Cryptoeconomic Systems • Volume 1, Issue 2

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

William George¹, Clément Lesaege¹

¹Kleros Cooperative

Published on: Oct 22, 2021

DOI: 10.21428/58320208.28ef9473

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

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 via smart contracts, requiring no trust in the attacker. In this work, we will analyze the “ $p + \epsilon$ attack” on Schelling point-based cryptoeconomic systems. Particularly, we will 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. 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 symmetric 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

Figure 1: An example payoff matrix for a voter $\mathcal{USR}$ in a symmetric coordination game for $p > 0$ constant.

However, as was noted by Thomas Schelling [1], 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. Then, a “Schelling game” is a coordination game where one option is such a focal point, that is distinguished for reasons external to the payoff structure of the game. This is in contrast to approaches based on participants coordinating based on asymmetries in the payoff structure of the game, such as when equilibria are Pareto-ranked; see *e.g.*, [2], [3], [4] for work in this direction.

Schelling games are 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 [5] it is pointed out that in consensus algorithms based on proof-of-work and the longest chain rule such as Bitcoin [6], the choice of which fork to follow can be thought of as a “vote” having the structure of a Schelling game. See Section 4 for further details on these examples.

An intriguing new attack, the $p + \epsilon$ attack, designed to warp the incentive structures of these games was proposed in a blog post by Vitalik Buterin [5], crediting the idea for the attack to Andrew Miller. In general, a $p + \epsilon$ 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 of the game of Figure 1, this yields a payoff matrix for the participant $\mathcal{USR}$ of:

	X wins	Y wins
$\mathcal{USR}$ votes X	p	0
$\mathcal{USR}$ votes Y	$p + \epsilon$	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.” Indeed, one observes that there are (at least) the following two equilibria in this game:

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

In [5], several defenses to attempt to resist such attacks are considered, including, notably, a defense that builds on the second type of equilibrium above, where one is at a “tipping point” between X and Y . Here, if a rational actor with v votes thinks, based

on her knowledge of how other participants were likely to vote, that she has the decisive vote, then she maximizes her 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. Then 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” considered in [5]. As pointed out in [5], a problem with counter-coordination in games of this form 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 [7].

More generally, 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, Buffet highlighted absurd aspects of the United States campaign finance system and arguing for campaign finance reform [8]. It has been remarked that situations similar to the $p + \epsilon$ have existed in other pre-blockchain situations, particularly surrounding share buying offers in acquisitions of companies; indeed, there is even an example of such a situation where the intended victim of the attack launched a second $p + \epsilon$ attack in support of its side [9]. Also, the $p + \epsilon$ attack has been considered in classifications of types of attacks on cryptoeconomic systems [10] particularly towards the perspective of whether these attacks are viable on proof-of-stake consensus systems as well as proof-of-work systems [11], [12]. Nonetheless, considering the widespread use of Schelling-based structures (explicitly or implicitly) in blockchain-based systems, this attack has so far received relatively little attention from the research community.

2. Possible structures of Schelling games

In this section, we will recall various elements that can be considered in the design of a Schelling game. In all of these situations we will take $p, \epsilon \in \mathbb{R}_{>0}$.

- Users can be required to place a deposit $d \geq 0$ which they lose if their vote disagrees with the ultimate result. Thus, this allows penalties to voters as well as rewards.

- We can also have appeals. Appeals result in delaying the result of a vote by an additional t blocks per appeal, up to some maximum number of appeal rounds $l_{**} \in \mathbb{N} \cup \{\infty\}$. Each appeal will typically consume some additional resources, with the idea that more resources can be put towards this ruling. So, ideally, this ruling is more likely to be correct. Concretely, if we suppose that there are M_i voters in the i th appeal round, a Schelling game of the form of [Figure 1](#) might require $O(M_i p)$ resources to provide for payouts to voters. Of particular note for our purposes, we will 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. Appeals can be triggered in a variety of ways, such as by allowing users to pay a fee, or by random chance.¹ Some of our results will assume the model where users can trigger an appeal in the i th appeal round by paying a fee of (at most) R_i . Nevertheless, a user that decides to trigger such an appeal may ultimately pay less than the R_i ; for example, if there are multiple users who decide to appeal, they might be able to split the fee between them, one of them might be randomly chosen to pay the full fee, or they might all have to pay the full fee with the excess amounts burned, depending on the appeal system used. While our results are general enough to handle these different possibilities, we nevertheless assume the following continuity property: if $USR_1, \dots, USR_N$ are users that appeal in round i with probabilities $y_{i,1}, \dots, y_{i,N} \in [0, 1]$ respectively, then

