An Analysis of $p+\varepsilon$ Attacks on Various Models of Schelling Game Based Systems

WILLIAM GEORGE, Kleros Cooperative
CLÉMENT LESAEGE, Kleros Cooperative

Smart contracts allow flexible new forms of trust-free cooperation. This gives mechanism designers a new platform in which they can apply game theoretical principles to engineer desirable outcomes. However, as we have seen with the emergence of "dark DAOs", hostile attackers can just as easily deploy smart contracts that coordinate collusion and otherwise attempt to undermine the goals of a given game theoretical structure. Indeed, attacks on incentive structures that would previously have seemed elaborate and difficult to organize can now execute automatically and requiring no trust in the attacker via smart contracts. In this work we analyze the "$p+\varepsilon$ attack" on Schelling point based cryptoeconomic systems. Particularly, we consider design choices that one can apply, slightly modifying these systems, to increase resistance to these types of attacks.

1. INTRODUCTION AND PAST WORK

In a coordination game, participants are allowed to choose from several options, and they have an incentive to choose the option that is also chosen by other participants [Cooper 1999]. A standard example of this phenomenon is the process of deciding whether to drive on the left or right side of a road. To the degree that the payoff matrix is symmetrical between the various options, which option participants coordinate on can be thought of as somewhat arbitrary.

	X wins	Y wins
$\mathcal{USR}$ votes X	p	0
$\mathcal{USR}$ votes Y	0	p

Fig. 1. An example payoff matrix for a voter $\mathcal{USR}$ in a symmetrical coordination game.

However, as was noted by Thomas Schelling [Schelling 1980], when one of the options is somehow distinguished, it can serve as a "focal point" or "Schelling point" on which participants can coordinate. This effect is particularly relevant when participants lack other means of communication that they could use to coordinate. This idea is the basis for many blockchain "oracles", i.e. systems used to import information onto a blockchain platform that is not checked as part of the consensus protocol of that blockchain. Here, the principle is that if a group of voters are asked to choose between options while reporting on some question about the external world, they will tend to gravitate towards the true response, if only because they view it as a distinguished answer that other voters will expect them to choose. Such oracles are important for reporting on-chain results of elections for prediction market contracts, rainfall amounts for drought insurance contracts, etc. Moreover, this structure also arises implicitly in fundamental ways in blockchain systems; in [Buterin 2015] it is pointed out that in consensus algorithms based on proof-of-work and the longest chain rule such as Bitcoin [Nakamoto 2009], choice of which fork to follow can be thought of as "votes" having the structure of a Schelling game. See Section 4 for further details on these examples.

An intriguing new attack, the $p+\varepsilon$ attack, designed to warp the incentive structures of these games was proposed in a blog post by Vitalik Buterin [Buterin 2015], crediting the idea for the attack to Andrew Miller. In general, a $p+\varepsilon$ attack that attempts to make the malicious choice Y win should pay to participants who vote Y an additional ε more than what they would receive for voting X, *contingent on X winning*. In the case the payoff matrix of Table 1, this yields a payoff matrix of:

	X wins	Y wins
$\mathcal{USR}$ votes X	p	0
$\mathcal{USR}$ votes Y	$p+\varepsilon$	p

Then voting Y is a weakly dominant strategy in this game, so the attacker might hope that the vast majority of participants will vote Y causing the attacker's choice to win even as she is not required to pay out any bribes. Hence, when participants are rational and self-interested this attack can warp the incentives of the system with no costs to the attacker, i.e. "for free". In fact, this idea of a bribe that is only paid out in the case that the bribe is unsuccessful in changing the outcome was already considered in a thought experiment proposed by Warren Buffett as part of an editorial written for the New York Times in 2000 highlighting absurd aspects of the United States campaign finance system and arguing for campaign finance reform [Buffett 2000]. Recently, analysis of blockchain governance has remarked that situations similar to the $p+\varepsilon$ have existed in more traditional governance situations, particularly surrounding share buying offers in acquisitions of companies [Abramowicz 2019]. Also, the $p+\varepsilon$ attack has been considered in classifications of types of attacks on cryptoeconomic systems [Deirmentzoglou et al. 2019] particularly towards the perspective of whether these attacks are viable on proof-of-stake consensus systems as well as proof-of-work systems [Wang 2019], [Sayeed and Marco Gisbert 2018]. Nonetheless, considering the widespread use of Schelling based structures (explicitly or implicitly) in blockchain based systems, this attack has so far received relativley little attention from the research community.

2. POSSIBLE STRUCTURES OF SCHELLING GAMES

In this section, we recall various aspects that can be considered in the design of a Schelling game.

— Users can be required to place a deposit d which they lose if their vote is incoherent with the ultimate result. Thus, this allows penalties to voter as well as rewards.
— We can also have appeals. Appeals result in delaying the result of a vote by an additional t Ethereum blocks per appeal, up to some maximum number of appeal rounds. Each appeal will typically consume some additional resources R, with the idea that more resources can be put towards this ruling, so ideally it is more likely to be correct. Concretely, if we suppose that there are M voters in an appeal round, a Schelling game of the form of that in the introduction can require up to $R \leq Mp$ resources to provide for payouts to voters. There are different models for how this amount is collected - it could be provided by the losing side of a previous appeal round, both sides could be compelled to provide a deposit that covers the amount required with the winning side refunded, in the context of voting in blockchain consensus mechanisms this amount may come from block rewards, etc. However, the exact model for how appeals are financed is not in the scope of this work; instead for a fixed total value of resources contributed to a decision across all appeal rounds, we compare the resistance of different structures to attacks. Of particular note for our purposes, we generally consider the winner of a game, as used for determining which payouts are made to be the option that receives the most votes in the *last* appeal round.
— Other contracts can be launched to a blockchain platform that interact with a Schelling game and with any $p+\varepsilon$ attack contract. Hence, it is possible for participants of a Schelling game to organize themselves or "counter-coordinate" in a way to counter-attack an attacker, changing the nature of the game. See Section 6.

Moreover, one can construct different payoff matrices for games that nonetheless keep the basic idea of a Schelling game. The payoff matrix that we saw in the Introduction, when modified to include deposits is:

Definition 2.1. We define a simple Schelling game to be one where a participant $\mathcal{USR}$ can vote one of two options with the payoffs given by:

	X wins	Y wins
$\mathcal{USR}$ votes X	p	$-d$
$\mathcal{USR}$ votes Y	$-d$	p

A $p+\varepsilon$ attack transforms the payoff matrix to:

	X wins	Y wins
$\mathcal{USR}$ votes X	p	$-d$
$\mathcal{USR}$ votes Y	$p+\varepsilon$	p

Concretely, this is achieved as the attacker pays to $\mathcal{USR}$ an amount of $p+d+\varepsilon$ if $\mathcal{USR}$ finds herself in the case of the bottom-left entry of the table.

Alternatively, we could also divide all amounts - both the positive sum resources that the participants can receive as well as the lost deposits - between coherent voters (in a given appeal round). This has the effect that coherent voters receive a larger reward in the event of narrow decisions.

Definition 2.2. We define a redistributive Schelling game to be one where a participant $\mathcal{USR}$ contributes a deposit d and can vote for one of two options with the payoffs given by:

	X wins	Y wins
$\mathcal{USR}$ votes X	$\frac{(M-x-1)d+Mp}{x+1}$	$-d$
$\mathcal{USR}$ votes Y	$-d$	$\frac{xd+Mp}{M-x}$

where x is the number of votes (other than that of $\mathcal{USR}$) for X.

A $p+\varepsilon$ attack on this game gives the following payoff table:

	X wins	Y wins
$\mathcal{USR}$ votes X	$\frac{(M-x-1)d+Mp}{x+1}$	$-d$
$\mathcal{USR}$ votes Y	$\frac{(M-x-1)d+Mp}{x+1}+\varepsilon$	$\frac{xd+Mp}{M-x}$

Note that, in a redistributive Schelling game, if a given voter is the only person to vote for X in her round and then X ultimately wins, she receives all of the Mp and all of the lost deposits for her round. Then in order for an attacker to make voting Y a dominant strategy and to incentivize to not take a chance on trying to get this "lone voice of reason jackpot" an attacker needs to being willing to offer bribes that are $O(M)$ to each voter in the most extreme case.

Finally we consider,

Definition 2.3. We define a symbiotic Schelling game to be one where a participant $\mathcal{USR}$ contributes a deposit d and can vote for one of two options with the payoffs given by:

	X wins	Y wins
$\mathcal{USR}$ votes X	$\frac{p(x+1)}{M}$	$-d$
$\mathcal{USR}$ votes Y	$-d$	$\frac{p(M-x)}{M}$

where x is the number of votes (other than that of $\mathcal{USR}$) for X.

