Draft: Using cryptoeconomic incentives to manage the revelation of vote information in Kleros

September 7, 2018

Abstract

Careful management of the flow of information is essential to systems that are designed to achieve desired outcomes through the use of Schelling points. In this work, we discuss incentives that can be put in place to encourage the appropriate diffusion of the knowledge of jurors' votes in Kleros. In particular, we analyze an anti-pre-revelation game proposed by Vitalik Butern, that penalizes jurors who pre-reveal their votes, in the context of Kleros, a Schelling point based dispute resolution platform. We point out an attack on this proposed anti-pre-revelation game and discuss parameter choices that can be made so that this attack is not profitable and has acceptable griefing factors. Moreover, we analyze the equilibria that arise in such a system, and discuss choices of parameters that would result in desirable equilibria.

Schelling points [8] have provided a framework for understanding how individuals make decisions when they have an incentive to be coherent with other members of a group with whom they cannot communicate. Proposed by Truthcoin [9], the basic idea is that, if voters are asked to choose between various outcomes for which there is some "true" response, lacking other information on how the other jurors will vote, each juror will reason that the other jurors will gravitate towards this "true" choice. Such a system can be used as an oracle, where the "true" outcome is relatively clear, such as in the system of Augur [7]. The core idea of Kleros [5] is that this same idea can be used where voters are asked to decide between outcomes in an arbitrary dispute in an arbitration system. Even though real-world disputes may not always have an absolutely clear outcome, a voter who attempts to vote what they believe a generic reasonable person would vote will be right on average.

However, the absence of communication in such systems is essential. If voters can share their votes among each other, they can attempt to influence the Schelling point towards hostile answers. In this work, we discuss managing these issues through structured cryptoeconomic incentives, specifically in the context of Kleros.

1 Introduction

1.1 Setup

We consider a system based on Kleros. Kleros is a smart contract based dispute resolution mechanism that relies on randomly selecting some fixed number of tokenholders of the Kleros token pinakion (PNK) to serve as jurors for a given dispute. (Already that the number of voters on each decision is fixed may introduce slight differences with other Schelling point based systems, such as Augur [7].) Then a system of rewards and punishments compensates these jurors for their work in examining the dispute and incentivizes them to vote coherently. Each juror has placed a deposit of d, in PNK, and if they vote incoherently, they lose this deposit. Then the coherent jurors divide these lost deposits proportionally, namely a coherent juror receives:

$$\frac{\#\text{ incoherent jurors} \cdot d}{\#\text{ coherent jurors}}.$$

Additionally, jurors are paid fees in ether (ETH) that are financed by the parties to the dispute, however the amount of these fees paid to the jurors does not depend on their vote. See [5] for further details. We will use the terms jurors and voters interchangeably to describe the participants of the Schelling point based consensus game, but the results of this work apply to any other Schelling point system that uses the same structures.

Voting is managed via a commit-and-reveal scheme [2]. While a dispute is being resolved, there is a commitment phase, when jurors commit to a blinded version of their vote, followed by a reveal phase where this vote is then made public. During the commitment phase, a juror selects a random r and computes $H(r, \text{vote}) = c$, where H is some hash function. Then the juror publishes c to the blockchain. During the reveal phase, the juror publishes the random value r and her vote, and the Kleros arbitrator contract verifies that these value produce a c that match the juror's commitment.

As we analyze this system, we make the following assumptions about the behavior of its participants:

- Actors are rational and self-interested.
- Actors which are not assumed to be an attacker ("honest actors") have no external interest in the outcome of the dispute beyond the payouts of the system.
- Attackers may have an external (financial) interest that motivates their attack, but we do not explicitly consider it. Particularly, we calculating griefing factors such interests are excluded.

1.2 Examples of pre-revelation attacks

The most basic kind of pre-revelation would be for a voter to publish on a public forum that she is going to vote a certain way. Such a declaration may not be

credible, particularly because reward for being coherent increases as the number of coherent voters decreases. Namely, a voter maximizes her reward by being on the winning side of a narrow decision.

In order to provide more convincing evidence of her vote, a juror could publish her secret r during the commitment period, allowing anyone to verify her vote against her commitment. Then one might put a system in place to penalize voters who release their r by allowing other jurors to use it to steal part of the attacker's deposit. This defense is described in the Truthcoin whitepaper [9]. However, pre-revelation attacks take myriad forms. As a result, defenses which are heavily tailored to the type of pre-revelation attack are unlikely to be effective.

For example, a juror who wants to pre-reveal her vote might do any of the following:

- Smart contract enforced, self-inflicted penalty - in this attack, a voter who wants to publicly signal her vote creates a smart contract in which she deposits some amount of value. Then the contract only allows her to get back this value if she votes in a certain way on the dispute. As such, observers will be able to see that in order to lie about her vote, the attacker must absorb some visible financial loss. See [3] for an implementation of this attack.

- Cut-and-choose - an attacker submits a list of commitments to the same vote using different random values:

  $$c_1 = H(r_1, \text{vote}), c_2 = H(r_2, \text{vote}), \ldots, c_n = H(r_n, \text{vote})$$

  and a deposit d^* to a smart contract she creates. The smart contract obtains a random value $i \in \{1, 2, \ldots, n\}$ from a random beacon which the attacker cannot predict prior to submitting the list of commitments. Then, the attacker reveals her vote and r_j to the smart contract for each $j \neq i$. If she fails, she lose her deposit. Then the attacker proceeds to submit $c_i = H(r_i, \text{vote})$ to the dispute. This attack is inspired by an attack used in the context of secure two-party computation [6]. An outside observer knows that if the attacker does not actually know r_i such that $H(r_i, \text{vote}) = c_i$, she took a $\frac{n-1}{n}$ percent chance risk of losing her deposit d^*. Namely, the chance of successfully lying with this attack is $\frac{1}{n}$, so the expected cost of a successful lie is nd^*. Hence for a fixed deposit size, the cut-and-choose attack is more convincing than the smart contract enforced self-inflicted penalty attack, even as it requires a shorter capital lockup period as the attacker can recover her deposit immediately after revealing the r_j. See [4] for an implementation of this attack.

- zk-SNARK proof - an attacker can issue a zk-SNARK proof that she possesses some r such that $H(r, \text{vote}) = c$, where c is the commitment to the attacker's vote.

Note that in none of these attacks does an observer learn r; nevertheless, they are provided with highly convincing evidence of the attacker's vote. Ultimately, a solution is required that can penalize an attacker for revealing any probabilistic information about her vote, no matter what form that information is transmitted in.

1.3 Summary of the proposed anti-pre-revelation game of [1]

In [1], Vitalik Buterin proposes a secondary game that can be added on to Schelling point schemes that can effectively penalize voters who pre-reveal, no matter what form this pre-revelation takes (releasing the pre-image of a committed vote, a zero-knowledge proof of that a committed vote corresponds to a given value, a smart contract enforced self-inflicted penalty that executes depending on the vote).

The proposal is the following: Suppose that voters are asked to make a binary choose between options X and Y. Suppose there are M total votes. Each voter will have made a deposit of N to be used in this game (beyond whatever deposits are required by the primary consensus game).

After the votes have been committed to but before they are revealed, Alice can place a bet that a given juror Bob voted X. Then after the commitments have been revealed, one calculates $P = \frac{\#\text{votes for} X}{M}$. If Bob did indeed vote for X, then Alice receives $N(1-P)$ from Bob. If Bob did not vote for X, then Alice pays NP to Bob.

This scheme has several nice properties:

- Bob is only penalized for being predictable in his vote *compared to the average voter*. For instance, if Alice can predict that Bob will vote X because X is the obvious answer and *everyone* is going to vote X, then $P = 1$ and the transfer from Bob to Alice is $N(1-1) = 0$.

- If Alice has no information about Bob's vote beyond being able to evaluate how the average voter would vote on the case, namely being able to guess P, then from Alice's perspective there is a P chance Bob will vote X and a $1-P$ chance that he will not. So the expected return on a bet that Bob votes X is

$$E(\text{transfer from Bob to Alice}) = P \cdot N(1-P) - (1-P) \cdot NP = 0.$$

  Namely, if Alice votes randomly the transfers to and from Bob will average out. So, Bob will not suffer a loss on average.