$$\begin{aligned} & f_k(y_{i,1}, \dots, y_{i,k-1}, y_{i,k+1}, \dots, y_{i,N}) \\ &= E[\text{amount of fees } USR_k \text{ pays to appeal}] \end{aligned} \tag{1}$$

is continuous in $y_{i,1}, \dots, y_{i,N}$. Furthermore, we assume

$$0 \leq f(y_{i,1}, \dots, y_{i,k-1}, y_{i,k+1}, y_{i,N}) \leq R_i.$$

In our work, we will only consider binary Schelling games: namely, games where a set of voters select between two different outcomes. Nonetheless, one can construct different payoff matrices for games that preserve 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 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 $p + \epsilon$ attack transforms the payoff matrix to:

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

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

Another alternative that is employed in some applications (see [Section 4](#)) is the following structure:

Definition 2.2: 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 $p + \epsilon$ 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	$\frac{p(x+1)}{M} + \epsilon$	$\frac{p(M-x)}{M}$

Finally, as another alternative, we could also divide all amounts—both the positive sum resources that the participants can receive *and* the lost deposits—between voters (in a given appeal round) on the winning side. This has the effect that rewards are larger in the event of narrow decisions.

Definition 2.3: 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 $p + \epsilon$ 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	$\frac{(M-x-1)d+Mp}{x+1} + \epsilon$	$\frac{xd+Mp}{M-x}$

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 voters to not take a chance on trying to get this “lone voice of reason jackpot,” an attacker needs to be willing to offer bribes that are $O(M)$ to each voter, for a total of $O(M^2)$, in the most extreme case.

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.

We will see in [Section 6](#) and [Section 7](#) that redistributive Schelling games with appeal systems have a certain level of robustness against $p + \epsilon$ attacks, essentially due to the nature of “lone voice of reason jackpot.” Hence, we also propose adapted attacks, which, while they can be applied to any of the payoff structures discussed above, are designed to preserve the spirit of a $p + \epsilon$ attack while removing or limiting the “lone voice of reason jackpot” in redistributive games:

- “Last round only attack”: an attacker can only pay bribes to participants in the last, decisive appeal round—without specifying what this round is in advance
- “Capped attack”: an attacker can offer a bribe in all rounds, but she caps how much a given voter can receive as a bribe by $H > 0$.

Deposits and (a form of) appeal systems,² were all already considered as possible defenses against $p + \epsilon$ attacks in [5] in the context of Schelling games of the form of [Definition 2.1](#). The possibility that participants can lose deposits has already been evoked as a basis for increased security against $p + \epsilon$ attacks in the context of proof-of-stake consensus algorithms [11] (compared to proof-of-work where miners do not typically stake a deposit³). However, the interplay between these elements, and particularly how they interact with the choice of games of the form of [Definition 2.1](#), [Definition 2.2](#), or [Definition 2.3](#), seems to have not been fully explored.

3. Our contribution

In this work we extend the analysis presented in [5] on simple Schelling games of the form of [Definition 2.1](#) to symbiotic Schelling games of the form of [Definition 2.2](#) and redistributive Schelling games of the form of [Definition 2.3](#). Moreover, we will see how various defenses against $p + \epsilon$ attacks, such as deposits and appeal systems, interact with each other and with the choice of payoff structure.

