all R. So, if he computes $I(R_0 - 1, \{z_i\}) \geq t_A$ based on his current priors, in any event consistent with E_{R_0-1}, namely where Alice does, in fact, receive at least a proportion of z_R in the Rth voting round for each $R \geq R_0$, insurers are willing to pay Alice's fees in all remaining rounds by the induction hypothesis. Then, as $z_i \geq 1/2$ for all $i \geq R_0$, in the event E_{R_0}, Alice wins all remaining rounds and her fees are always paid; hence she ultimately wins. Thus, if Isaac computes $\mathrm{Prob}(E_{R_0-1}) \geq I(R_0 - 1, \{z_i\}) \geq t_A$, he is willing to pay Alice's fees in the $R_0 - 1$st round, completing the induction.

Thus, in any round where Isaac computes $\mathrm{Prob}(E_R) \geq I(R_0, \{z_i\}) \geq t_A$, he concludes that $t_A \leq \mathrm{Prob}(E_R) \leq p_A$ and he should be willing to pay Alice's fees.

□

Heuristic 1. *Suppose one tries to choose $z_R \in (1/2, 1)$ such that*

$$\int_0^{z_R} \frac{y^{a_R-1}(1-y)^{b_R-1}}{B(a_R, b_R)} \approx \epsilon$$

is small. Then, attempt to choose a decreasing sequence $z_i \in (1/2, 1)$ for $i = R+1, \ldots, L$ such that

$$z_{i+1} \leq \frac{x}{\sqrt{a_R + b_R + (2^{i+3} - i) - (2^{R+3} - R)}} + \frac{a_R + \sum_{j=R+1}^{i}(2^{j+2} - 1)z_j}{a_R + b_R + (2^{i+3} - i) - (2^{R+3} - R)}.$$

Then if

$$\int_0^{z_{i-1}} \frac{y^{a_R-1+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j} \cdot (1-y)^{b_R+2^{i+3}-i-2^{R+3}+R-2-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}}{B\left(a_R + \sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j, b_R + 2^{i+3} - i - 2^{R+3} + R - 1 - \sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j\right)} dy \tag{5}$$

remains small for all i, i.e. ϵ or smaller, we will have that Isaac will insure Alice in the Rth round if

$$(1-\epsilon)^{L-R} \prod_{i=R+1}^{L} \left(1 - \frac{(1-z_i)z_{i-1}}{\left[2^{i+2} - 1\right](z_i - z_{i-1})^2}\right) \geq t_A.$$

In particular, if additionally one has

$$|z_{i-1} - z_i| > \left(\sqrt{\frac{1+\delta}{2}}\right)^i,$$

then it is sufficient that

$$\left(1 - \frac{1}{2(1+\delta)^R}\right)^{\frac{1+\delta}{\delta}(1-(1+\delta)^{R-L})} \cdot (1-\epsilon)^{L-R} \geq t_A.$$

using the bound of Proposition 2.

This process can fail for three reasons

1. *The integral computed in 5 does not remain less than ϵ.*
2. *The z_i that are computed fall out of $(1/2, 1)$.*
3. *The ultimate lower bound computed*

$$\left(1 - \frac{1}{2(1+\delta)^R}\right)^{\frac{1+\delta}{\delta}(1-(1+\delta)^{R-L})} \cdot (1-\epsilon)^{L-R} < t_A.$$

By Proposition 3, a failure of reason 1 means that $R < R_0$. In this case, it may be possible to choose the first few terms of z_i ad-hoc attempting to make the integral of 5 small for each, then using the recursion for larger i. Also, note that R_0 depends on z_R, so another choice of z_R could conceivably work, even with the same R. Failures of reasons 2 and 3 indicate that the prior a_R, b_R is not strong enough (relative to R and z_R) to use this lower bound (which is not necessarily sharp). A failure of reason 3 may also indicate that the z_i are too close to one another, in which case

$$\prod_{i=R+1}^{L} \left(1 - \frac{(1-z_i)z_{i-1}}{\left[2^{i+2}-1\right](z_i - z_{i-1})^2}\right)$$

blows up. To deal with this, δ can be adjusted.