A $p+\varepsilon$ attack on this game gives the following payoff table:

	X wins	Y wins
$\mathcal{USR}$ votes X	$\frac{p(x+1)}{M}$	$-d$
$\mathcal{USR}$ votes Y	$\frac{p(x+1)}{M}+\varepsilon$	$\frac{p(M-x)}{M}$

Note that these three structures are normalized so that they have the same maximum total payout to voters - Mp. However, the full Mp is only guaranteed to be paid out in the redistributive Schelling game case.

Deposits, (a form of) appeal systems[1], and counter-coordination were all already considered as possible defenses against $p+\varepsilon$ attacks in [Buterin 2015] in the context of Schelling games of the form of Definition 2.1. The possibility that participants can lose deposits has already been evoked

[1] In [Buterin 2015] the idea that one "use[s] round $N+1$ to determine who should be rewarded during round N," is considered.

as a basis for increased security against $p+\varepsilon$ attacks in the context of proof-of-stake consensus algorithms [Wang 2019] (compared to proof-of-work where miners do not stake a deposit). However, the interplay between these elements, and particularly how they interact with the choice of games of the form of Definition 2.1, 2.2, or 2.3 seems to have not been fully explored.

In this work, our attacker will generally have the ability to offer some type of $p+\varepsilon$ bribe. In situations where there is an appeal structure, we typically think of this structure as being a defense; however, is a double-edged sword as we then also allow our attacker the ability to make malicious appeals. However, we do not consider attackers with the ability to tamper directly with votes. In particular, we do not consider attackers capable of performing 51% attacks or denial of service attacks on the consensus protocol of an underlying blockchain platform where the vote is taking place, nor do we consider attackers that have votes themselves or that are capable of directly controlling users' votes, e.g. via malware. Voters are assumed to be rational (possibly assigning costs to immoral actions, see Section 5) and either uncoordinated or coordinated in specified ways.

3. OUR CONTRIBUTION

In this work we extend the analysis presented in [Buterin 2015] on simple Schelling games of the form of Definition 2.1 to Schelling games of the form of Definitions 2.2 and 2.3. We see how various defenses against $p+\varepsilon$ attacks, such as deposits, appeal systems, and counter-coordination interact with each other and with the choice of payoff structure. In Section 5, we consider the situation where rational participants assign a moral (or cognitive effort) cost to accepting the bribe and we consider equilibria where the attack sometimes fails. In Section 7, we see that, under a variety of measures of difficulty of an attack such as expected cost of the attack and budget required, redistributive games, particularly when there is an appeal structure, have increased resistance to $p+\varepsilon$ attacks. We see that, under certain conditions, a "pure" $p+\varepsilon$ attack on a redistributive Schelling game with an appeal system can have budget requirements which grow quadratically in the number of participants and are hence likely prohibitive. Moreover, whereas [Buterin 2015] saw that counter-coordinating is a dominated strategy for games of the form of Definition 2.1, we see that equilibria where counter-coordinating occurs can exist for games of the form of Definition 2.2. However, in Section 8, we consider several ways in which the attacker can adapt the spirit of a $p+\varepsilon$ attack to redistributive games so that the attacks are less prohibitive. Nevertheless, these adapted versions seem to be less dangerous than the "free" $p+\varepsilon$ attacks on simple Schelling game systems. Finally we consider the results of some preliminary experiments around $p+\varepsilon$ attacks conducted on the Kleros platform.

4. EXAMPLES

A number of blockchain-based systems involve voting that, implicitly or explicitly, can be seen in the context of Schelling point based games.

— A number of token-curated lists involve voting on candidates for admission to the list where voters are rewarded via a Schelling point game, see [Goldin 2017]. Notable examples include lists of true or fake news for decentralized news aggregation services or lists of products that meet certain criteria for sale in decentralized marketplaces. Here the voters for the list, typically a pool of token holders, vote on whether a given entry deserves admission to the list, with rewards (and in some cases penalties) applied to voters based on whether or not they agree with the majority vote. This process serves as an oracle for blockchain applications that use these lists to gain information about whether submissions satisfy the listing criteria. See [Goldin et al. 2017], [Civil 2017] for examples.
— Truthcoin [Sztorc 2015] proposes a more general purpose oracle system based on a Schelling point mechanism; however, its payoff structures are somewhat more complicated than those considered in Section 2 as Truthcoin creates an interdependence of votes on different questions, which are then resolved using a singular value decomposition on the matrix of votes.
— Kleros [Lesaege and Ast 2018] is a blockchain based system for dispute resolution that functions on a Schelling point based mechanism. It is then a type of oracle that is specialized in bring-

ing on-chain information necessary to resolve disputes. In any given case, for example between two parties to a small scale freelancing dispute, a number of "jurors" are drawn randomly from among a pool of token holders. Then these "jurors" vote on the result of the dispute, and they are rewarded or penalized based on whether they vote for the majority result. Note that Kleros uses a redistributive Schelling game structure and has an explicit appeal mechanism satisfying the properties described in Section 1.

— One of the original examples of [Buterin 2015] is that of the proof-of-work consensus mechanism. A given block producer can expect to receive p in block rewards, but only if they are on their blocks are ultimately in the longest chain. Then suppose that we have two rival chains on which one can mine, one where the transaction X has been accepted and one where the conflicting transaction Y has been accepted. Over short time frames (compared to the difficulty adjustment period) then mining a block on each chain would require approximately equal resources and produce a constant reward dependent only on the corresponding chain ultimately winning out. Hence for each block someone accepting the $p+\varepsilon$ bribe mines that includes Y on a chain that eventually settles on X as the valid transaction, the attacker must pay the block rewards plus ε. This structure then resembles that of Definition 2.1.

  However, imagine heuristically that we have very short block times (relative to the period of time in which the attacker needs to wait to have a sufficient number of block confirmations) and that difficulty adjusts continuously. Then suppose a single miner Bob that represents a proportion of z of all mining resources is mining on the transaction X chain for some period of time while all other minings, representing $1-z$ resources are mining on the Y chain. This can be thought of as z percent of the "votes" being cast for X and $1-z$ percent being cast for Y over this period. By considering longer amounts of time (heuristically discretizing them into distinct periods) we can think of the votes as being broken into appeal periods in the sense of Section 1. From the perspective of the X chain Bob should get all of the block rewards over this period of time. If the option X ultimately gives the longest chain (beyond the number of blocks required for "coinbase maturity") then all of these rewards that Bob receives will be usable. In order to incentivize rational miners who are willing to accept bribes but do not think that the attacker has the resources to maintain the attack over the long term to not be tempted by this "lone voice of reason jackpot", it is not sufficient for the attacker to pay the miners who support Y on a per block basis. Rather she must pay them the amount that they would have received by mining on the other chain during this period, up to the total block rewards plus epsilon. This structure resembles that of Definition 2.2.

— The future Serenity proof-of-stake version of Ethereum [Buterin 2019] is planned to use the symbiotic Schelling game model of Definition 2.3. This choice was motivated to make the protocol less vulnerable to "discouragement attacks" [Buterin 2019], [Buterin 2018]. Note that as this protocol, in contrast to proof-of-work, is designed to have finalization guarantees, one cannot consider waiting extra blocks as an "appeal" in the same way.

5. MORAL AND COGNITIVE COSTS

Even though, in the abstract, accepting a $p+\varepsilon$ attack is a (weakly) dominant strategy, in practice, there are non-negligible cognitive costs associated with accepting the bribe. One must first understand the game theory of the attack by reading an article such as [Buterin 2015]. Moreover, while a given attack can be smart contract enforced, a user would not be able to immediately detect whether the smart contract guaranteeing the bribe is bug-free or otherwise has some trapdoor to benefit the attacker. Performing an audit of the contract requires a significant effort. Moreover, a user may simply be reluctant to accept a bribe for moral reasons. As a simple model for this phenomenon, one can attach a value to this moral cost of accepting a bribe, see [Burguet et al. 2016] for a summary of theoretical and experimental results along these lines.