In [Section 6](#) and [Section 7](#), we see that, under a variety of measures of difficulty of an attack, redistributive games, particularly when there is an appeal structure, have increased resistance to $p + \epsilon$ attacks. Specifically, in [Section 6](#), we see that, under certain conditions, a $p + \epsilon$ attack on a redistributive Schelling game with an appeal system has budget requirements which grow quadratically in the number of participants and hence are likely to be prohibitive in many cases. In [Section 7](#), we see that there are “counterattacks” against $p + \epsilon$ attacks on redistributive Schelling games with appeal that are stable in a mixed strategy equilibrium and result in the attack failing with positive probability greater than a lower bound that we are able to calculate explicitly.⁴

However, in [Section 9](#), we will analyze the “last round only” and “capped” $p + \epsilon$ attacks, where the attacker adapts the spirit of a $p + \epsilon$ attack to avoid some of these defenses. While we will see that these adapted versions are less prohibitive, even in the context of a redistributive Schelling game with an appeal system, accepting the bribe is no longer a dominant strategy, and hence these attacks are no longer “free.” Thus, these attacks seem to be somewhat less dangerous. In [Section 8](#), we consider the situation where rational participants assign a moral (or cognitive effort) cost to accepting the bribe, and we observe mixed strategy equilibria where the attack fails some percentage of the time.

Finally, in [Section 10](#), we will consider the results of some preliminary experiments around $p + \epsilon$ attacks as 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 [\[13\]](#). 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 [\[14\]](#), [\[15\]](#) for examples.
- Truthcoin [\[16\]](#) 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 [\[17\]](#) 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 bringing 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. These “jurors” then vote on the result of the dispute, and they are rewarded or penalized based on whether or not 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 [\[5\]](#) is that of the proof-of-work consensus mechanism. A given block producer can expect to receive p in block rewards, but only if her blocks are ultimately in the longest chain. Then suppose that we have two rival chains on which one can mine, chains X and Y , where on each chain a transaction has been accepted that conflicts with the other chain. 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 $p + \epsilon$ bribe mines on the Y chain assuming the X chain ultimately becomes the longest chain, 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 z of all mining resources is mining on the X chain for some period of time while all other miners, 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 2.3](#).

- The Serenity proof-of-stake version of Ethereum [\[18\]](#) (currently in Phase 0 of its release) uses the symbiotic Schelling game model of [Definition 2.2](#). This choice was motivated to make the protocol less vulnerable to “discouragement attacks” [\[18\]](#), [\[19\]](#). Note that as this protocol, in contrast to proof-of-work, is designed to have finalization guarantees [\[20\]](#), one cannot consider waiting extra blocks as an “appeal” in the same way.

5. Model of actors

In this work, our attacker will generally have the ability to offer some type of $p + \epsilon$ bribe. In situations where there is an appeal structure, one typically thinks of this structure as being a defense; however, is a double-edged sword as we then may 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.

Moreover, in [Section 7](#) where we discuss “counterattacks” against $p + \epsilon$ attacks, we will need to assume a more rigid pattern for attacker behavior. Indeed, there we consider a $p + \epsilon$ attack in a multi-round situation, and we model the attacker as offering the bribe in each round until

- Her attack succeeds and Y wins the game; or
- The attacker stops, either because she has reached the maximum appeal round or because the required bribes exceed predetermined some budget,⁵

and not taking other actions.

For some results we will require specific behavioral assumptions of voters. A particular set of hypotheses that we will often make on voters is the following:

Definition 5.1: Voters are said to satisfy * if:

- Each vote (across all rounds) is controlled by a unique entity.⁶
- Voters are economically rational, namely, they take actions that maximize their expected return in a risk-neutral fashion (possibly assigning extra, specified internal costs to immoral or cognitively difficult actions, *see* [Section 8](#)).
- The structures of the games, including the presence of attacks, is common knowledge among the voters.⁷
- If the payoff matrix is symmetric between the “honest choice” X and the “dishonest choice” Y , for example for participants in an appeal round where the $p + \epsilon$ bribe is no longer provided, voters vote for the “honest choice” X with probability one.