2 Our contributions

We provide an analysis of the anti-pre-revelation game proposed in [1]. In particular, we point out a general attack, that was unnoticed in [1] and discuss how

to adjust fees and parameters so that this attack is not viable. Subsequently, we analyze properties and examine the equilbria of the game created by combining a primary consensus game and a secondary anti-pre-revelation game, particularly for a primary consensus game which is structured as that of Kleros.

3 Betting strategies that take advantage of knowing one's own votes.

In the comments of [1], Paul Sztorc makes the following interesting remark regarding the proposed anti-pre-revelation game:

"However, this limiting of the gamble is tough, because I can always pretend to be someone else (someone ratting me out), and rat myself out. Is that not a fundamental limitation? Also, I'll have to think more about my controlling 30% votes, lying with 4% (and not telling anyone) and then betting "against" the 96%."

Vitalik responds to the first point, that a voter can bet on his own vote in such a way to drain the deposit preventing *other* people from betting against him, with:

"So, bet against yourself as a way of draining your security deposit to prevent others from betting against you? Okay, I accept that; that is a good reason why bets need to have non-zero "spread" that is absorbed by the system to make them costly." Such a spread can be created by imposing a fee on the bettor so that the game is no longer zero-sum. Then even if the pre-revealor manages to drain his entire deposit, she still bears some cost.

Note that we have not yet specified how to treat situations where several people bet on the outcome of a single vote. At least for the purposes of this work we choose the following framework:

Design Choice 1. *Suppose that multiple parties each place the required deposit to bet on a juror's vote. Suppose* $P = \frac{\#vote\ for\ X}{M}$*, where* M *is the total number of jurors. Then if a bettor correctly predicts that the juror votes* X*, they receive*

$$\frac{bettor's\ deposit}{total\ deposits\ predicting\ juror\ votes\ X} \cdot N(1-P),$$

and if the bettor is incorrect their contribution to the payment to the juror is

$$\frac{bettor's\ deposit}{total\ deposits\ predicting\ juror\ votes\ X} \cdot NP.$$

Furthermore, any fees that bettors pay (so that betting has "a non-zero spread") are similarly divided proportionally.

Note that some bettors might bet that Alice will vote X while others will predict that she votes Y. These bets can be handled separately, each with

payoffs according to Design Choice 1, and as Alice will only lose one of them, a single deposit of N will suffice for both.

Alternatively, we could have taken the first bet that arrived on a juror's vote, and disallowed further bets beyond that. This model would be somewhat more difficult to analyze as we would have to discuss the chances of a given bet arriving first. Moreover, under the assumptions of Design Choice 1, if an attacker bets on herself to drain her security deposit, the amount that she is able to recover will be diluted. As we will see later, if we also assume Design Choice 2, the amount that an attacker will be able to drain from her security deposit will be capped in terms of the number of votes she controls.

However, there is also the additional problem, suggested in the second half of Sztorc's comment, that a voter knowing their own vote tells them something about P, and hence tells them something about how *everyone else* will vote relative to P.

Suppose that an attacker controls A of the M votes. For the moment, assume that all of the attacker's vote in the same direction. Then $P = \frac{\#\text{votes for} X}{M}$ is calculated including the attacker's votes. However, the attacker then knows that if she votes X, less than a proportion of P of the other $M - A$ votes voted for X. Then, absent any fees/based on the game as described so far, it makes sense for the attacker to bet that each of the $M - A$ other voters voted for Y. Similarly, if the attacker voted Y, then the other $M - A$ votes voted for X at a proportion higher than P, so it makes sense for the attacker to bet that each of the other $M - A$ votes voted for X.

Concretely,

Attack 1. *Eve uses her A votes to vote for Y, then there were MP votes for X and $M(1 - P) - A$ votes for Y excluding those of Eve. Eve bets that all of the $M - A$ votes she does not control voted for X.*

Then, if no one else places any bets, Attack 1 has a return of:

$$MP \cdot N(1 - P) - [M(1 - P) - A] \cdot NP = NPA.$$

Example 1. *Suppose that $M = 5$ and $A = 1$. The four votes that Eve does not control vote X, X, X, Y. (Of course, Eve does not know that in advance.)*

If Eve votes Y, bets that the other four voters will vote X, and no other actors place any bets, then $P = 2/3$, so Eve wins $N/3$ from each of the three voters who vote X and loses $N \cdot 2/3$ to the voter who votes Y for a net gain of $N/3 = NPA$.

If Eve votes X and bets that the other four voters will vote Y, then $P = 1/5$, so Eve wins $N \cdot 4/5$ from each of the voter who votes Y and loses $N \cdot 1/5$ to each of the three voters who vote X for a net gain of $N/5 = NPA$.

Note that Eve's gain is being drawn in two very different ways depending on whether she votes X or Y, namely whether she votes coherently or incoherently. In this example, if Eve votes incoherently by voting Y, the NPA she obtains is being obtained by taking a small amount from each of the voters who voted

X. Namely, Eve is taking a small amount from the voters who *were right*. Of course, Eve will lose her deposit d for the primary coherence game by being incoherent, so if d is sufficiently large relative to N, Eve will still take a net loss. If Eve votes coherently by voting X, the NPA is obtained by taking a small amount from each of the voters who voted incoherently by voting Y. Eve in fact makes a payment to the other voters who voted X. So, by pursuing this strategy, Eve is essentially forcing everyone to double down on the coherence game. Eve would not lose her deposit d in this case.

Neither of these results (being able to grief coherent voters in exchange for a lost deposit or being able to up the ante on the coherence game) are not necessarily catastrophic in themselves. However, note, in any event and regardless of how Eve votes, betting that all other voters will vote the opposite of her vote is a dominant strategy (unless the other bettors place bets in such a way that dilutes the attacker's rewards from correct bets while not diluting her rewards from incorrect bets; namely unless the other bettors place bets that correspond to having more information about the various juror's bets than what is available to the attacker). Thus, voters will be encouraged to make bets constantly when they do not actually have any individual knowledge of the other voter's decisions.

The initial idea to correct this problem is to set a fee for betting that makes the above strategy non-profitable. We imagine that a bettor must pay F, which can be some function of N, M, P, etc (as these are known values when it comes time to make the payouts for this game).

Remark 1. *For the moment we assume that these fees are simply burnt rather than being redistributed to the actors in this vote, where the attacker might get back part of the fees she paid mitigating the cost of the attack. In the case of Kleros, the burnt fees, paid in PNK, could then be used to fund the minting of new PNK while maintaining the same total number of tokens in circulation. By distributing this new PNK to the coherent jurors over a large number of cases/over the cases over a large period of time, arbitration work could essentially be subsidized, but an individual attacker's lost fees would only have a negligible influence on the future subsidies she would receive.*

Remark 2. *Note that when analyzing the fees necessary for Attack 1 to not be profitable, etc, we can assume without loss of generality that we have at most one attacker in "each direction;" namely one attacker who has voted Y and is betting that the other voters are voting X, and one attacker who has voted X and is betting that the other voters are voting Y. This is because if there were several attackers making bets in the same direction, due to Design Choice 1, they each pay a proportion of the fees and receive a proportion of the payouts. In fact, multiple attackers will be collectively less efficient in griefing the honest jurors as the attackers are placing bets on each others votes, whereas a single attacker would avoid the extra fees that this implies. If there are two attacks in opposite directions, as these bets are handled separately for the purposes of payouts, they do not affect each other in our analysis.*

Then ideally we would take,

$$F > \frac{NPA}{M-A}.$$

This would result in the fee paid per bet exceeding the average return from the $M-A$ bets the attacker must make to execute the above strategy.