Suppose that we combine these two types of costs together; then we assume a penalty of m if a participant votes for Y either due to moral or cognitive costs. This gives the payoff matrix of:

	X wins	Y wins
$\mathcal{USR}$ votes X	p	$-d$
$\mathcal{USR}$ votes Y	$p+\varepsilon-m$	$p-m$

If $\varepsilon-m$ and $p+d-m$ have the same sign there is again a dominant strategy (to either always or never take the bribe as the case may be). Suppose that $\varepsilon-m$ and $p+d-m$ have different signs, so there is no dominant strategy. Then there exists a mixed strategy equilibrium when voting for X gives the same expected value as voting for Y. We see that:

PROPOSITION 5.1. *Suppose the cognitive/moral cost m is constant for all voters, $\varepsilon < m < p+d$. Let $\delta > 0$. Then there exists an equilibrium and some M_δ such that if the number of voters M is greater than M_δ,*

$$\frac{p+d-m-(p+d)\delta}{p+d-\varepsilon} < Prob(X\ wins) < \frac{p+d-m}{p+d-\varepsilon}+\delta$$

in that equilibrium.

PROOF.

We see that there exists some $\pi_X \in (0,1)$ such that each participant adopting a mixed strategy, voting X with probability π_X independent of each other, gives rise to an equilibrium. To see this, we define

$$f(\pi_X) = E[\text{vote } X] - E[\text{vote } Y]$$

$$= (p+d)\sum_{\lfloor M/2\rfloor}^{M-1}\binom{M-1}{i}\pi_X^i(1-\pi_X)^{M-1-i} - d - \varepsilon\sum_{\lfloor M/2\rfloor+1}^{M-1}\binom{M-1}{i}\pi_X^i(1-\pi_X)^{M-1-i} - (p-m).$$

Then as f is continuous in π_X, $f(0) = -p-d+m < 0$, and $f(1) = m-\varepsilon > 0$, there exists the claimed value π_X such that $f(\pi_X) = 0$, which implies that these strategies yield an equilibrium, by the Intermediate Value Theorem.

Then, if the number of voters M is sufficiently large, considering the (binomial) density function of how many votes X receives, one has

$$0 \le \text{Prob}(X \text{ wins}|\,\mathcal{USR} \text{ votes } X) - \text{Prob}(X \text{ wins}|\,\mathcal{USR} \text{ votes } Y) < \delta.$$

Denoting $x = \text{Prob}(X \text{ wins}|\,\mathcal{USR} \text{ votes } Y)$, and using $f(\pi_X) = 0$, we have

$$(p+d)\text{Prob}(X \text{ wins}|\,\mathcal{USR} \text{ votes } X) - d = \varepsilon x + p - m$$

$$\Rightarrow (p+d)x - d \le \varepsilon x + p - m, \text{ and}$$

$$\varepsilon x + (p-m) < (p+d)(x+\delta) - d.$$

Then

$$\Rightarrow \frac{p+d-m-(p+d)\delta}{p+d-\varepsilon} < x = \text{Prob}(X \text{ wins}|\,\mathcal{USR} \text{ votes } Y) \le \frac{p+d-m}{p+d-\varepsilon}, \text{ and}$$

$$\frac{p+d-m-(p+d)\delta}{p+d-\varepsilon} < \text{Prob}(X \text{ wins}|\,\mathcal{USR} \text{ votes } X) < \frac{p+d-m}{p+d-\varepsilon}+\delta.$$

The result follows by the Law of Total Probability. □

Hence we see that the amount of the ε, namely the bribe that a corrupt voter can receive beyond the normal payout, needs to be large relative to cognitive and moral costs in order for the $p+\varepsilon$ attack to be effective. In this way, this attack begins to resemble more traditional bribes.

6. COMMENTS ON COUNTER-COORDINATION

The idea of voters organizing themselves *against* the attacker, engineering a failure of the attack that would require the attacker to pay out the maximum amount of bribes increasing total voter payoffs, was already considered in [Buterin 2015]. In this section, we recall and formalize these ideas so as to be able to analyze the effectiveness of counter-coordination under the various structures for Schelling games in Section 7.

PROPOSITION 6.1.
The following equilibria exist in a simple Schelling game under a $p+\varepsilon$ attack:

— *all voters vote Y*
— $\lfloor \frac{N}{2} \rfloor$ *votes for Y and* $\lfloor \frac{N}{2} \rfloor + 1$ *votes for X.*

If a rational actor with v votes thought, based on their knowledge of how other participants were likely to vote, that they had the decisive vote, then they maximize their return by splitting up their v votes so that X wins by a single vote. Thus as many votes as possible receive the additional ε while remaining in the situation where the attacker loses and actually has to pay out the bribes. This idea, when considered from the perspective of a group of voters coordinating rather than a single participant, is the basis for the response of "counter-coordination". As pointed out in [Buterin 2015], a problem with counter-coordination in games of the form Definition 2.1 is that each counter-coordinator, if they believe that the counter-coordination will succeed and the attack will fail independently of whether they individually participate or not, is strictly better off taking the full bribe. Hence counter-coordination can only function to the degree that participants are either altruistic or believe their participation is likely to make a difference in the success of the endeavor. Thus, this situation exhibits characteristics of a tragedy of the commons [Hardin 1968].

Also, we see:

PROPOSITION 6.2. *Suppose that counter-coordinators participate in a contract that only activates if at least $\lfloor \frac{M}{2} \rfloor + F$ participants place a deposit of D. Then, unless a future appeal round reverses ultimately rules in favor of the attacker, counter-coordinating will yield at least as high a payoff as voting coherently as long as*

$$D \geq \frac{(p+d)\lfloor \frac{M}{2} \rfloor}{F}$$

in the case of a simple Schelling game and

$$D \geq \frac{\left(\frac{\lfloor \frac{M}{2} \rfloor d + Mp}{M - \lfloor \frac{M}{2} \rfloor} + d \right) \lfloor \frac{M}{2} \rfloor}{F} \sim \frac{(p+d)M}{F}$$

in the case of a redistributive Schelling game.

PROOF. In both cases, we merely calculate the total amount that is required to compensate each of the at most $\lfloor \frac{M}{2} \rfloor$ non-defecting counter-coordinators, and divide it by the at least F defectors. □

In particular, note that the counter-coordination contract can be written such that $F = O(M)$, so that the deposit D is asymptotically constant as M grows. Furthermore, note that if one is guaranteed to be in the last round (for example because there is no appeal system or because the appeal system is set so that the maximum number of appeals has been exhausted), counter-coordinating can be constructed so that counter-coordinators cannot be worse off than if they merely attempted to vote coherently without counter-coordination. Last round effects are a double edge sword for whether they facilitate or hinder $p+\varepsilon$ attacks and counter-coordination, as we will explore in Section 8.

If one were to apply the ideas of Section 5 to a context where participants have the option of counter-coordinating, a participant might view counter-coordinating as more moral than accepting the bribe but it would likely have an even higher cognitive cost.

7. COSTS OF $P+\varepsilon$ ATTACKS ON VARIOUS STRUCTURES

In this section we consider measures of the difficulty for an attacker to launch $p+\varepsilon$ attacks in the models of Definitions 2.1, 2.2, and 2.3 with and without appeal. Particularly, we calculate the required capital that must be locked up in each of these scenarios and, where applicable in situations where a counter-coordination is possible that rational, non-altruistic players will participate in with some probability in an equilibrium, the expected cost in paid bribes due to the cases where the counter-coordination succeeds.

PROPOSITION 7.1.

— *A $p+\varepsilon$ attack on a simple Schelling game of the form of Definition 2.1 with M voters and no appeal requires a budget of $\lfloor \frac{M}{2} \rfloor (p+d+\varepsilon)$. More generally, a $p+\varepsilon$ attack on a simple Schelling game of the form of Definition 2.1 with successive appeal rounds of $M_1, M_2, \ldots, M_k$ voters respectively requires a budget of $\left(\lfloor \frac{M_k}{2} \rfloor + \sum_{i=1}^{k-1} M_i \right) (p+d+\varepsilon)$.*

— *A $p+\varepsilon$ attack on a redistributive Schelling game of the form of Definition 2.2 with M voters and no appeal requires a budget of*

$$\left(\frac{\left(M - \lceil \frac{M}{2} \rceil - 1\right) d + Mp}{\lceil \frac{M}{2} \rceil + 1} + d + \varepsilon \right) \left(M - \lceil \frac{M}{2} \rceil \right) \sim (2d + 2p + \varepsilon) \frac{M}{2}$$

with the asymptotic as M increases.
More generally, a $p+\varepsilon$ attack on a redistributive Schelling game of the form of Definition 2.2 with successive appeal rounds of $M_1, M_2, \ldots, M_k$ voters respectively requires a budget of

$$\left(\frac{\left(M_k - \lceil \frac{M_k}{2} \rceil - 1\right) d + M_k p}{\lceil \frac{M_k}{2} \rceil + 1} + d + \varepsilon \right) \left(M_k - \lceil \frac{M_k}{2} \rceil \right) + \sum_{i=1}^{k-1} (M_i^2 (d+p) + M_i \varepsilon).$$

— *A $p+\varepsilon$ attack on a symbiotic Schelling game of the form of Definition 2.3 with M voters and no appeal requires a budget of $\lfloor \frac{M}{2} \rfloor \left(\frac{p}{2} + d + \varepsilon \right)$. More generally, a $p+\varepsilon$ attack on a symbiotic Schelling game of the form of Definition 2.3 with successive appeal rounds of $M_1, M_2, \ldots, M_k$ voters respectively requires a budget of $\lfloor \frac{M_k}{2} \rfloor \left(\frac{p}{2} + d + \varepsilon \right) + \sum_{i=1}^{k-1} M_i (p+d+\varepsilon)$.*

PROOF.