6. Budget requirements of $p + \epsilon$ attacks on various structures

In this section we will begin to consider measures of the difficulty for an attacker to launch $p + \epsilon$ attacks in the models of [Definition 2.1](#), [Definition 2.2](#), and [Definition 2.3](#), with and without appeal. Particularly, we will calculate the required capital that must be locked up in each of these scenarios.

Proposition 6.1:

- A $p + \epsilon$ attack on a simple Schelling game of the form of [Definition 2.1](#) with M voters and no appeal requires a budget of $\lfloor \frac{M}{2} \rfloor (p + d + \epsilon)$. More generally, a $p + \epsilon$ attack on a simple Schelling game of the form of [Definition 2.1](#) with successive appeal rounds of $M_1, M_2, \dots, M_k$ voters respectively requires a budget of $\left(\lfloor \frac{M_k}{2} \rfloor + \sum_{i=1}^{k-1} M_i \right) (p + d + \epsilon)$.
- A $p + \epsilon$ attack on a symbiotic Schelling game of the form of [Definition 2.2](#) with M voters and no appeal requires a budget of $\lfloor \frac{M}{2} \rfloor \left(\frac{p}{2} + d + \epsilon \right)$. More generally, a $p + \epsilon$ attack on a symbiotic Schelling game of the form of [Definition 2.2](#) with successive appeal rounds of $M_1, M_2, \dots, M_k$ voters respectively requires a budget of $\lfloor \frac{M_k}{2} \rfloor \left(\frac{p}{2} + d + \epsilon \right) + \sum_{i=1}^{k-1} M_i (p + d + \epsilon)$.
- A $p + \epsilon$ attack on a redistributive Schelling game of the form of [Definition 2.3](#) with M voters and no appeal requires a budget of

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

with the asymptotic as M increases.

More generally, a $p + \epsilon$ attack on a redistributive Schelling game of the form of [Definition 2.3](#) with successive appeal rounds of $M_1, M_2, \dots, M_k$ voters respectively requires a budget of

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

PROOF: We consider the redistributive Schelling game case. Take x_i to be the number of participants who vote for X in the i th round. The attacker must pay out $\frac{(M_i - x_i - 1)d + M_i p}{x_i + 1} + d + \epsilon$ to each of the $M_i - x_i$ participants who vote Y in the i th 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 $\lfloor \frac{M}{2} \rfloor (p + \epsilon)$ to $M(p + \frac{\epsilon}{2})$. The difference in budget required between a

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 $i < k$.

Example 6.2: We consider the implications of [Proposition 6.1](#) on some of the examples we considered in [Section 4](#), particularly those related to consensus mechanisms. Namely, for each proof-of-work as it exists for both Bitcoin and Ethereum, the heuristic version of proof-of-work presented in [Section 4](#) where difficulty adjusts continuously, and the Serenity proof-of-stake system, we will consider the budget required to do each of two different $p + \epsilon$ attacks: a censoring attack and a double spend attack. See [Figure 2](#) for a summary of the results of these calculations.