However, this runs into the following difficulty. Imagine that Eve uses her A votes to vote for Y and all of the other $M-A>0$ votes are for X. Then

$$P = \frac{\#\text{ votes for } X}{M} = \frac{M-A}{M} \Rightarrow 1-P = 1-\frac{M-A}{M} = \frac{A}{M}.$$

Then

$$\frac{NPA}{M-A} = \frac{N\frac{M-A}{M}A}{M-A} = N\cdot\frac{A}{M} = N(1-P). \tag{1}$$

Note that from the perspective of an observer, $M-A$ "honest" voters voting for X and A attackers voting for Y is indistinguishable from any result where X receives $M-A$ votes and Y receives A votes. (Conceivably, it may be possible to differentiate these situations based on the pattern of which votes are the targets of anti-pre-revelation bets and adjust the fees for that, however so far that seems complicated.) Thus, if we admit the possibility of attackers with an arbitrary number A of votes, we must require that $F \geq N(1-P)$. Namely, the fee needs to at least equal the reward for winning the game and the anti-pre-revelation game becomes non-viable.

It might be natural at this point to assume that $A \leq M/2$, as if Eve is capable of executing a 51% attack one cannot expect security. However, this constraint does not help very much. Let $R \in \mathbb{N}$. Then:

$$A \leq \frac{M}{R} \Rightarrow M-A \geq \frac{R-1}{R}M \Rightarrow \frac{1}{M-A} \leq \frac{1}{\frac{R-1}{R}M}.$$

So

$$\frac{NPA}{M-A} \leq \frac{NP\frac{M}{R}}{\frac{R-1}{R}M} = \frac{NP}{R-1}.$$

In particular, if $R=2$, ie $A \leq M/2$,

$$\frac{NPA}{M-A} \leq NP.$$

In the extreme case where Eve makes $A = M/2$ votes for Y and all of the other votes are for X, all of the inequalities are equalities and

$$NP = N/2 = N(1-P).$$

So to recap, assuming that $A \leq M/2$ allows one to take the fees

$$F \geq \min\{N(1-P), NP\}$$

which is strictly less than $N(1-P)$ and results in the game being viable for $P < 1/2$, but fees increase to become quite close to $N(1-P)$ as P approaches $1/2$. In the next section we will consider improvements on this model.

4 Asymmetric fees and the case where $P \leq 1/2$

Note that it is not necessary to apply the same fee on a bettor who makes a correct prediction of a voter's vote versus a bettor who makes an incorrect prediction. In fact, when structuring the payments it is important that

- a correct better receives $N(1-P)$ from the target and an incorrect bettor pays NP to the target so that the game will average out from the perspective of an honest target who has not revealed his vote.
- that a correct bettor does not pay so much in fees that betting resulting in them losing money, namely that the fee for a correct bet be less than $N(1-P)$.

At first it seems that the fee on an incorrect answer would also be bounded by $N(1-P)$, as this is the amount left over for an incorrect bettor after paying NP to the target. Whereas taking a fee of $N(1-P)$ from a correct bettor removes any incentive to play the anti-prerelevation game, the entire $N(1-P)$ from an incorrect bettor could be reasonable applied to fees. In fact, we could reasonably have the bettor submit a deposit of $K \geq N$ and if they are incorrect pay NP to the target and then pay $K - NP$ in fees.

Note that increasing the bettor's deposit like this cannot entirely remove the attack described above as we saw in Equation 1 that in an extreme case the attacker can make an average return from her victims of $N(1-P)$, and if $P > 1/2$ in that extreme case, the attacker will never place an incorrect bet. So we also impose a fee of F on correct bets. However, with assumptions about the resources of the attacker $A < M/R$ we will provide bounds on attacker gains and griefs that are tunable in terms of F, R, and K.

Now the amount gained by the above attack is $N(1-P) - F$ from each of the MP many targets where the bettor is right, with a loss of K for each of the $M(1-P) - A$ targets where the bettor is incorrect (note that the structure of the attack assumes that the attacker uses all of A votes to vote the same way and bets in the opposite way so that $A \leq M(1-P)$). Thus the total gain is

$$MP \cdot [N(1-P) - F] - [M(1-P) - A] \cdot K = M(1-P)[NP - K] - MPF + AK.$$

For this gain to be non-positive,

$$F \geq \frac{M(1-P)(NP-K) + AK}{MP}$$

(note we can exclude the case $P = 0$ as in this case the attacker places no successful bets).

Note that one cannot tell from the results of a vote the value of A, so F cannot depend on this value. However, we assume that $A \leq M/R$, so

$$\frac{M(1-P)(NP-K) + AK}{MP} \leq \frac{M(1-P)(NP-K) + \frac{M}{R}K}{MP}$$

$$= \frac{(1-P)(NP-K)+\frac{K}{R}}{P}.$$

So if we take

$$F \geq \frac{(1-P)(NP-K)+\frac{K}{R}}{P}, \tag{2}$$

this will suffice.

Notice that $F \leq N(1-P)$ (so that a correct bettor does not suffer a loss) when

$$(1-P)(NP-K)+\frac{K}{R} \leq N(1-P)P \Leftrightarrow \frac{1}{R} \leq 1-P.$$

Ideally we would like to take $R = 2$ so that our security assumptions on the attacker are the same as the assumptions preventing a 51% attack, but in this case this fee structure is only workable up to $P = 1/2$. Moreover, a choice of $P = 1/2$ requires relatively large choices of F and K. Beyond $P = 1/2$ we will have to depend on the attacker's lost deposits in the primary coherence game.

The grief inflicted on the other jurors is NPA, which as we saw above is the net amount of money the attacker takes from them before including the fees the attacker must pay and which are burnt. The cost of the attacker is just the negation of the gain we calculated above. Then, the griefing factor (assuming the attacker does not lose any deposits in the primary coherence game) is:

$$\begin{aligned}\text{griefing factor} &= \frac{NPA}{-M(1-P)(NP-K)+MPF-AK} \\ &\leq \frac{NPA}{-M(1-P)(NP-K)+MP\frac{(1-P)(NP-K)+\frac{K}{R}}{P}-AK} \\ &= \frac{NPA}{\frac{MK}{R}-AK}\end{aligned}$$

The numerator is always non-negative. Assuming $K > 0$, the denominator is positive, and hence the griefing factor is non-negative when

$$A < \frac{M}{R}.$$

Moreover, the griefing factor is bounded by 1 when

$$NPA \leq \frac{MK}{R} - AK \Leftrightarrow A \leq \frac{MK}{R(NP+K)}.$$

Summarizing the results of this section:

Proposition 1. *Suppose*

$$F \geq \frac{(1-P)(NP-K)+\frac{K}{R}}{P}$$

and $A \leq \frac{M}{R}$*, then Attack 1 is not profitable. Moreover, if* $A \leq \frac{MK}{R(NP+K)}$*, this attack has a griefing factor upper bounded by* 1*.*

5 Attacker is incoherent case - $P > 1/2$

Now we look at situations where the attacker votes in the minority, and is betting that others, by voting in the opposite way, are voting for the majority. Namely, P, the proportion of voters who vote for the option that the attacker bets is larger that $1/2$. As discussed above, the dynamic here is substantially different. While the above inequalities above still apply, we saw above that in order for the attack to be non-profitable based on fees, we need

$$\frac{1}{R} \leq 1 - P \Leftrightarrow P \leq 1 - \frac{1}{R}. \tag{3}$$

So already for $R = 1/2$, fees alone cannot make the attack nonprofitable in this case.

However, as the attacker is voting incoherently, she is of course losing deposits in the primary coherence game. So we can consider these in the cost of the attack/in the griefing factors, accepting that by doing so the "true penalty" for being incoherent may be partially hedged as she uses her lost incoherence deposits to finance some other attack that recoups part of that value. We will calculate bounds on the "true cost" of incoherence.

If we include the lost deposit d from the primary coherence game as party of the cost of the attack via the anti-pre-revelation game then, the M voters are divided into three groups:

- MP honest voters who vote X, where $P = \frac{\#\text{ votes for X}}{M} > \frac{1}{2}$ - Eve gains $N(1-P) - F$ by correctly betting on each of these voters
- $M(1-P) - A$ honest voters who vote Y - Eve loses K by incorrectly betting on each of these voters
- A voters controlled by the attacker who vote Y - Eve does not bet on her own votes, but she loses d for each of these votes in the primary coherence game.