We consider the redistributive Schelling game case. Take x_i to be the number of participants who vote for X in the ith round. The attacker must pay out $\frac{(M_i - x_i - 1)d + M_i p}{x_i + 1} + d + \varepsilon$ to each of the $M_i - x_i$ participants who votes Y in the ith round. However, the attacker only has to make these payments if $x_k \geq \lceil \frac{M_k}{2} \rceil$, causing the attack to fail. However, one can take any $x_i \geq 0$ for $i < k$. Note that this upper bound is in fact realized for $x_i = 0$ for $i < k$ and $x_k = \lceil \frac{M_k}{2} \rceil$. The other cases are similar.

□

Hence even when $d = 0$ and with no appeal mechanism, moving from the simple Schelling game to the redistributive model roughly doubles the budget required by the attacker from $\lfloor \frac{M}{2} \rfloor (p+\varepsilon)$ to $M(p + \frac{\varepsilon}{2})$. The difference in budget required between redistributive Schelling game versus the other models is much more dramatic in the case of appeals. The simple Schelling game has a required budget that grows linearly in M_i for all i, whereas the redistributive Schelling game has a required budget that grows quadratically in M_i for $i < k$.

In the context of performing a $p+\varepsilon$ attack on a proof-of-work system, one can compare these deposit sizes to those of other smart-contract enforced attacks described in [Judmayer et al. 2019]. Note that in some such applications on proof-of-work systems, particularly those in which a payout is made to "uncle blocks" as in Ethereum, an attack could reduce their payouts by these amounts, hence reducing the costs of a failed attack and hence the deposits required in a manner similar to the uncle-mining based attacks presented in [McCorry et al. 2019].

Remark 7.2. Performing a 51% attack, either by percentage of staked tokens or of mining power, tends to require a budget of $O(M)$ in the examples we considered in Section 4. Hence by having quadratic budget requirements in the redistributive case one can have situations where the $p+\varepsilon$ attack requires a much larger budget than obtaining 51% of all tokens, even if only a relatively small percentage of token holders are participating in a given vote.

Remark 7.3. The large budget requirements of $p+\varepsilon$ attacks on redistributive Schelling games with an appeal system represent an implicit cost to the attacker from the opportunity costs of having their capital locked up during the vote.

Remark 7.4.
One might criticize the perspective of the attacker offering bribes to participants in the first round when the first round is ultimately overruled by the appeal round as being artificial. We will consider more sophisticated strategies that the attacker can employ in Section 8. Note, however, that the counter-coordinators have an advantage in that the attacker is required to "move first" by launching a contract that commits her to paying out bribes according to clear, programmable conditions. The counter-coordinators can launch their contract in response.

We further see that the large budgets required to bribe participants in a redistributive Schelling game with appeals can be used as a weapon against the attacker via a well-designed counter-coordination.

THEOREM 7.5. *Suppose that an attacker that has a budget distributed according to an exponential distribution as $B \sim Exp(\lambda)$ for some $\lambda > 0$ launches a $p+\varepsilon$ attack on a multi-round redistributive Schelling games such as in Definition 2.2. Suppose the number of voters per round $M_i \geq 8$ grows at fastest exponentially, i.e. $M_{i+1} \leq c_1 M_i$ for all i, and at slowest linearly, i.e. $M_i \geq c_2 i$ for all i and suppose that appeal fees to appeal from the $i-1$st round to the ith round are no more than $c_3 M_i p$. We suppose that $d > \max\{c_4 p, 1\}$ for some fixed constant c_4 that depends only on c_1, c_2, c_3, and λ. Furthermore we suppose that $\varepsilon < \min\left\{\frac{2d}{3c_1}, d\right\}$.*

Suppose that each vote (across all rounds) is controlled by a unique entity. Let $\varepsilon_1 > 0$ and $l \in \mathbb{N}$. We consider the situation where counter-coordinators in the ith round, each of whom place a deposit of

$$D \geq \left\{ \begin{array}{ll} c_3 M_{i+1} p & : \ i \leq l-2 \\ c_3 M_l p + \lceil \frac{3M_l}{4} \rceil \varepsilon + M_l \varepsilon_1 & : \ i = l-1 \\ 8(p+d) & : \ i = l \end{array} \right\},$$

launch a contract that encourages participants to vote Y until round $l-1$ and then attempts to engineer a narrow loss for the attacker in round l by ordering $\lfloor \frac{3M_l}{4} \rfloor$ voters to vote X, as in Proposition 6.2. We organize this contract so that counter-coordinator(s) pay any required appeal fees from their deposit(s) and counter-coordinators in the $l-1$st round pay required appeal fees and an additional $\lceil \frac{3M_l}{4} \rceil \varepsilon + M_l \varepsilon_1$. Then, there exist choices of ε_1 and l, and a mixed strategy equilibrium where

$$Prob(X \text{ wins}) \geq 1 - e^{-\lambda M_l^2 (d+p)} - \sum_{j=1}^{l-1} \frac{R_j}{((M_j - 1)(d+p) + \varepsilon)\left(1 - e^{-\lambda M_l^2 (d+p)}\right)} > 0,$$

where

$$R_j = \left\{ \begin{array}{ll} c_3 M_{j+1} p & : \ j \leq l-2 \\ c_3 M_l p + \lceil \frac{3M_l}{4} \rceil \varepsilon + M_l \varepsilon_1 & : \ j = l-1 \end{array} \right\}.$$

Moreover, if we are in a model where the counter-coordinator do not need to fund the appeal fees themselves, then there exists an equilibrium where

$$Prob(X\ wins) \geq 1 - e^{-\lambda M_l^2(d+p)} - \frac{\lceil \frac{3M_l}{4} \rceil \varepsilon + M_l \varepsilon_1}{((M_{l-1}-1)(d+p)+\varepsilon)\left(1-e^{-\lambda M_l^2(d+p)}\right)} > 0.$$

Remark 7.6.

While this counter-coordination requires no altruism on the part of its participants, there are several practical limitations to this approach. The deposits required of the counter-coordinators in non-terminal rounds are quite large (though much smaller than the budget required by the attacker as in Proposition 7.1). So even though we will see in the proof that it suffices to have a single counter-coordinator in each of these rounds for the counter-coordination to succeed, a model where we imagine the whole pool of voters as willing to employ a mixed strategy where they pay that deposit with some probability is likely not realistic. Also, an important element of this scheme is that early round counter-coordinators are incentivized to participate because their potential reward is very large relative to the chance that their individual participation changes the result. This depends on the "lone voice of reason jackpot" that comes from all voters in these rounds voting Y. In practice, a substantial number of voters will vote X (and not counter-coordinate) either because they are altruistic or they are unaware of the attack. To the degree that the percentage of such voters is significant but less than half, this actually works against the counter-coordination. Also, similar to our comments in Section 6, while participating in such a counter-coordination scheme may be considered by participants as more moral than accepting a bribe, this would likely require high cognitive costs, even compared to the more straightforward counter-coordination considered previously. Finally, the assumption that ε is bounded in terms of d is not ideal as an attacker can choose a very large ε to avoid this defense, at the expense of even higher budget requirements. It is unclear if there is a version of counter-coordination for which this assumption can be removed.

PROOF OF THEOREM 7.5. To continue a $p+\varepsilon$ attack against a vote that is in the lth appeal round, the attacker needs to commit at least $M_l^2(d+p)$ resources to be locked up by Proposition 7.1 beyond the resources in appeal fees and bribes that are required in rounds up to that point. By the memory-less property of the exponential distribution, the probability that the attacker has the resources for the lth round, given that she has the resources for the first $l-1$ rounds, is at most

$$\text{Prob}(B \geq M_l^2 p) = e^{-\lambda M_l^2(d+p)},$$

using properties of the exponential distribution.