- In a censorship attack on a proof-of-work system, the attacker incentivizes all miners to mine a fork where a given transaction $\mathcal{TX}$ is not included; denote this branch by Y . A priori, starting from any block in Y a miner could insert $\mathcal{TX}$ in a child block resulting in several different branches competing with Y to be accepted by the network. However, we consider the worst case for the attacker, where all mining resources that accept $\mathcal{TX}$ are concentrated in a single branch X . Then, if the attacker wants to censor $\mathcal{TX}$ in the branch accepted by the network at a fixed time T , this is equivalent to having Y win a simple Schelling game with no appeal and M equal to the number of blocks until T , allowing one to apply [Proposition 6.1](#) directly. As of April 2021, for Bitcoin these values are an average block time of roughly 9 minutes and a block reward of 6.25 BTC plus average fees of approximately .6 BTC, and for Ethereum one has an average block time of 13.3 seconds with an average block reward of roughly 3.1 ETH, including fees. [\[21\]](#)
- A double spend $p + \epsilon$ attack on a proof-of-work network has a similar structure. Here an attacker wants two conflicting transactions $\mathcal{TX}_1$ and $\mathcal{TX}_2$ to receive enough confirmations K to be accepted by two different vendors so that she receives products corresponding to each of these transactions, while only paying for one of them. We analyze the budget required for an attacker strategy where the attacker waits for one of the transactions, for example $\mathcal{TX}_1$, to receive K confirmations on branch X and then launches a $p + \epsilon$ attack to incentivize miners to revert to the last block where $\mathcal{TX}_1$ was not included and mine an alternative branch Y until it surpasses X to become the longest chain and also has (at least) K confirmations. Suppose the attacker offers the $p + \epsilon$ bribe to miners who mine blocks $K + 1, \dots, 2K + 1$ (the attacker could offer the bribe over a wider range, allowing for her to still win even if some minority of miners mine on X at the expense of requiring

a larger budget). Then, this corresponds to a simple Schelling game between X and Y with $M = 2K + 1$ votes where the first K votes are already set to X. Due to this restriction on the voting pattern, one cannot apply [Proposition 6.1](#) directly; nonetheless one sees by a similar argument that the payouts required by this attack in the most extreme case where exactly K of the remaining $K + 1$ participants mine on Y gives the same required total payouts as those in *Proposition 6.1*. For the values for Bitcoin that we present in [Figure 2](#), we take the standard choice of $K = 6$, [\[6\]](#) and for Ethereum we take $K = 20$, corresponding to the number of confirmations required by large exchanges to process deposits [\[22\]](#).

- We now consider similar attacks under the heuristic model of proof-of-work with rapidly adjusting difficulty discussed in [Section 4](#). We develop on this heuristic, noting additional approximations as necessary. Particularly, we suppose that the collective mining power of the community consists of M identical ASICs each of which can attempt t nonces per minute. This collective mining power remains fixed over the period of the attack. Furthermore, as the difficulty of different blocks varies over the course of a confirmation period, we assume that nodes verifying the confirmation status of a block check for a total amount of difficulty K in subsequent blocks rather than simply a given number of subsequent blocks. Suppose T is the time at which the attacker wants the transaction $\mathcal{TX}$ to be censored. Denote by T_0 the time at which $\mathcal{TX}$ becomes known, and denote by time T_1 a point at which the difficulty on the fork X where $\mathcal{TX}$ is included has adjusted to the point so that a given nonce has a $\frac{1}{t}$ chance of producing a block. (In particular, this supposes that enough miners mine on X during the transition between T_0 and T_1 to trigger the difficulty adjustment.) Then suppose that the attacker manages to incentivize all miners to mine on the fork Y where $\mathcal{TX}$ is not included until time T_2 when all miners switch to mining the fork X. For the attacker to incentivize miners to not defect, she needs to provide them with a reward of at least

$$E \left[\begin{array}{c} \text{Mine on X} \\ \text{at a Single} \\ \text{Time Increment} \end{array} \right]$$

at each of the t increments per minute at which they mine between T_1 and T_2 . Hence, to cover bribes in this scenario, the attacker needs

$$\begin{aligned}
& (T_2 - T_1) \cdot M \cdot t \cdot E \left[\begin{array}{c} \text{Mine on X} \\ \text{at a Single} \\ \text{Time Increment} \end{array} \right] \\
& \approx \left(\frac{1}{2}T - \frac{\text{Difficulty Transition}}{\text{Periods}} \right) \\
& \cdot M \cdot t \cdot \frac{\text{Avg. Block Reward}}{\tau},
\end{aligned}$$

where we note that, as all miners are mining on Y between T_1 and T_2 and are all mining on X after T_2 , discounting the assumed to be short periods of difficulty adjustment, these periods that produce roughly equal total difficulty are of roughly equal length.