Notice that Eve loses at least $\min\{K, d\}$ for each of the $M(1-P)$ many votes for Y, so

$$\text{Eve's return} \leq MP \cdot [N(1-P) - F] - M(1-P) \cdot \min\{K, d\}.$$

So, in order for this attack to be non-profitable to Eve it suffices to have

$$F \geq \frac{NP(1-P) - \min\{K, d\}(1-P)}{P}, \tag{4}$$

which implies

$$\text{Eve's return} \leq 0.$$

Then we have

$$\text{griefing factor} = \frac{\text{grief}}{\text{cost}} \leq \frac{NPA}{-MP \cdot [N(1-P) - F] + M(1-P) \cdot \min\{K, d\}}.$$

It is still true that $A \leq \frac{M}{R}$. However, as the attacking votes must vote in the opposite direction of the bet, we also have $A \leq M(1-P)$, which can provide a useful bound in this case as $P > \frac{1}{2}$ and does not depend on hypotheses about the size of R. Then we have

$$\text{griefing factor} \leq \frac{NP(1-P)}{-P \cdot [N(1-P) - F] + (1-P) \cdot \min\{K, d\}}.$$

Under the assumption on F of inequality 4 the griefing factor is non-negative. Moreover, to have griefing factor ≤ 1 it suffices to have

$$F \geq \frac{2NP(1-P) - \min\{K, d\}(1-P)}{P}.$$

Of course, by counting the attacker's lost deposits in the primary coherence game as part of the cost of this grief, we are allowing the attacker to hedge against incoherent votes via the secondary anti-pre-revelation game. However, as we saw in the discussion around inequalities 3, this is inevitable as if we exclude the lost primary game deposits from the cost of this attack, the attack is profitable and the griefing factors are infinite.

However, we could define C_I, the "true cost of incoherence," to be some proportion of d. Then we can repeat the above calculations where we imagine that each incoherent vote in the primary coherence game contributes $d - C_I$ to the cost of this attack and it is this much of d that can be hedged through such a strategy. Then when analyzing incentives around the primary coherence game, we can assume that an incoherent vote costs the voter C_I. In this framework, repeating the calculations above we have:

Proposition 2. *Suppose Eve performs Attack 1, where she votes for Y and bets that all of the votes she does not control vote for X, and suppose $P = \frac{\#\text{ votes for } X}{M} > \frac{1}{2}$. This attack is not profitable if:*

$$F \geq \frac{NP(1-P) - \min\{K, d - C_I\}(1-P)}{P}.$$

Moreover, the griefing factor for this attack is upper bounded by 1 if:

$$F \geq \frac{2NP(1-P) - \min\{K, d - C_I\}(1-P)}{P}.$$

A complete analysis, which we will pursue in future work, would consider how incentives and the griefing factors of different attacks interact between primary coherence game and the anti-pre-revelation game as C_I is taken to vary.

However, in any event, it seems that the F required in this case are fairly reasonable. Indeed, if

$$2N \leq \min\{K, d - C_I\}$$

then already the griefing factor is bounded by 1 regardless of the choice of F.

6 Properties of this game

In the preceding sections we have discussed how to choose parameters for the anti-pre-revelation game such that

- The attack discussed in Section 3, Attack 1,
  - is not a dominant/even viable strategy under any but the most extreme circumstances (A close to $1/2$) and even then it has bounded impact
  - has reasonable griefing factors, again assuming bounds on A
  - reasonable fees
- Someone who bets correctly on a user's vote will not lose money on the bet after fees - namely $F \leq N(1-P)$.

In this section, we will observe other properties of this game, particularly those that can be guaranteed with judicious choices of parameters.

Proposition 3. *A juror who votes coherently will not lose money in total between the primary coherence game and the anti-pre-revelation game if $d \geq N$.*

Proof. Suppose P is the percentage of jurors who vote in the majority. Then there are MP jurors who vote coherently and $M(1-P)$ who vote incoherently. So the transfer of the incoherent jurors' deposits gives

$$d\frac{M(1-P)}{MP} = d\frac{1-P}{P}$$

to each coherent juror.

On the other hand, jurors can only lose $N(1-P)$ via the anti-pre-revelation game. So a coherent juror that is penalized for pre-revealing gains in total as long as

$$d\frac{1-P}{P} \geq N(1-P) \Leftrightarrow d > NP.$$

Particularly, the statement is satisfied if $d \geq N$.

□

It is debatable whether this property is actually desirable, as it implies a rather low limit on penalty for pre-revelation. On the other hand, this would likely lead to less frustration among honest jurors not actively participating in the anti-pre-revelation game. Note that this property show that costs are concentrated on attacks that *fail* to change the outcome.

We will now analyze the strategies available and the equilibria of the game consisting of the combination of the primary consensus and secondary anti-pre-revelation games. Up to this point we have not discussed any restrictions on who can play the anti-pre-revelation game; in particular, a priori, actors not involved in the primary consensus game could place bets on jurors. We now restrict this:

Design Choice 2. *Only jurors to a dispute may participate in the anti-pre-revelation game for that dispute by placing bets on others' votes.*

We make this restriction essentially so that the amount Eve loses due to being bet upon in the anti-pre-revelation game is directed to the honest voters and can influence their incentives. As the amount lost $N(1-P)$ in the anti-pre-revelation game is larger the smaller P is, the other voters stand to gain more by their bet if fewer voters vote in the same way as Eve. This gives them an incentive to vote in the *opposite* direction from Eve's pre-revealed vote, providing a sort of counteraction to the effect on the incentives of knowing Eve's vote. However, voters must balance this incentive with the primary consensus game, so they still have to imagine what strategy the other honest voters will take.

Restricting betting in this way is reasonable, as ultimately the only communication that matters to the Schelling point is from one voter to another. (If the attacker publishes her vote in a forum but the other voters do not see it, that is not really a problem.)

Each vote has the ability to place a bet on each other vote, but not on itself. If an attacker controls several votes (potentially attached to different addresses), each vote can still place bets on the others. Hence, an attacker can water down their losses from others betting on their vote as discussed in Section 1.3, but only in a way that is limited by the number of votes controlled by this attacker.

7 The $M = 3$ jurors case

We begin by considering the situation for $M = 3$, where our three voters Alice, Bob, and Eve must choose between choices X and Y. Eve votes Y and pre-reveals this vote in a way that is communicated to Alice and Bob and that is absolutely convincing. For now we assume that Eve does not bet on the votes of Alice or Bob (an assumption to be revisited in Remark 3), but that Alice and Bob both (correctly) bet that Eve votes Y, hence dividing the $N(1-P)$ reward from Eve between them.

Below is the payoff tables of this game based on the strategies of Alice and Bob. Note that the game is symmetric for Alice and Bob, so we simply indicate the payoffs to Alice. Here and in our remaining discussion, we denote by $F(x)$, the fee charged for placing a correct bet when $P = x$ (highlighting the dependence of these fees on P, and suppressing their dependence on d, N, M, etc which are viewed as having been chosen prior to the beginning of the game).