We exhibit voter strategies such that it is dominant to counter-coordinate in the lth round *assuming the vote reaches that round*, and in all previous rounds there will be equilibria where voters counter-coordinate with non-zero probability. Note that all participants who do not counter-coordinate will, in equilibrium, accept the bribe and vote Y as this choice weakly dominates voting X (without counter-coordinating).

In the lth round, the counter-coordination contract attempts to find enough committed participants, each of whom submits the deposit of D, so that even if F defect the lth round still votes for X, where F is chosen so that $\lfloor \frac{M_l}{2} \rfloor + F = \lfloor \frac{3M_l}{4} \rfloor$. By Proposition, 6.2 using our assumptions that $D \geq 8(d+p)$ and $M_l \geq 8$, no participant that has so committed will have an incentive to vote other than as instructed by the counter-coordination contract.

We construct our counter-coordination so that earlier rounds of counter-coordinators subsidies the last round (and any appeal fees as necessary). Take

$$\varepsilon_1 = \frac{(4d+4p)e^{-\lambda M_l^2(d+p)}}{1-e^{-\lambda M_l^2(d+p)}}.$$

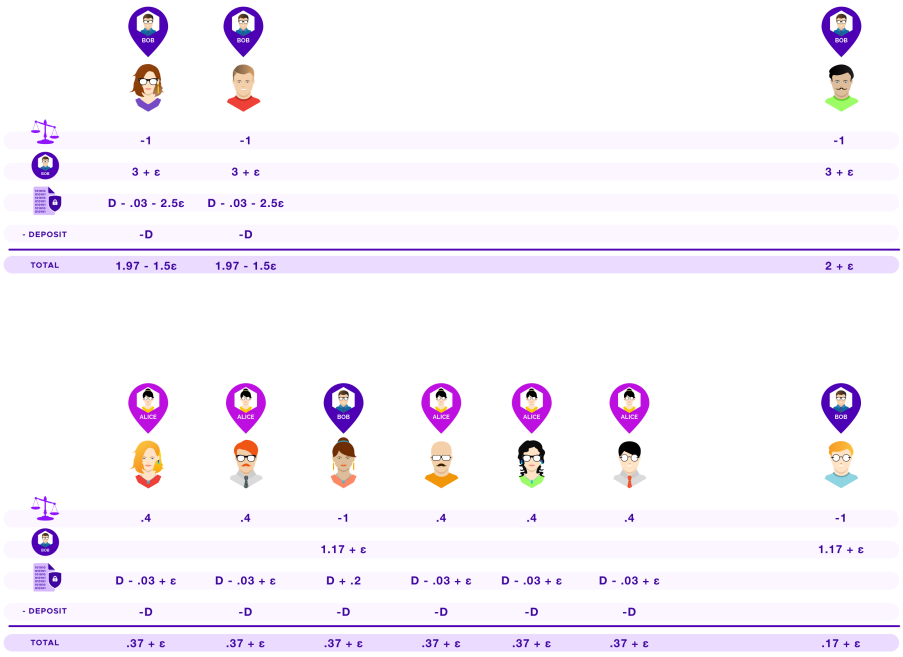

Fig. 2. Example of multi-round counter-coordination in a redistributive Schelling game where voters can choose options favoring Alice or Bob. Voters receive payouts from the coherence game, the attack contract, and the counter-coordination contract, which are symbolized by the scale, the image of Bob, and the file respectively. Participants are grouped horizontally by appeal round with counter-coordinators on the left and non-counter-coordinators on the right. Here $d = 1$, $p = 0$; hence if the attacker won in a later round the first round participants who unanimously voted for the attacker would receive 0. If the non-counter-coordinator in the last round had counter-coordinated, all last round participants including him would have received $.33 + \varepsilon$, an improvement over his current payoff. Note that for small M_l relative to ε it can be the case that counter-coordinating in the last round has a higher payoff than voting Y even before the subsidy from the previous rounds.

Then if we arrive to the lth round, the subsidy will pay for the at most $c_3 M_l p$ appeal fees, as well as provide an additional $\lceil \frac{3}{4} M_l \rceil \varepsilon + M_l \varepsilon_1$ to be split between round l, (non-defecting) counter-coordinators. Then if $K \geq \lfloor \frac{3M_l}{4} \rfloor$ counter-coordinators participate, do not defect, and the attacker does not have sufficient resources to appeal to the $l+1$st round, the $K - \lfloor \frac{3M_l}{4} \rfloor$ who vote Y receive ε more from the attack contract than the others. Then using the $\lceil \frac{3}{4} M_l \rceil \varepsilon + M_l \varepsilon_1$, each counter-coordinator receives the coherence payment for correctly voting X plus an additional $\varepsilon + \varepsilon_1$.

Thus, counter-coordinating yields a greater return than not counter-coordinating in the lth round by an additional ε_1 as long as the attacker does not have the resources to appeal to the $l+1$st round. However, a counter-coordinator can lose at most d if the attacker manages to win in a subsequent appeal, and a non-countercoordinating participant that votes Y would receive at most

$$\frac{\lfloor \frac{3M_l}{4} \rfloor d + M_l p}{M_l - \lceil \frac{3M_l}{4} \rceil} \leq 3d + 4p$$

as the number of counter-coordinating participants voting X is at most $\lfloor \frac{3M_l}{4} \rfloor$ in equilibrium as voting X outside of the context of counter-coordination is a dominated strategy.

However,

$$\varepsilon_1 = \frac{(4d+4p)e^{-\lambda M_l^2 (d+p)}}{1 - e^{-\lambda M_l^2 (d+p)}}$$

$$\Rightarrow 1-e^{-\lambda M_l^2(d+p)} \geq \frac{4d+4p}{4d+4p+\varepsilon_1}$$

$$\Rightarrow \text{Prob}\left(\begin{array}{c}\text{attacker cannot appeal}\\ \text{to } l+1\text{st round}\end{array}\right) \geq \frac{4d+4p}{4d+4p+\varepsilon_1}$$

$$\Rightarrow \varepsilon_1 \text{Prob}\left(\begin{array}{c}\text{attacker cannot appeal}\\ \text{to } l+1\text{st round}\end{array}\right) \geq \text{Prob}\left(\begin{array}{c}\text{attacker can appeal}\\ \text{to } l+1\text{st round}\end{array}\right)(4d+4p)$$

Hence we see that we have an equilibrium where honestly counter-coordinating has a higher expected return than voting Y outside of counter-coordination. So, assuming that a vote reaches the lth round, all participants in that round are incentivized to honestly counter-coordinate.

Now we consider earlier rounds of the counter-coordination. We assume the deposit of any counter-coordinator includes a subsidy of

$$R_i = \left\{\begin{array}{ll} c_3 M_{i+1} p & : \ i \leq l-2 \\ c_3 M_l p + \lceil \frac{3M_l}{4} \rceil \varepsilon + M_l \varepsilon_1 & : \ i = l-1 \end{array}\right\}$$

which covers any required appeal fees and the amount that is transfered to the counter-coordinators in the last round. Here all counter-coordinators will be instructed to vote Y; due to the $p+\varepsilon$ attack voting X either independently or in defecting from an instructed counter-coordination vote is a dominated strategy. This gives the payoff table:

	X wins	Y wins
Counter-coordinate	$\geq (M_i-1)d + M_i p + \varepsilon - R_i$	$\geq p - R_i$
Don't Counter-coord.	$(M_i-1)d + M_i p + \varepsilon$	p

Note that the inequalities in the payoff table are due to the possibility that if there are multiple counter coordinators in a given round, the R_i can be split between them. For simplicity we can also just think of any additional R_i contributions from having multiple counter-coordinators in a given round as being burned; we will see that even in this less favorable setting counter-coordination can still be incentivized.

Note that the chance that a single user's participation in the counter-coordination or not can make a difference in whether X or Y wins, when combined with the elevated rewards in the event that X wins, leads to an equilibrium where the counter-coordination can succeed, similar to the phenomenon discussed in Proposition 6.1.