In general, finding a choice of z_i that optimize the lower bound on Prob(E_R), and hence creating a more succinct characterization of when this lower bound is greater than t_A, seems non-obvious. In practice, one can attempt the choices of this heuristic, possibly supported by numerical methods, to find a given lower bound, without a guarantee that it will provide a global optimum. Nonetheless, any sequence z_i that one finds that satisfy the conditions of Theorem 1 is provably sufficient.

Remark 8. *Note the ways in which we have used the assumption that there are only a finite number, L, of future rounds. In order to prove a theorem in this mold where insurers are able to estimate the probability that future insurers would continue to pay Alice's fees indefinitely, one would need to do the following:*

- *Replace the constraint on the sequence z_i:*

$$z_{i+1} \leq \frac{x}{\sqrt{a_R + b_R + (2^{i+3} - i) - (2^{R+3} - R)}} + \frac{a_R + \sum_{j=R+1}^{i} (2^{j+2} - 1) z_j}{a_R + b_R + (2^{i+3} - i) - (2^{R+3} - R)}.$$

  with a more strict constraint that would be sufficient for

$$\sum_{i=R+1}^{\infty} Prob \left(X_{a_i, b_i} \leq z_{i+1} \middle| \frac{a_j^*}{a_j^* + b_j^*} \geq z_j,\, \forall j = R, R+1, ..., i \right)$$

  to converge to a sufficiently small value (rather than for each term to be bounded by the same small constant)

- *choose z_i that then also satisfies $|z_i - z_{i+1}| > \left(\sqrt{\frac{1+\delta}{2}}\right)^i$ for some $\delta \in (0, 1)$, to show that*

$$\prod_{i=R+1}^{\infty} \left(1 - \frac{(1 - z_i) z_{i-1}}{[2^{i+2} - 1] (z_i - z_{i-1})^2} \right)$$

  is lower bounded appropriately

- *find a sequence z_i subject to these constraints such that $z_i \in (1/2, 1)$ for all $i > R$.*

- *Adapt the induction proof that insurers are willing to pay Alice's fees for any round $R \geq R_0$ for which they compute $Prob(E_R) \geq I(R, \{z_i\}) \geq t_A$ as there is no longer a last round from which to work backwards - indeed, in the case of an infinite number of rounds it is likely that (even if Alice's chances of winning future roundss by wide margins as estimated by $Prob(E_R)$ is high), there can be distinct equilibria where rational insurers are willing to pay her fees or not depending on their expectations of how insurers will behave in future rounds.*

Note, however, that there is a fixed total number of PNK (and a given PNK can only be subdivided up to a fixed number of decimal places), so there can only be a finite collection of possible jurors at any given time. Hence, in practice, it would be of limited utility to allow appeals indefinitely in any event. As PNK is chosen with replacement, larger and larger appeals from the same (finite) pool of PNK would result in less variability in how the drawn sample reflects the broader population; however, beyond a certain point a given appeal reasonably reflects a census of all PNK holders.

Example 3. *We consider a case where $R = 5$ and Isaac has a prior of $a_R = 98$ and $b_R = 2$. Suppose $L = 25$, which for realistic choices of juror fees would put the required appeal fees in the last round on the order of the market capitalization of Ethereum. Then if one takes z_i as below, Isaac will find a lower bound on his winning chances by estimating the future rounds as follows:*