Alice's payoffs

Alice \ Bob	vote X, no bet on Alice	vote Y, no bet on Alice	vote X, bet Alice votes X	vote Y, bet Alice votes X	vote X, bet Alice votes Y	vote Y, bet Alice votes Y
vote X, no bet on Bob	$d/2$ $+\frac{N}{3}-\frac{F(1/3)}{2}$	$-d$ $+\frac{N}{6}-\frac{F(2/3)}{2}$	$d/2-N/3$ $+\frac{N}{3}-\frac{F(1/3)}{2}$	$-d-2/3N$ $+\frac{N}{6}-\frac{F(2/3)}{2}$	$d/2+N/3$ $+\frac{N}{3}-\frac{F(1/3)}{2}$	$-d+2N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$
vote Y, no bet on Bob	$d/2$ $+\frac{N}{6}-\frac{F(2/3)}{2}$	0 $+\frac{-F(1)}{2}$	$d/2+N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$	0 $+\frac{-F(1)}{2}$	$d/2-N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$	0 $+\frac{-F(1)}{2}$
vote X, bet Bob votes X	$d/2$ $+\frac{N}{3}-\frac{F(1/3)}{2}$ $+\frac{N}{3}-F(2/3)$	$-d$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $-K$	$d/2-N/3$ $+\frac{N}{3}-\frac{F(1/3)}{2}$ $+\frac{N}{3}-F(2/3)$	$-d-2/3N$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $-K$	$d/2+N/3$ $+\frac{N}{3}-\frac{F(1/3)}{2}$ $+\frac{N}{3}-F(2/3)$	$-d+2N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $-K$
vote Y, bet Bob votes X	$d/2$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $+\frac{2N}{3}-F(1/3)$	0 $+\frac{-F(1)}{2}$ $-K$	$d/2+N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $+\frac{2N}{3}-F(1/3)$	0 $+\frac{-F(1)}{2}$ $-K$	$d/2-N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $+\frac{2N}{3}-F(1/3)$	0 $+\frac{-F(1)}{2}$ $-K$
vote X, bet Bob votes Y	$d/2$ $+\frac{N}{3}-\frac{F(1/3)}{2}$ $-K$	$-d$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $+\frac{N}{3}-F(2/3)$	$d/2-N/3$ $+\frac{N}{3}-\frac{F(1/3)}{2}$ $-K$	$-d-2/3N$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $+\frac{N}{3}-F(2/3)$	$d/2+N/3$ $+\frac{N}{3}-\frac{F(1/3)}{2}$ $-K$	$-d+2N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $+\frac{N}{3}-F(2/3)$
vote Y, bet Bob votes Y	$d/2$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $-K$	0 $+\frac{-F(1)}{2}$ $-F(1)$	$d/2+N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $-K$	0 $+\frac{-F(1)}{2}$ $-F(1)$	$d/2-N/3$ $+\frac{N}{6}-\frac{F(2/3)}{2}$ $-K$	0 $+\frac{-F(1)}{2}$ $-F(1)$

Note that as one must restrict $F(P) \leq N(1-P)$, as discussed in Section 4, $0 \leq F(1) \leq N(1-1) = 0$. So the payoff table simplifies to:

Alice's payoffs

Alice \ Bob	vote X, no bet on Alice	vote Y, no bet on Alice	vote X, bet Alice votes X	vote Y, bet Alice votes X	vote X, bet Alice votes Y	vote Y, bet Alice votes Y
vote X, no bet on Bob	$d/2+N/3$ $-\frac{F(1/3)}{2}$	$-d+N/6$ $-\frac{F(2/3)}{2}$	$d/2$ $-\frac{F(1/3)}{2}$	$-d-N/2$ $-\frac{F(2/3)}{2}$	$d/2+2N/3$ $-\frac{F(1/3)}{2}$	$-d+5N/6$ $-\frac{F(2/3)}{2}$
vote Y, no bet on Bob	$d/2+N/6$ $-\frac{F(2/3)}{2}$	0	$d/2+N/2$ $-\frac{F(2/3)}{2}$	0	$d/2-N/6$ $-\frac{F(2/3)}{2}$	0
vote X, bet Bob votes X	$d/2+2N/3$ $-\frac{F(1/3)}{2}$ $-F(2/3)$	$-d+N/6$ $-\frac{F(2/3)}{2}$ $-K$	$d/2+N/3$ $-\frac{F(1/3)}{2}$ $-F(2/3)$	$-d-N/2$ $-\frac{F(2/3)}{2}$ $-K$	$d/2+N$ $-\frac{F(1/3)}{2}$ $-F(2/3)$	$-d+5N/6$ $-\frac{F(2/3)}{2}$ $-K$
vote Y, bet Bob votes X	$d/2+5N/6$ $-\frac{F(2/3)}{2}$ $-F(1/3)$	0 $-K$	$d/2+7N/6$ $-\frac{F(2/3)}{2}$ $-F(1/3)$	0 $-K$	$d/2+N/2$ $-\frac{F(2/3)}{2}$ $-F(1/3)$	0 $-K$
vote X, bet Bob votes Y	$d/2+N/3$ $-\frac{F(1/3)}{2}$ $-K$	$-d+N/2$ $-\frac{3}{2}F(2/3)$	$d/2$ $-\frac{F(1/3)}{2}$ $-K$	$-d-N/6$ $-\frac{3}{2}F(2/3)$	$d/2+2N/3$ $-\frac{F(1/3)}{2}$ $-K$	$-d+7N/6$ $-\frac{3}{2}F(2/3)$
vote Y, bet Bob votes Y	$d/2+N/6$ $-\frac{F(2/3)}{2}$ $-K$	0	$d/2+N/2$ $-\frac{F(2/3)}{2}$ $-K$	0	$d/2-N/6$ $-\frac{F(2/3)}{2}$ $-K$	0

We now analyze the equilbria of this game. There is the fairly obvious pure strategy equilibrium:

Proposition 4. *The (pure) strategy of voting Y and not placing a bet on the*

other player is a Nash equilibrium in this game when $N - 3F(2/3) \leq 2d$.

Proof. Note that if both Alice and Bob vote Y and do not place bets on each other, they receive a payoff of 0. Then, assuming that Bob plays the strategy of voting Y and not betting on Alice, it is not advantageous for Alice to switch to another strategy as

$$\frac{N}{2} - \frac{3}{2}F(2/3) \leq d$$

and moreover

$$\frac{N}{6} - \frac{F(2/3)}{2} \leq \frac{d}{3} \leq d$$

by our assumption.

□

On the other hand, we also find:

Proposition 5. *Suppose*

- $d > \frac{N}{2}$
- $F(1/3) - F(2/3) \leq \frac{N}{3}$
- $F(1/3) + F(2/3) \leq \frac{2}{3}N$.

Let $\epsilon > 0$. *Then there exists* K_ϵ *such that if* $K \geq K_\epsilon$, *then the above game has a Nash equilibrium such that the probability of voting for* X *is greater than* $1 - \epsilon$.

Note that the conditions on d, N, $F(1/3)$, and $F(2/3)$ are not severe. By equation 2, if K is sufficiently large, $F(1/3)$ can be chosen to be zero and the discussion of Section 4 concerning the profitability and the griefing factor of the attack will still hold (though, one still needs to choose a higher $F(1/3)$ to combat attackers draining their own deposits as discussed in Section 3). Then, as we already impose the constraint $F(2/3) \leq N(1 - 2/3) = \frac{N}{3}$, the conditions on $F(1/3)$ and $F(2/3)$ would hold. Moreover, we saw at the end of Section 5 that if $2N \leq \min\{K, d - C_l\}$, $F(2/3)$ can also be chosen to be any non-negative value while still obtaining a griefing factor of 1.

Proof. Note that the strategy of voting Y and betting the other party also votes Y is weakly dominated by the strategy of voting Y with no bet (regardless of the size of K). Moreover, in an equilibrium where Alice and Bob believe that there is not a positive probability of the other voting and betting Y, for sufficiently large K, the strategy of voting X and betting that the other juror will vote Y is dominated by the strategy of voting Y with no bet as $d > \frac{N}{2}$. Similarly, if Alice and Bob believe that they are in an equilibrium in which there is not a positive probability that the other party bets that they vote Y, the strategy of voting Y and betting that the other votes X weakly dominates the strategy of voting X and betting that the other votes X, using our assumed conditions.

So under these assumptions and for sufficiently large K, Alice and Bob are not advantaged by deviating from an equilibrium consisting of the strategies

$$\{\text{vote } X\text{, no bet; vote } Y\text{, no bet; vote } Y\text{, bet other votes } X\}$$

to one of the other three strategies.