- We take the assumptions related to difficulty adjustment from the heuristic above and have the same structure of double spending attacks as for standard proof-of-work. Similarly to above, we denote by T_0 the point at which $\mathcal{T}\mathcal{X}_1$ is mined in a block and by T_1 the point at which the chain X has accumulated enough difficulty to be viewed as confirmed and the attacker launches the attack. Denote by T_2 the time at which the difficulty on chain Y has adjusted to reflect the entire community mining on it and at which the difficulty on chain X has adjusted so that a given nonce has a $\frac{1}{\tau}$ chance of producing a block. We consider a scenario where all N miners mine on Y until some time T_3 when exactly $K - 1$ total difficulty is mined on chain Y, at which point all miners switch back to mining on X. For the attacker to incentivize miners to not defect, she now needs to be able to cover total bribes of

$$\begin{aligned}
& (T_3 - T_2) \cdot M \cdot t \cdot E \left[\begin{array}{c} \text{Mine on X} \\ \text{at a Single} \\ \text{Time Increment} \end{array} \right] . \\
& \approx \left(\frac{\text{Avg. Typical}}{\text{Confirmation Time}} - \frac{\text{Difficulty Transition}}{\text{Periods}} \right) \\
& \cdot M \cdot t \cdot \frac{\text{Avg. Block Reward}}{\tau},
\end{aligned}$$

where we note that, as all miners are mining on Y between T_2 and T_3 and discounting the assumed to be short periods of difficulty adjustment, $T_3 - T_2$ is roughly the time that would normally be required for a transaction to be confirmed.

- We now consider a similar censorship attack on the proof-of-stake system of Serenity [18], adapted to account for its somewhat different structure. Here blocks are grouped into “epochs” and the participants of the validator pool are tasked with providing an attestation to one block per epoch. Then the protocol attempts to determine a sequence of “checkpoint” blocks, consisting of a single checkpoint per epoch. Attestations contain references: to the previous block, a “target” checkpoint for the current epoch, and a “source” checkpoint from a previous epoch [23]. Then a checkpoint b is considered “justified” if at least two thirds of validators (weighted by deposit) provide an attestation with target b and source a , where a is a justified checkpoint for a previous epoch. An epoch is considered “finalized” if its checkpoint a is justified and at least two thirds of validators attest to a checkpoint in the immediately following epoch with a provided as the source [24]. In the event of a fork, (self-consistent) attestations to the different branches would disagree on the previous block, the previous block and the target checkpoint, or all three elements, depending on the depth of the divergence point of the forks; validators are rewarded or penalized based on how many of these elements they correctly specify, conforming to the state of the chain as it is eventually finalized in some future epoch.⁸

In the censorship attack that we consider, we suppose that the attacker’s goal is to prevent a given transaction $\mathcal{TX}$ from appearing in a block in a finalized epoch by a fixed time T . Suppose that the transaction $\mathcal{TX}$ to be censored⁹ appears in epoch i_0 . Then, considering that in order to be finalized, a checkpoint on some fork must be included as a target and then a source in two-thirds of attestations of two consecutive epochs, to prevent finalization of an epoch including $\mathcal{TX}$ it is sufficient for an attacker to offer $p + \epsilon$ bribes to validators who attest a fork Y that does not include $\mathcal{TX}$ in every other epoch: $i_0 + 1, i_0 + 3, \dots$. Hence we assume that the attacker only offers bribes for these epochs.¹⁰ Note that if a fork X that includes $\mathcal{TX}$ in epoch i_0 ultimately finalizes, then any attestation for Y will be incorrect on the previous block, the target, and the source (assuming that i_0 is justified) in any of the epochs considered. Serenity has a base reward B that is calculated in terms of the number of active validators [18]. It uses¹¹ a symbiotic Schelling game under “normal operations” with the choice between X and Y corresponding to $p = 3B$ and $d = 3B$. However, if more than E epochs pass without being finalized, then the incentive structure changes to a simple Schelling game with $p = 0$ and $d = 6B + l$, where l is the “inactivity leak,” an amount proportional to the validator’s remaining balance