i	$a_i \geq$	$b_i \leq$	z_i	$Prob\left(\pi_A \geq z_i \middle\vert \frac{a_j^*}{a_j^*+b_j^*} \geq z_j : j = R+1,...,i-1\right)$
5	98	2	.9	$3.54 * 10^{-4}$
6	212	15	.839	$5.57 * 10^{-6}$
7	426	56	.7855	$7.10 * 10^{-9}$
8	827	166	.7385	$5.91 * 10^{-13}$
9	1583	483	.7015	$2.20 * 10^{-11}$
10	3019	1044	.67	$2.15 * 10^{-24}$
11	5762	2396	.64	$4.12 * 10^{-37}$
12	11004	5344	.615	$5.41 * 10^{-54}$
13	21080	11652	.59	$1.58 * 10^{-89}$
14	40413	28056	.574775648	$1.22 * 10^{-16}$
15	151937	110168	.555111572	$3.30 * 10^{-142}$
16	297456	226792	.548901697	$3.17 * 10^{-160}$
17	585238	463297	.54429633	$4.98 * 10^{-179}$
18	1155973	941137	.540880899	$6.10 * 10^{-199}$
19	2290282	1903979	.531785952	$2.75 * 10^{-748}$
20	4520754	3867810	.5245595765	$1.30 * 10^{-1510}$
21	8921381	7855790	.518911423	$4.50 * 10^{-2413}$
22	17627270	15927116	.514417556	$3.22 * 10^{-3481}$
23	34888258	32220559	.510864843	$3.83 * 10^{-4740}$
24	69171817	65045863	.508056176	$6.28 * 10^{-6243}$
25	137361962	131073445	.50583573	$4.47 * 10^{-8060}$

Then

$$\prod_{i=R+1}^{L}\left(1-\frac{(1-z_i)z_{i-1}}{\left[2^{i+2}-1\right]\left(z_i-z_{i-1}\right)^2}\right)\cdot\int_{z_{i-1}}^{1}\frac{y^{a_R-1+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}\cdot(1-y)^{b_R+2^{i+3}-i-2^{R+3}+R-2-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j}}{B\left(a_R+\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j,\, b_R+2^{i+3}-i-2^{R+3}+R-1-\sum_{j=R+1}^{i-1}(2^{j+2}-1)z_j\right)}dy$$

$$\geq .504012138\cdot .99964599288 = 0.50383371411 > .5 = t_A.$$

Note that, at least in this example, the more rigid constraint on the z_i seems to be to get

$$\prod_{i=R+1}^{L}\left(1-\frac{(1-z_i)z_{i-1}}{\left[2^{i+2}-1\right]\left(z_i-z_{i-1}\right)^2}\right)$$

to converge with z_i remaining in $(1/2,1)$ rather than to get

$$Prob\left(\pi_A \geq z_i \middle\vert \frac{a_j^*}{a_j^*+b_j^*} \geq z_j : j = R+1,...,i-1\right)$$

to stabilize/decrease. Indeed, the z_i were adapted up to $i = 14$ towards this end.

One might hope to make small improvements to these bounds. Moreover, it is worth noting that this scenario, in which Isaac evaluates his chances of winning against a determined bank attacker who is willing to appeal indefinitely, is rather pessimistic. In practice, Isaac would probably factor in some expectation of how many times a given case is likely to be appealed, which should make the prospect of providing fee insurance more attractive that what is indicated by Theorem 1. Nonetheless, we see that with a fairly strong prior, even under these pessimistic assumptions, one could expect insurers to participate.

References

[1] William Feller. *An Introduction to Probability Theory and its Applications, Volume I, Third Edition.* 1950.

[2] William George. Why Kleros needs a native token. Online, `https://medium.com/kleros/why-kleros-needs-a-native-token-5c6c6e39cdfe`.

[3] Han Liu and Larry Wasserman. *Statistical Machine Learning.* 2014. `http://www.stat.cmu.edu/~larry/=sml/Bayes.pdf`.

[4] Alex Tabarrok. When can token curated registris actually work? Online, `https://medium.com/wireline/when-can-token-curated-registries-actually-work-%C2%B9-2ad908653aaf`.