We claim for sufficiently large K, there is an equilibrium whose support is only

$$\{\text{vote } X\text{, no bet; vote } Y\text{, bet other votes } X\}\,.$$

Let p_X and p_Y be the probabilities with which these strategies are adopted. Note that for such an equilibrium to exist, Alice and Bob have to be indifferent between the two strategies, namely:

$$p_X\left(\frac{d}{2}+\frac{N}{3}-\frac{F(1/3)}{2}\right)+p_Y\left(-d-\frac{N}{2}-\frac{F(2/3)}{2}\right)$$

$$=p_X\left(\frac{d}{2}+\frac{5N}{6}-\frac{F(2/3)}{2}-F(1/3)\right)+p_Y(-K)$$

and

$$p_X+p_Y=1.$$

This results in

$$p_X=\frac{K-d-\frac{N}{2}-\frac{F(2/3)}{2}}{K-d-\frac{N}{6}-\frac{F(1/3)}{2}-F(2/3)}$$

and

$$p_Y=1-\frac{K-d-\frac{N}{2}-\frac{F(2/3)}{2}}{K-d-\frac{N}{6}-\frac{F(1/3)}{2}-F(2/3)}.$$

Note that these values are between 0 and 1 when K is sufficiently large and

$$-d-\frac{N}{2}-\frac{F(2/3)}{2}<-d-\frac{N}{6}-\frac{F(1/3)}{2}-F(2/3),$$

which holds by the assumption that $F(1/3)+F(2/3)\leq\frac{2}{3}N$. Moreover, p_X approaches 1 as K grows.

Then it suffices to see that in such an equilibrium, the expected value of deviating to vote Y with no bet is less than the expected payoff of staying in the equilibrium.

$$E(\text{vote } Y \text{ with no bet})=\left(\frac{d}{2}+\frac{N}{6}-\frac{F(2/3)}{2}\right)\frac{K-d-\frac{N}{2}-\frac{F(2/3)}{2}}{K-d-\frac{N}{6}-\frac{F(1/3)}{2}-F(2/3)}$$

whereas

$$E(\text{equilibria strategies})=$$

$$\left(\frac{d}{2}+\frac{5N}{6}-\frac{F(2/3)}{2}-F(1/3)\right)\frac{K-d-\frac{N}{2}-\frac{F(2/3)}{2}}{K-d-\frac{N}{6}-\frac{F(1/3)}{2}-F(2/3)}$$

$$+(-K)\left(\frac{\frac{N}{3}-\frac{F(2/3)}{2}-\frac{F(1/3)}{2}}{K-d-\frac{N}{6}-\frac{F(1/3)}{2}-F(2/3)}\right).$$

So it suffices to see that

$$\left(\frac{2N}{3}-F(1/3)\right)\frac{K-d-\frac{N}{2}-\frac{F(2/3)}{2}}{K-d-\frac{N}{6}-\frac{F(1/3)}{2}-F(2/3)}-K\left(\frac{\frac{N}{3}-\frac{F(2/3)}{2}-\frac{F(1/3)}{2}}{K-d-\frac{N}{6}-\frac{F(1/3)}{2}-F(2/3)}\right)$$

is positive, which holds when K is sufficiently large and

$$F(1/3)-F(2/3)<\frac{2}{3}N.$$

□

So there are Nash equilibria where Alice and Bob both choose Y or are both very likely to choose X (potentially there are other equilibria as well). Then their votes will be determined by which of the viable strategies each thinks the other will adopt, where they have an incentive to be coherent. Namely, despite Eve's pre-revealed vote, each of Alice and Bob have an incentive to vote as another (a priori honest but self-interested) juror would vote. Note that this is already the case for $M = 3$, where a third of the vote is controlled by a hostile pre-revealing party. As we continue this work, we will look at larger M from a model where jurors have some Bayesian priors on what other jurors will consider the honest vote to be, and observe how large A needs to be relative to M for a pre-revelation attack to transform voting with the attacker into a dominant strategy.

Remark 3. *If Eve bets on each of Alice and Bob to vote X, then this changes the payoff table examined above. In particular, this can alter whether an equilibrium of the type seen in Proposition 5 exists. However, by pushing Alice and Bob to vote Y, this bet functions as Eve paying K to give Alice and Bob each a smaller bribe. Moreover, Eve's decision to make this bet will only occur after each of Alice and Bob have casted their votes, so for it to affect Alice and Bob's voting patterns, it would have to be credible to them that Eve will bet in this way in the future even if at that point Eve has no strict incentive to make such bets. We expect it to make more sense for Eve to make more traditional bribes, though we will further analyze this behavior.*

8 Pure strategy equilibria for larger M

We now consider pure strategy equilbria when the number of jurors is $M > 3$. We suppose that the attacker Eve controls $V_E < M/2$ votes which are pre-revealed as voting for Y. We will denote by V_A the largest number of votes controlled by any of jurors other than Eve. (We think of V_A as being the number of votes controlled by Alice.) Our results will include hypotheses on V_A.

We assume that there are no additional pre-revelation attacks or bribes that otherwise influence the payoffs. We assume that each of the $M-1$ other vote places a correct bet on each of Eve's votes (including the bets from the other votes controlled by Eve as discussed in Section 1.3 and in the comments following Design Choice 2). However, as in Section 7 and Remark 3, we assume that Eve does not bet on the other jurors.

Similar to Proposition 4, we find:

Proposition 6. *Suppose that $V_A < M/2$ and $V_E < M/2$. Then the (pure) strategy of voting Y and placing a bet that all other players will vote Y is a Nash equilibrium when*

$$d \geq 2N \left(\frac{1}{M} + \frac{M}{4(M-1)} + \frac{M}{M-2} \right).$$

Proof. Suppose that all of the $M - V_A$ votes are for Y and that the $M - V_A - V_E$ non-attacking votes not controlled by Alice are accompanied by bets that all other players vote Y.

We consider the payoffs available to Alice as she places her V_A votes. These payoffs decompose into

- the payoff from the primary coherence game
- payoffs from bets placed on Eve's votes
- payoffs from bets Alice places on voters other than Eve (including potentially betting on her own votes) and
- gains and losses from bets other voters place on Alice.

As $M - V_A > M/2$, Alice's votes can not change the result of the dispute. Hence, if Alice casts V_{AX} votes for X and $V_A - V_{AX}$ votes for Y, her return from the primary coherency game is

$$-d \cdot V_{AX} + \frac{V_{AX} \cdot d}{M - V_{AX}}(V_A - V_{AX}) = V_{AX} \cdot d \left[\frac{V_A - M}{M - V_{AX}} \right]. \tag{5}$$

The collective payoff to Alice for betting on Eve's V_E votes is

$$V_E \cdot \left[N \frac{V_{AX}}{M} - F\left(1 - \frac{V_{AX}}{M}\right) \right] \cdot \frac{V_A}{M-1}, \tag{6}$$

whereas the payoff for betting that the $M - V_A - V_E$ votes controlled neither by Alice nor Eve vote Y is

$$(M - V_A - V_E) \cdot \left[N \frac{V_{AX}}{M} - F\left(1 - \frac{V_{AX}}{M}\right) \right] \cdot \frac{V_A}{M - V_E - 1}. \tag{7}$$

(The difference between these expressions is due to the fact that Eve bets on her own votes when possible, so each of Eve's votes is bet upon by each of the other

$M-1$ votes, whereas Eve is assumed not to bet on the other votes.) Note that Alice has strictly no incentive to place a bet that any of these votes will be for X as this would result in receiving a smaller share of the $N\frac{V_{AX}}{M} - F\left(1-\frac{V_{AX}}{M}\right)$ than she otherwise would as well as assuming a cost of K while not otherwise influencing Alice's payouts.