that increases with the number of epochs that are not finalized. Then, in the most extreme case where all M validators accept the bribe in each epoch j for $j = i_0 + 1, i_0 + 3, \dots, j \leq \lfloor \frac{T}{\text{Epoch Time (min)}} \rfloor$, but X is finalized by having checkpoints in epochs $\lfloor \frac{T}{\text{Epoch Time (min)}} \rfloor - 1$ and $\lfloor \frac{T}{\text{Epoch Time (min)}} \rfloor$ justified, then the attacker must pay out:

$$\begin{aligned}
 & M \cdot (6B + \epsilon) \lfloor \frac{E}{2} \rfloor + \\
 & M \cdot (6B + \epsilon) \left(\left\lfloor \frac{T}{2 \cdot \text{Epoch Time (min)}} \right\rfloor - \lfloor \frac{E}{2} \rfloor - 1 \right) \\
 & + \sum_{j=i_0+1, i_0+3, \dots, j < \lfloor \frac{T}{\text{Epoch Time (min)}} \rfloor} l_j + \delta,
 \end{aligned}$$

where taking the continuous approximation to the inactivity leak of [23]

$$\begin{aligned}
 & \sum_{j=i_0+1, i_0+3, \dots, j < \lfloor \frac{T}{6.4 \text{min}} \rfloor} l_j \approx \\
 & \frac{1}{2} \cdot M \cdot \text{Initial Validator Balance} \cdot \left(1 - e^{-\lfloor \frac{T}{6.4 \text{Epoch Time (min)}} \rfloor^2 / 2\alpha} \right),
 \end{aligned}$$

with α a fixed parameter and δ is the total bribe that must be paid out in the last round. As the last round finalizes and hence converts back to a symbiotic Schelling game, we have¹² $\delta \leq M \cdot (6B + \epsilon)$. In Figure 2, we take $M = 120000$, $B = 8100$ gwei (based on current values as of April 2021 [25] calculating B from M using the formulas of [23]), $E = 4$, Initial Validator Balance=32 ETH, Epoch Time=6.4 min, and $\alpha = 2^{24}$, the (anticipated, long-term) value of α [23].

- To the degree that a double spend on a system using Serenity would require two conflicting transactions to be finalized, this is only possible insofar as it leads to a permanent fork of the network. One could consider attacks that attempt to trigger such a split by incentivizing validators to attest in such a way such that the inactivity leak eventually results in each fork perceiving the validators on the other forks as being penalized to the point where the remaining validators suffice to finalize [18], [24]. However, as such an attack requires some, but not all, validators voting for each fork, it does not fit into structures that we have considered in this section.