Suppose all voters in the ith round adopt a mixed strategy where they counter-coordinate with probability z_i. Then $K_i \sim \text{Binom}(M_i, z_i)$ is the number of ith round voters who choose to counter-coordinate and $K_i' \sim \text{Binom}(M_i - 1, z_i)$ is the number of counter-coordinators other than a given selected voter $\mathcal{USR}$. We prove by induction that there exist $z_i, \ldots, z_{l-1} \in (0,1]$ such that there is an equilibrium where all voters in the jth round counter-coordinate with probability z_j, which yields

$$\text{Prob}\left(\begin{array}{c|c} X & \text{in } i\text{th} \\ \text{wins} & \text{round}\end{array}\right) \geq 1 - e^{-\lambda M_l^2(d+p)} - \sum_{j=i}^{l-1} \frac{R_j}{((M_j-1)(d+p)+\varepsilon)\left(1-e^{-\lambda M_l^2(d+p)}\right)} > 0 \quad (1)$$

and for $i > 1$

$$\text{Prob}\left(\begin{array}{c|c} X & \text{in } i\text{th} \\ \text{wins} & \text{round}\end{array}\right) \geq \frac{R_{i-1}}{(M_{i-1}-1)(d+p)+\varepsilon}. \quad (2)$$

Assume that we have some $1 \leq i < l-2$ such that the induction hypothesis holds for $i+1$. For $\mathcal{USR}$ in the ith round, define

$$f(z_i) = E\left[\begin{array}{c}\text{Counter-coord.}\\ \text{in } i\text{th round}\end{array}\right] - E\left[\begin{array}{c}\text{Do not counter-}\\ \text{coord. in } i\text{th round}\end{array}\right].$$

We denote

$$P(X|C) = \text{Prob}\left(\begin{array}{c} X \text{ wins if } \mathcal{USR} \\ \text{counter-coordinates} \end{array}\right)$$

$$P(X|N) = \text{Prob}\left(\begin{array}{c} X \text{ wins if } \mathcal{USR} \text{ doesn't} \\ \text{counter-coordinate} \end{array}\right)$$

etc.

This allows us to calculate

$$E\left[\begin{array}{c} \text{Do not counter-} \\ \text{coord. in } i\text{th round} \end{array}\right] = P(X|N)((M_i-1)d + M_i p + \varepsilon) + P(Y|N)p$$

and

$$E\left[\begin{array}{c} \text{Counter-coord.} \\ \text{in } i\text{th round} \end{array}\right] = -R_i + P(X|C)((M_i-1)d + M_i p + \varepsilon) + P(Y|C)p.$$

The system is structured so that a single counter-coordinator in the ith round is sufficient for the game to progress to the $i+1$st round. Then

$$P(X|C) = \text{Prob}\left(\begin{array}{c|c} X & \text{in } i+1\text{st} \\ \text{wins} & \text{round} \end{array}\right).$$

Moreover,

$$P(X|N) = \text{Prob}\left(\begin{array}{c|c} \text{progress to} & \mathcal{USR} \text{ does not} \\ i+1\text{st round} & \text{counter-coordinate} \end{array}\right) \cdot \text{Prob}\left(\begin{array}{c|c} X & \text{in } i+1\text{st} \\ \text{wins} & \text{round} \end{array}\right)$$

$$= \text{Prob}(K_i' \geq 1) P(X|C)$$

Note that $P(X|C)$ is independent of z_i (though it depends on the z_j for $j > i$), and $\text{Prob}(K_i' \geq 1)$ is continuous in z_i by properties of binomial distributions. So $f(z_i)$ is continuous. Furthermore, $f(0) = P(X|C)\left[(M_i-1)(d+p+\varepsilon)\right] - R_i > 0$, while $f(1) = -R_i < 0$. So the required $z_i \in (0,1)$ that yields an equilibrium exists.

Moreover,

$$P(X|C) \geq \prod_{j=i+1}^{l-1} \text{Prob}(K_j \geq 1) \cdot \left(1 - e^{-\lambda M_l^2 p}\right).$$

So

$$E\left[\begin{array}{c} \text{Counter-coordinate} \\ \text{in } i\text{th round} \end{array}\right] = E\left[\begin{array}{c} \text{Do not counter-} \\ \text{coordinate in } i\text{th round} \end{array}\right]$$

$$\Leftrightarrow R_i = \left((M_i-1)(d+p)+\varepsilon\right) P(X|C)\left(1 - \frac{P(X|N)}{P(X|C)}\right)$$

$$\Leftrightarrow \text{Prob}(K_i' \geq 1) = 1 - \frac{R_i}{\left((M_i-1)(d+p)+\varepsilon\right) P(X|C)}.$$

Then, as $\mathcal{USR}$ counter-coordinates with the same probability as the other voters in her round,

$$\text{Prob}(K_i \geq 1) \geq \left(1 - \frac{R_i}{\left((M_i-1)(d+p)+\varepsilon\right) P(X|C)}\right)^{M_i/(M_i+1)}.$$

$$\geq 1 - \frac{R_i}{((M_i-1)(d+p)+\varepsilon)\prod_{j=i+1}^{l-1}\text{Prob}(K_j \geq 1)\cdot\left(1-e^{-\lambda M_l^2(d+p)}\right)}.$$

So

$$\prod_{j=i}^{l-1}\text{Prod}(K_i \geq 1) \geq \prod_{j=i+1}^{l-1}\text{Prod}(K_j \geq 1) - \frac{R_i}{((M_i-1)(d+p)+\varepsilon)\left(1-e^{-\lambda M_l^2(d+p)}\right)}$$

$$\geq 1 - \sum_{j=i}^{l-1}\frac{R_j}{((M_j-1)(d+p)+\varepsilon)\left(1-e^{-\lambda M_l^2(d+p)}\right)},$$

where the last step uses another induction argument.

As we saw above that counter-coordination succeeds if it makes it to the lth round unless the attacker's resources allow her to appeal to an $l+1$st round, we have

$$\text{Prob}\left(\begin{array}{c|c} X & \text{in } i\text{th} \\ \text{wins} & \text{round}\end{array}\right) \geq 1 - e^{-\lambda M_l^2(d+p)} - \sum_{j=i}^{l-1}\frac{R_j}{((M_j-1)(d+p)+\varepsilon)\left(1-e^{-\lambda M_l^2(d+p)}\right)}.$$

$$\geq 1 - e^{-\lambda M_l^2(d+p)} - \frac{\lceil\frac{3M_l}{4}\varepsilon\rceil}{((M_{l-1}-1)(d+p)+\varepsilon)\left(1-e^{-\lambda M_l^2(d+p)}\right)} - O\left(\frac{lp+\varepsilon_1}{(d+p)\left(1-e^{-\lambda M_l^2(d+p)}\right)}\right),$$

where the implicit constant in the O only depends on the c_k's. Then, as

$$\frac{\varepsilon_1}{(d+p)\left(1-e^{-\lambda M_l^2(d+p)}\right)} = O\left(\frac{e^{-\lambda M_l^2(d+p)}}{\left(1-e^{-\lambda M_l^2(d+p)}\right)^2}\right),$$

taking all of the terms of the lower bound except for $O\left(\frac{lp}{(d+p)\left(1-e^{-\lambda M_l^2(d+p)}\right)}\right)$, we have a quantity that, as $u = \lambda M_l^2(d+p) \geq \lambda M_l^2$ grows, approaches a limit lower bounded by $1 - \frac{3c_1\varepsilon}{4d} \geq 1/2$, using our assumptions on ε relative to d. Then we can choose l sufficiently large in terms of λ and the c_k's. Moreover, we see that if d is a sufficiently large multiple of p (for fixed values of l, λ, and the c_k's), the contribution of the $O\left(\frac{lp}{(d+p)\left(1-e^{-\lambda M_l^2(d+p)}\right)}\right)$ term can be made suffciently small so that the probability that X wins is lower bounded by a constant. In particular,

$$\frac{R_{i-1}}{(M_{i-1}-1)(d+p)+\varepsilon} \leq \frac{c_1c_3M_{i-1}p}{(M_{i-1}-1)(d+p)+\varepsilon} \leq \frac{2c_1c_3p}{d}$$

can be made smaller than this constant for $d > c_4p$.

Then we complete the induction argument by considering the $i = l-1$ case. The bound required by inequality (2) is already seen to hold as part of our argument above lower bounding $\text{Prob}\left(\begin{array}{c|c} X & \text{in } i\text{th} \\ \text{wins} & \text{round}\end{array}\right)$. Then, similar to the general case, one can use the Intermediate Value Theorem to find z_{l-1} and show the bound of inequality (1).

Note that in the case where the counter-coordinators are not required to cover the appeal fees of a subsequent round, the only amount that is required for them to contribute is the direct subsidy to the

counter-coordinators in the lth round. Then we can think of the counter-coordination as beginning in the $l-1$st round, which gives the result in this case.