Then, we have assumed that the $M - V_A - V_E$ other honest participants adopt the strategy of betting that everyone (including Alice) will vote Y. For each of Alice's V_{AX} votes for X, Alice will receive a reward from these bets, and for her remaining $V_A - V_{AX}$ she will have to make a payout on these bets, mitigated somewhat by her ability to bet on each of her own votes with her $V_A - 1$ other votes. Thus the net effect of these bets on Alice is

$$V_{AX}\cdot N\left(1-\frac{V_{AX}}{M}\right)+(V_A-V_{AX})\cdot\left[-N\frac{V_{AX}}{M}+\left(N\frac{V_{AX}}{M}-F\left(1-\frac{V_{AX}}{M}\right)\right)\cdot\frac{V_A-1}{M-V_E-1}\right]. \tag{8}$$

Adding expressions 5 - 8 and simplifying, we get Alice's total payoff as

$$V_{AX}\cdot d\left(\frac{V_A-M}{M-V_{AX}}\right)+\left(1-\frac{V_A}{M}\right)\cdot N\frac{V_{AX}}{M}$$
$$+\left[N\frac{V_{AX}}{M}-F\left(1-\frac{V_{AX}}{M}\right)\right]\cdot\left(\frac{V_A\cdot V_E}{M-1}+\frac{MV_A-V_AV_E-V_AV_{AX}-V_A+V_{AX}}{M-V_E-1})\right). \tag{9}$$

Then, using the assumptions that V_A and V_E are bounded by $M/2$ and $V_{AX} \le V_A \le M$, we have $V_{AX}\cdot d\left(\frac{V_A-M}{M-V_{AX}}\right) \le -\frac{d}{2}\cdot V_{AX}$. Similarly, $\left(1-\frac{V_A}{M}\right)\cdot N\frac{V_{AX}}{M} \le \frac{N}{M}\cdot V_{AX}$. Hence, using $F\left(1-\frac{V_{AX}}{M}\right) \ge 0$ and again $V_{AX} \le V_A$, the payout is upper-bounded by

$$\left[-\frac{d}{2}+\frac{N}{M}+\frac{N}{M}\left(\frac{V_A\cdot V_E}{M-1}+\frac{M\cdot V_A}{M-V_E-1}\right)\right]\cdot V_{AX}.$$

Using again that V_A and V_E are bounded by $M/2$, the payout is bounded by

$$\left[-\frac{d}{2}+N\left(\frac{1}{M}+\frac{M}{4(M-1)}+\frac{M}{M-2}\right)\right]\cdot V_{AK}.$$

Thus, by our assumption on N, d, and M, this is non-positive. However, by expression 8, the payoff is zero when $V_{AK} = 0$ as $F(1) = 0$ by the comments of Section 6. Hence the strategy of using all of one's votes to vote Y and bet that the other jurors will also vote Y is an equilibrium.

□

Proposition 7. *Suppose V_E, $V_A \ge 1$ and $V_E + V_A < M/2$ Then the (pure) strategy of voting X and placing a bet that all other votes other than those pre-revealed by Eve will vote X is a Nash equilibrium when*

$$d \ge N\left(\frac{2M}{M-2}+3\right)+\frac{M^2}{8}N\left(\frac{1}{2M-2}+\frac{1}{M^2-2M}+\frac{M-2}{M^2}\right).$$

Proof. We follow a similar argument to that of Proposition 6. Suppose that all of the $M - V_A - V_E$ non-attacking votes not controlled by Eve are for X and are accompanied by a bet that all other all of these other votes will also vote X. Suppose Alice casts V_{AY} votes for Y from among her V_A total votes. As $V_E + V_A < M/2$, Alice can again not change the outcome of the dispute. So, her return from the primary coherence game is

$$-d \cdot V_{AY} + d\frac{V_E + V_{AY}}{M - V_E - V_{AY}}(V_A - V_{AY}).$$

The payoff resulting from Alice's bets that Eve's V_E many votes vote Y is

$$V_E \cdot \left[N\left(1 - \frac{V_{AY} + V_E}{M}\right) - F\left(\frac{V_{AY} + V_E}{M}\right)\right] \cdot \frac{V_A}{M-1}.$$

Similarly, the payoff from betting that the $M - V_A - V_E$ votes controlled neither by Alice nor by Eve vote X is

$$(M - V_A - V_E) \cdot \left[N\frac{V_{AY} + V_E}{M} - F\left(1 - \frac{V_{AY} + V_E}{M}\right)\right] \cdot \frac{V_A}{M - V_E - 1}.$$

Then the net payout from bets placed upon Alice's votes (including by herself) is

$$V_{AY} \cdot N\left(1 - \frac{V_{AY} + V_E}{M}\right)$$

$$+(V_A - V_{AY}) \cdot \left[-N\frac{V_{AY} + V_E}{M} + \left(N\frac{V_{AY} + V_E}{M} - F\left(1 - \frac{V_{AY} + V_E}{M}\right)\right) \cdot \frac{V_A - 1}{M - V_E - 1}\right].$$

Similar comments amount the lack of incentives to bet that the other honest jurors vote Y apply as in Proposition 6, so it suffices to show that this payoff is maximized when $V_{AY} = 0$.

We take the sum of these terms to be Payoff(v_{AY}). Suppose that there exists some $v^*_{AY} \in \mathbb{Z}_{\geq 0}$ strictly greater than zero that gives a higher payout that Payout(0).

We have assumed little about the function F beyond that $0 \leq F(p) \leq N(1-p)$ for all $p \in [0,1]$. We decompose

$$\text{Payout}(V_{AY}) = \text{Payout}_0(V_{AY}) - \text{Fees portion}(V_{AY}),$$

where

$$\text{Fees portion}(V_{AY}) =$$

$$\frac{V_E \cdot V_A}{M-1}F\left(\frac{V_{AY} + V_E}{M}\right) + \frac{(M - V_A - V_E) \cdot V_A}{M - V_E - 1}F\left(1 - \frac{V_{AY} + V_E}{M}\right) + \frac{(V_A - V_{AY}) \cdot (V_A - 1)}{M - V_E - 1}F\left(1 - \frac{V_{AY} + V_E}{M}\right).$$

Note that

$$\text{Fees portion}(0) - \text{Fees portion}(V^*_{AY})$$

$$\leq \frac{V_E \cdot V_A}{M-1} \cdot N\left(1-\frac{V_E}{M}\right) + \left(\frac{(M-V_A-V_E)\cdot V_A}{M-V_E-1} + \frac{(V_A-V_{AY})\cdot(V_A-1)}{M-V_E-1}\right)\cdot N\left(\frac{V_E}{M}\right),$$

as, in the most extreme case $F\left(\frac{V_E}{M}\right)$ and $F\left(1-\frac{V_E}{M}\right)$ could attain their maximum values of $N\left(1-\frac{V_E}{M}\right)$ and $N\left(\frac{V_E}{M}\right)$ respectively while F could be zero at all other values. Simplifying this and using the assumptions that $V_A + V_E \leq M/2$ which in particular implies that $V_A \cdot V_E \leq M^2/16$, we have

$$\text{Fees portion}(0) - \text{Fees portion}(V_{AY}^*) \leq \frac{M^2}{16} N\left(\frac{1}{2M-2} + \frac{1}{M^2-2M} + \frac{M-2}{M^2}\right).$$

Note that as $V_E + V_{AY} < M/2$, Payout_0 is differentiable with respect to V_{AY} for $V_{AY} > 0$.

We compute the derivative

$$\frac{d}{dV_{AY}}\left(-d\cdot V_{AY} + d\frac{V_E+V_{AY}}{M-V_E-V_{AY}}(V_A-V_{AY})\right) = dM\frac{-M+V_E+V_A}{(M-V_E-V_{AY})^2} \leq -\frac{d}{2},$$

where we use the fact that $V_E + V_A < M/2$.