□

Hence the structure of this game has in some way allowed individual action to be "leveraged". Compared to the tragedy of the commons type situation we observed in Section 6, here voters are incentivized to participate in early rounds because their participation is sufficiently likely to change the result.

8. ADAPTATIONS OF $P+\varepsilon$ ATTACKS

In Section 7 we saw that the budget required to perform a $p+\varepsilon$ attack on a redistributive Schelling game across multiple appeal rounds grows quadratically in the number of voters and that counter-coordination becomes more viable in this setting. Ultimately, both of these effects come from the "lone voice of reason jackpot" phenomenon in these games as discussed in Section 2, where participants receive very large payouts if they are one of very few voters who choose a given option that goes on to win in appeal, so for the attacker to make voting for her a dominant strategy, she must offer correspondingly high bribe.

We have so far identified two natural approaches for the attacker to preserve the spirit of a $p+\varepsilon$ attack while removing or limiting this effect:

— An attacker can only pay bribes to participants in the last, decisive appeal round - without specifying what this round is in advance
— An attacker can offer a bribe in all rounds, but she can cap how much a given voter can receive as a bribe by B.

In this section we consider the implications of these two approaches compared to "pure" $p+\varepsilon$ attacks.

8.1. Last-round only $p+\varepsilon$ attacks

With a last round only bribe, participants are incentivezed to take the bribe if

— they believe they are in the last round, or
— if they think that the voters in the ultimate round will take the bribe and then they should vote coherently with them.

If the voter pool believes that the attack will likely get appealed beyond the point where the attacker can provide enough resources to maintain it, and hence ultimately lose, they are not incentivized to vote with the attacker. As a result, the chances of this type of attack are heavily dependent on whether voters expect that the attack is likely to succeed or not.

8.2. Capped $p+\varepsilon$ attacks

In the alternative, where the bribe payout is capped at $B < Md$, it is no longer a equilibrium for voters to adopt the pure strategy of accepting the bribe. Indeed, if a voter thinks that all of the other voters in a round will accept the bribe but that the decision will ultimately be overturned, then she can get a reward of $(M-1)d$ by voting coherently (without losing her deposit of d). In this section, we study an equilibrium that arises.

Denote $\pi_Y = \text{Prob}(\text{voter takes bribe})$. Furthermore, we denote by B the cap of the amount the attacker pays. Then a participant that accepts the bribe in an attack that fails can net no more than $B-d$ after losing her coherence payout. This gives rise to the following payoff table:

	X wins	Y wins
$\mathcal{USR}$ votes X	$\frac{(M-x-1)d+Mp}{x+1}$	$-d$
$\mathcal{USR}$ votes Y	$\min\left\{\frac{(M-x-1)d+Mp}{x+1}+\varepsilon, B-d\right\}$	$\frac{xd+Mp}{M-x}$

We see:

PROPOSITION 8.1. *Suppose*

— $d \leq B < Md + Mp$,
— all voters have exactly one vote
— the voters all expect the attack to eventually fail, possibly after future appeals.

Let $\delta > 0$. Then there exists M_0 such that, in equilibrium, if $M \geq M_0$,

$$\pi_Y \leq 1 - \frac{d+p}{B} + \delta.$$

PROOF.

Consider voters adopting a mixed strategy where they vote for X with probability π_X. Similar to the proof of Proposition 5.1, one can define $f(\pi_X) = E[\text{vote } X] - E[\text{vote } Y]$, which is continuous with,

$$f(0) = (M-1)d + Mp - \min\{(M-1)d + Mp + \varepsilon, B - d\},$$

$$f(1) = p - \min\{p + \varepsilon, B - d\}.$$

Note we have used here the assumption that all voters expect X to eventually win, hence one need not consider the column in the payoff table corresponding to Y winning. By our assumptions on d, B, M, and p, we have $f(0) > 0$ and $f(1) < 1$, so by the Intermediate Value Theorem there exists some $\pi_X \in (0,1)$ such that $f(\pi_X) = 0$, yielding an equilibrium.

Using the assumption that the final ruling is X, we compute

$$E[\text{vote } X] = \sum_{k=0}^{M-1} \frac{(M-k-1)d + Mp}{k+1} \binom{M-1}{k} \pi_X^k (1-\pi_X)^{M-k-1}$$

Note that, for $k \leq M-2$,

$$\frac{(M-k-1)}{k+1} \binom{M-1}{k} = \binom{M-1}{k+1}.$$

Also the $k = M-1$ term of the sum in $E[\text{vote } X]$ is 0. So taking $j = k+1$,

$$E[\text{vote } X] = p + \frac{1-\pi_X}{\pi_X} \left(\sum_{j=1}^{M-1} (d+p) \binom{M-1}{j} \pi_X^j (1-\pi_X)^{M-j-1} \right)$$

$$= p + \frac{1-\pi_X}{\pi_X} (d+p) \left(1 - (1-\pi_X)^{M-1}\right).$$

Then $E[\text{vote}X] = E[\text{vote } Y] \leq B - d$ implies

$$\frac{1-\pi_X}{\pi_X} (d+p) \left(1 - (1-\pi_X)^{M-1}\right) \leq B - d - p.$$

As $\pi_X < 1$, we know that if M is sufficiently large, $(1-\pi_X)^{M-1} < \delta_0$. Hence,

$$\frac{1-\pi_X}{\pi_X} (d+p)(1-\delta_0) \leq B - d - p$$

$$\Rightarrow \frac{(d+p)(1-\delta_0)}{B - (d+p)\delta_0} \leq \pi_X.$$

By continuity, for sufficiently small δ_0, this implies

$$\frac{d+p}{B} \leq \pi_X + \delta.$$

As $\pi_Y = 1 - \pi_X$, this gives the desired result.

□

In order for the attacker to have a reasonable chance of success as M increases, she should choose B so that $\pi_Y > 1/2$. Then we see by Proposition that she should, heuristically, choose $B > 2(d+p)$. Note that, if this attack is performed in successive appeal rounds with $M_1, \ldots, M_k$ voters respectively where $B > 2(d+p)$ is constant across all rounds, this requires a budget of at least

$$\max\left\{\left(\frac{(M_k - \lceil \frac{M_k}{2} \rceil - 1)d + M_k p}{\lceil \frac{M_k}{2} \rceil + 1} + d + e\right), B\right\}\left(M_k - \lceil \frac{M_k}{2} \rceil\right) + \sum_{i=1}^{k-1} 2M_i(d+p).$$

Compared to Proposition 7.1, this budget requirement is comparable to twice that of a series of $k-1$ redistributive Schelling games with no appeal and four times that of a series of $k-1$ simple Schelling games with no appeal. Considering our arguments in Remark 7.2 about how this attack compares to a 51% attack, these differences in budget requirements can make a significant difference in the viability of an attack. Moreover, we expect that both of these adaptations will require a more complicated thought process from voters, adding to the cognitive costs that we already saw were a defense in Section 5.

9. EXPERIMENTAL/STRESS TEST RESULTS

Finally, we mention a number of test $p+\varepsilon$ attacks that we launched on the Kleros "Doges on Trial" pilot in August and September of 2018. This platform maintained a curated list of "Doge" images (meme images of dogs, particularly Shiba Inus) and images showing cats were considered hostile in the context of the list. The $p+\varepsilon$ attacks were attached to images of cats, hence providing voters with an unambiguous sense of what an "honest" vote was in this context. These attacks were enforced with a smart contract that was made available to participants in advance.

All of these trials were redistributive Schelling games as in Definition 2.2 with the same p and d as these choices were fixed for the Kleros "Doges on Trial" pilot. These values were roughly $p = 4$ USD and $d = 3.5$ USD at the time. Particularly, this represented a "low-stakes situation". However the ε varied and both attacks where the bribe was guaranteed in non-terminal rounds as well as attacks where the bribe was only guaranteed in the terminal round were launched. (In both cases, it was not known to participants in advance how many rounds there would be.) Note that in this platform participants are pseudo-anonymous and if they have large percentages of the total token pool, they may be drawn multiple times for a given vote; in Tables 3 and 4 resulting number of votes is given with this multiplicity. There were 17 total Ethereum addresses (and hence likely this many distinct individuals) that participated as voters at least once; however, as these 17 addresses controlled a large percentage of the token pool, we will argue below that our observations from this test nonetheless provide insight about the state of the platform at the time of the tests.