Then

$$\text{Payoff}_0'(V_{AY}) \leq -\frac{d}{2} - V_{AY}\frac{N(V_E-1)(V_A-1)}{M(M-V_E-1)} + \text{CNST},$$

where

$$\text{CNST} = (M-V_E-V_A)\frac{V_A\cdot N}{(M-V_E-1)M} - \frac{V_E\cdot V_A\cdot N}{(M-1)M} + N - \frac{N\cdot V_E(V_E-1)}{M(M-V_E-1)} + V_A\left(-\frac{N}{M} + \frac{V_A-1}{M-V_A-1}\cdot\frac{N}{M}\right).$$

Note that using our bounds that $V_A + V_E \leq M/2$ we have

$$\text{CNST} \leq N\left(\frac{M}{M-2} + \frac{3}{2}\right).$$

By the Mean Value Theorem, there exists some $c > 0$ such that

$$\text{Payout}_0(V_{AY}^*) - \text{Payout}_0(0) = \text{Payout}_0'(c)\cdot V_{AY}^*.$$

So

$$\text{Payout}(V_{AY}^*) - \text{Payout}(0) = \text{Payout}(V_{AY}^*) - \text{Payout}_0(V_{AY}^*) + \text{Payout}_0(V_{AY}^*) - \text{Payout}_0(0) + \text{Payout}_0(0) - \text{Payout}(0)$$

$$= \text{Payout}_0(V_{AY}^*) - \text{Payout}_0(0) + \text{Fees portion}(0) - \text{Fees portion}(V_{AY}^*)$$

$$\leq \text{Payout}_0'(c)\cdot V_{AY}^* + \frac{M^2}{16}N\left(\frac{1}{2M-2} + \frac{1}{M^2-2M} + \frac{M-2}{M^2}\right)$$

As V_{AY}^* is assumed to be a positive integer and by our bound on Payout_0', and using the fact that our assumed bound on d implies

$$-\frac{d}{2} + N\left(\frac{M}{M-2} + \frac{3}{2}\right) \leq 0,$$

we have

$$\text{Payout}_0'(c)V_{AY}^* \leq -\frac{d}{2} + N\left(\frac{M}{M-2} + \frac{3}{2}\right).$$

Then

$$\text{Payout}(V_{AY}^*) - \text{Payout}(0) \leq -\frac{d}{2} + N\left(\frac{M}{M-2} + \frac{3}{2}\right) + \frac{M^2}{16}N\left(\frac{1}{2M-2} + \frac{1}{M^2-2M} + \frac{M-2}{M^2}\right) \leq 0.$$

So the the strategy of taking $V_{AY} = 0$ is an equilibrium.

□

Note that the constraint on d in Proposition 7 in terms of M can become problematic as M becomes large. This is essentially due to the contribution to the bound of the worst case handling of the fees. However, this was a rather extreme case, where the fee at $F\left(\frac{V_E}{M}\right)$ and $F\left(1 - \frac{V_E}{M}\right)$ was maximized, and the fee at every other value was zero. Namely, this reflects a very sharp discontinuity in the fees.

In exchange for making additional assumptions on F we can prove:

Proposition 8. *Suppose V_E, $V_A \geq 1$ and $V_E + V_A < M/2$. Furthermore, let $R > 0$ and assume that*

$$|F(a) - F(b)| \leq \frac{1}{R}|N(1-a) - N(1-b)| = \frac{N}{R}|a - b|$$

for all possible vote percentages a and b. Then the (pure) strategy of voting X and placing a bet that all other votes other than those pre-revealed by Eve will vote X is a Nash equilibrium when

$$d \geq N\left(\frac{2M}{M-2} + 3\right) + \frac{N}{2R}\left(\frac{M^2}{8M-8} + M\right).$$

Proof. The proof of this proposition is exactly like that of Proposition 7 except that we will have a more precise bound on

$$\text{Fees portion}(0) - \text{Fees portion}(V_{AY}^*)$$

$$= \frac{V_E \cdot V_A}{M-1}\left[F\left(\frac{V_E}{M}\right) - F\left(\frac{V_{AY}^* + V_E}{M}\right)\right] + V_A\left[F\left(1 - \frac{V_E}{M}\right) - F\left(1 - \frac{V_{AY}^* + V_E}{M}\right)\right] - \frac{V_{AY}^*(V_A - 1)}{M - V_E - 1}F\left(1 - \frac{V_{AY}^* + V_E}{M}\right)$$

So by our continuity assumptions on the fees,

$$\text{Fees portion}(0) - \text{Fees portion}(V_{AY}^*) \leq |\text{Fees portion}(0) - \text{Fees portion}(V_{AY}^*)|$$

$$\leq \frac{V_E \cdot V_A}{M-1}\left[\frac{NV_{AY}^*}{RM}\right] + V_A\left[\frac{NV_{AY}^*}{RM}\right] \leq \frac{N}{4R}\left(\frac{M^2}{8M-8} + M\right),$$

where we have used the assumption that $V_E + V_A < M/2$.

Then

$$\text{Payout}(V^*_{AY}) - \text{Payout}(0) \leq -\frac{d}{2} + N\left(\frac{M}{M-2} + \frac{3}{2}\right) + \frac{N}{4R}\left(\frac{M^2}{8M-8} + M\right).$$

This is non-positive by out assumptions on d. So the the strategy of taking $V_{AY} = 0$ is again an equilibrium.

□

The choice of the parameter R, namely the strength of the continuity assumptions imposed on F, can be tuned to the application. A choice of $R = M$ might be generally acceptable. This forces assumptions on d and N that are reasonable regardless of the choice of M. Assuming that

$$|F(a) - F(b)| \leq \frac{N}{M}|a - b|$$

for all a and b imposes substantial constraints on the size of F when M is large; however, we have seen in Section 3 that the principal role of positive valued F is to limit the ability of users to drain their own anti-pre-revelation deposit, particularly as Attack 1 can be controlled through adjusting K. If M is large, the pool of people who have the ability to bet on an attacker's vote is large enough that the threat of deposit draining attacks should be correspondingly reduced.

9 Conclusion

As we saw in Section 1.2 that there are highly varied pre-revelation attacks that are possible, we have reiterated the need for a flexible and general solution to the problem of pre-revelation in Schelling point based crypto-economic systems. We have provided a analysis of the solution proposed in [1], which appears to be the most promising response to this problem. Particularly, while we have pointed out an attack on this proposal, we see how to adjust fees in such a way that this attack is not viable. Moreover, we have explored parameter choices that incentivize honest jurors to still tend toward equilibria where they agree among each other even in the presence of a pre-revelation attack. We have even shown that such equilibria exist in the case of a three juror vote where one juror pre-reveals. Namely, under these appropriate parameter choices, we have seen that the honest jurors have an incentive to consider what the Schelling point would be among the remaining jurors that have not pre-revealed. This analysis lends support to the idea that an anti-pre-revelation game will be a workable mechanism for systems such as Kleros.

References

[1] Vitalik Buterin. On anti-pre-revelation games. Online, `https://blog.ethereum.org/2015/08/28/on-anti-pre-revelation-games/`. August 2015.

[2] Karl Floersch. Learning solidity part 2: Commit-reveal voting. Online, https://karl.tech/learning-solidity-part-2-voting/.

[3] Clément Lesaege. Attack on early reveal by deposit. https://github.com/kleros/kleros-attacks/blob/master/contracts/early-reveal-penalization/DepositForEarlyReveal.sol, 2018. commit: 5243a90f882225b370b5cdecaaa8f017176ff9f5.

[4] Clément Lesaege. Cut and choose attack for early revelation. https://github.com/kleros/kleros-attacks/blob/master/contracts/early-reveal-penalization/CutAndChooseForEarlyReveal.sol, 2018. commit: 52c9b4837dc65b9388aaa3170099af7b542352d1.

[5] Clément Lesaege and Federico Ast. Kleros: Short paper v1.0.5. https://kleros.io/assets/whitepaper.pdf. January 2018.

[6] Yehuda Lindell and Benny Pinkas. An efficient protocol for secure two-party computation in the presence of malicious adversaries. *J. Cryptology*, 28(2):312–350, 2015.

[7] Jack Peterson and Joseph Krug. Augur: a decentralized, open-source platform for prediction markets. https://bravenewcoin.com/assets/Whitepapers/Augur-A-Decentralized-Open-Source-Platform-for-Prediction-Markets.pdf. January 2018.

[8] T.C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1980.

[9] Paul Sztorc. Truthcoin: Peer-to-peer oracle system and prediction marketplace. https://www.truthcoin.info/papers/truthcoin-whitepaper.pdf.