When considering the danger of these $p+\varepsilon$ attacks, we are mostly concerned with whether 50% of the token holder pool from which the voters are drawn corresponds to people that would accept the bribe. In the case of Kleros, the voter selection mechanism makes independent, random choices of tokens to choose the voters. So it makes sense to imagine that each given token would vote one way or the other if presented with a given $p+\varepsilon$ attack, namely that there is some percentage y of all tokens that would accept the bribe. Hence, the number of votes in a given round of that attack should be distributed as Binomial(number of votes, y). Then we can ask if the number of votes seen accepting the bribe can realistically occur by random chance under the null hypothesis that exactly $y = .5$. Taking the last round of the "last round only attack" alone we see if $X \sim \text{Binomial}(122, .5)$, then $\text{Prob}(X <= 24) < 0.000001$. Similarly, if we consider all of the guaranteed p+epsilon bribes together, we see if $X \sim \text{Binomial}(163, .5)$, then $\text{Prob}(X <= 33) < 0.000001$.

ε (approx. USD)	votes accepting bribe	votes rejecting bribe
.02	0	9
.17	1	6
.87	1	4
1.7	2	4
5.2	3	3
17	1	7

Fig. 3. Vote totals accepting or rejecting $p+\varepsilon$ bribes at various values of ε when the bribe was guaranteed in non-terminal rounds.

ε (approx. USD)	votes accepting bribe	votes rejecting bribe
5.2 (before last round)	30	69
5.2 (last round)	24	98

Fig. 4. Vote totals accepting or rejecting $p+\varepsilon$ bribes when the bribe was only guaranteed in the last round, separated into totals before the last rounds and in the last round. Here the number of rounds was not announced in advance, but decision was appealed all the way to the last round possible under the parameters of the smart contract in the underlying platform. Hence, in the last round the bribe was known to be guaranteed to participants.

So far, we have found no observable relationship between the percentage of votes accepting the bribe and either value of ε or whether the bribe was guaranteed versus not guaranteed in the last-round version of the attack. (In fact, a higher percentage of participants accepted the last-round only bribe in the non-terminal rounds than in the last round.) Indeed, among the 10 Ethereum addresses that were selected for at least two different votes, 8 either always took the bribe or never took the bribe.

These results allow us to conclude with great confidence that the majority of token holders (weighted by the number of tokens they held) at the time of these attacks were not willing to accept bribes for these values of ε (and p and d). However, these results cannot be extrapolated to different values of parameters (indeed, several participants when solicited for feedback on their decisions after the end of the attack reported that the size of the bribe was not sufficient for them to accept it, lending greater credence to the idea of moral costs as a barrier to $p+\varepsilon$ attacks as in Section 5). Moreover, as these participants were chosen randomly from among a pool of token holders and not from a broader population, we can only really infer conclusions about the population of Kleros token holders. This is an appropriate context for a stress test on the resistance of a given plateform to $p+\varepsilon$ attacks; however, as the pool of token holders evolves, it might exhibit different behavior, particularly as it may adapt to the experiences it has witnessed with previous $p+\varepsilon$ attacks. As a result, we belief that further experiments regarding peoples' behavior when faced with a $p+\varepsilon$ attack in varying conditions should be conducted. However, the model provided by this kind of test can serve as a stress test for cryptoeconomic platforms that are based on Schelling games and which are potentially vulnerable to $p+\varepsilon$ attacks by periodically seeing if the population of token holders is willing to accept these bribes in conditions given by the typical parameters on that platform.

10. CONCLUSION

We observed that there exist viable counter-coordination that are stable in equilibrium that can defeat redistributive $p+\varepsilon$ attacks on systems where appeal is possible. This suggests the dangers for an attacker in the cat and mouse game that these attacks can become. Moreover, even in contexts where counter-coordination in redistributive Schelling games is not practical, we have seen that the budget required for $p+\varepsilon$ attacks on these systems is likely prohibitive. We have considered ways in

which an attacker can adapt. However, the resulting attacks present a danger that is more nuanced than was the case for pure $p+\varepsilon$ attacks and they likely require higher cognitive costs to understand, whereas cognitive costs are already an important defense against this class of attacks. Finally, we considered the results of some experiments/stress tests in which we launched $p+\varepsilon$ attacks on a real platform, and while these tests are preliminary and limited to the specific context of that platform, we observed that these attacks did not succeed.

11. ACKNOWLEDGEMENTS

We wish to acknowledge the contribution of Daniel Babbev towards writing the attack smart contract used in the tests discussed in Section 9. For the code of this contract see [Babbev 2018].

References

Michael Abramowicz. 2019. The Very Brief History of Decentralized Blockchain Governance. *Vanderbilt Journal of Entertainment and Technology Law (symposium contribution). Forthcoming; GWU Law School Public Law Research Paper No. 2019-14* (February 2019).

Daniel Babbev. 2018. p-epsilon.sol. https://github.com/kleros/kleros-attacks/blob/master/contracts/p-epsilon.sol, *Github repository* (August 2018).

Warren Buffett. 2000. The Billionaire's Buyout Plan. Online, https://www.nytimes.com/2000/09/10/opinion/the-billionaire-s-buyout-plan.html, *New York Times* (September 2000).

Roberto Burguet, Juan-José Ganuza, and José Garcia Montalvo. 2016. *The microeconomics of corruption: A review of thirty years of research*. Economics working papers, Barcelona GSE working paper series working paper no. 908. https://EconPapers.repec.org/RePEc:upf:upfgen:1525

Vitalik Buterin. 2015. The P + epsilon Attack. Online, https://blog.ethereum.org/2015/01/28/p-epsilon-attack/, *Ethereum blog* (January 2015).

Vitalik Buterin. 2018. Discouragement Attacks. Online, https://github.com/ethereum/research/blob/master/papers/discouragement/discouragement.pdf, *Github repository* (December 2018).

Vitalik Buterin. 2019. Serenity Design Rationale. Online, https://notes.ethereum.org/@vbuterin/rkhCgQteN?type=view, (Consulted November 2019).

Civil. 2017. The Civil Cryptoeconomic Whitepaper. Online, https://medium.com/@Join_Civil/the-civil-cryptoeconomic-whitepaper-1a42a7ff038d, *Medium* (October 2017).

Russell Cooper. 1999. *Coordination Games*. Cambridge University Press. https://EconPapers.repec.org/RePEc:cup:cbooks:9780521570176

Evangelos Deirmentzoglou, Georgios Papakyriakopoulos, and Constantinos Patsakis. 2019. A Survey on Long-Range Attacks for Proof of Stake Protocols. *IEEE Access* PP (February 2019).

Mike Goldin. 2017. Token-Curated Registries 1.0. Online, https://medium.com/@ilovebagels/token-curated-registries-1-0-61a232f8dac7, *Medium* (September 2017).

Mike Goldin, Ameen Soleimani, and James Young. 2017. The AdChain Registry. Online, https://adtoken.com/uploads/white-paper.pdf, (May 2017).

G. Hardin. 1968. The Tragedy of the Commons. *Science* 162, 3859 (December 1968), 1243–1248.

Aljosha Judmayer, Nicholas Stifter, Alexei Zamyatin, Itay Tsabary, Ittay Eyal, Peter Gazi, Sarah Meiklejohn, and Edgar Weippl. 2019. Pay-To-Win: Incentive Attacks on Proof-of-Work Cryptocurrencies. Online, https://eprint.iacr.org/2019/775, *Cryptology ePrint Archive, Report 2019/775* (2019).

Clément Lesaege and Federico Ast. 2018. Kleros: Short Paper v1.0.5. https://kleros.io/assets/whitepaper.pdf, (January 2018).

Patrick McCorry, Alexander Hicks, and Sarah Meiklejohn. 2019. Smart Contracts for Bribing Miners. In *Financial Cryptography and Data Security*. Springer Berlin Heidelberg, Berlin, Heidelberg, 3–18.

Satoshi Nakamoto. 2009. Bitcoin: A Peer-to-Peer Electronic Cash System. (March 2009).

Sarwar Sayeed and Hector Marco Gisbert. 2018. On the effectiveness of blockchain against cryptocurrency attacks. In *The Twelfth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies)*. International Academy, Research, and Industry Association, United States, 9–14.

T.C. Schelling. 1980. *The Strategy of Conflict*. Harvard University Press. https://books.google.ca/books?id=7RkL4Z8Yg5AC

Paul Sztorc. 2015. Truthcoin: Peer-to-Peer Oracle System and Prediction Marketplace. https://www.truthcoin.info/papers/truthcoin-whitepaper.pdf, (December 2015).

Kyle Wang. 2019. Cryptoeconomics: Paving the Future of Blockchain Technology. Online, https://hackernoon.com/cryptoeconomics-paving-the-future-of-blockchain-technology-13b04dab971, *Hackernoon* (October 2019).