	PoW (Bitcoin)	PoW (Heuristic: Quickly Adjusting Difficulty)
Censorship	$\left\lfloor \frac{\text{Time to Censor (min)}}{2 \cdot \text{Avg. Block Time (min)}} \right\rfloor \left(\text{Avg. Block Reward} + \epsilon \right)$ $= \left\lfloor \frac{\text{Time to Censor (min)}}{18} \right\rfloor (6.85 + \epsilon) \text{ BTC}$	$\approx \left(\frac{\text{Time to Censor (min)}}{2} - \text{Difficulty Transition Periods} \right) \cdot M \cdot t \cdot \frac{\text{Avg. Block Reward}}{\tau}$
Double spend	$\text{Number of Blocks for Confirmation} \left(\text{Avg. Block Reward} + \epsilon \right)$ $= 41.1 + 6\epsilon \text{ BTC}$	$\approx \left(\text{Avg. Typical Confirmation Time (min)} - \text{Difficulty Transition Periods} \right) \cdot M \cdot t \cdot \frac{\text{Avg. Block Reward}}{\tau}$
	PoW (Ethereum)	Serenity PoS
Censorship	$\left\lfloor \frac{\text{Time to Censor (min)}}{2 \cdot \text{Avg. Block Time (min)}} \right\rfloor \left(\text{Avg. Block Reward} + \epsilon \right)$ $= \left\lfloor \frac{\text{Time to Censor (min)}}{.44} \right\rfloor (3.1 + \epsilon) \text{ ETH}$	$\approx M(6B + \epsilon) \left(\left\lfloor \frac{\text{Time to Censor (min)}}{2 \cdot \text{Epoch Time (min)}} \right\rfloor - 1 \right)$ $+ \frac{M}{2} \text{Init. Validator Balance} \cdot \left(1 - e^{-\left\lfloor \frac{\text{Time to Censor (min)}}{\text{Epoch Time (min)}} \right\rfloor^2 / 2\alpha} \right) + \delta$ $\approx (5.8 + 1.2 \cdot 10^5 \epsilon) \left\lfloor \frac{\text{Time to Censor (min)}}{12.8} \right\rfloor$ $+ 1.9 \cdot 10^6 \cdot \left(1 - e^{-\left\lfloor \frac{\text{Time to Censor (min)} \cdot 12.8}{6.4} \right\rfloor^2 / 2^{25}} \right) \text{ ETH}$
Double spend	$\text{Number of Blocks for Confirmation} \left(\text{Avg. Block Reward} + \epsilon \right)$ $= 62 + 20\epsilon \text{ ETH}$	NA

Figure 2: Budget requirements of various $p + \epsilon$ attacks on block production protocols.

So to the degree that the specific censorship attacks we consider on the PoW and the PoS versions of Ethereum are comparable, we see in [Figure 2](#) that censorship attacks on the PoS version require smaller budgets at short intervals of time, but substantially larger budgets on longer time scales due to the inactivity leak. On the other hand, the structure of Serenity provides greater resistance to double spending attacks, even removing the types of attacks we consider on PoW as possibilities. Note that it remains possible for an attacker to launch adapted versions of these attacks that might have different budget requirements; see [Section 9](#) and particularly [Example 9.1](#), where we consider some modified versions of the censorship attack that provide a different set of tradeoffs for the attacker.

In the context of performing a $p + \epsilon$ attack on a proof-of-work system, one can compare these deposit sizes to those of other smart contract-enforced attacks described in [\[26\]](#). 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 in a manner similar to the uncle-mining based attacks presented in [\[27\]](#). Nonetheless, as uncle block rewards on Ethereum are only made to the extent that the uncle blocks are subsequently referenced in blocks in the main chain [\[28\]](#), and are hence not guaranteed, we have not taken them into account in the budget requirements shown in

[Figure 2](#). Note that the quadratic costs manifest themselves in the heuristic version of PoW that we consider with quickly adjusting difficulty in that, even when one accounts for the balancing average block reward against τ to produce some average reward per unit time, there is an additional term of M , the number of active miners, in the required budgets.

Remark 6.3: Note that in systems such as [\[29\]](#) and [\[17\]](#) where deposits are made in a system token, $p + \epsilon$ bribes must also be paid in this token.¹³ Then, by having quadratic budget requirements in the redistributive case, one can have situations where the budget requirement of a sufficiently high appeal round exceeds the number of total tokens in circulation and is hence impossible. For example, in the Kleros “general court” where (as of April 2021) $d = 620$ PNK tokens of which the total circulating supply is 524 million [\[30\]](#), an appealable redistributive Schelling game vote with $M = 905$ participants would require a bribe budget of at least $905^2 \cdot 620$, exceeding the supply.

Remark 6.4: The large budget requirements of $p + \epsilon$ 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.

7. Counterattacks against $p + \epsilon$ attacks

We will 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. Specifically, we describe a counterattack that is possible in Schelling games where the appeal system is such that a voter in round i can trigger an appeal from the i th round to the next round by paying R_i . Here the actions available to voters in round i are:

- Vote X or vote Y
- Independent of her choice of vote, to trigger an appeal or not

See [Section 5](#) for our model of possible actions taken by the attacker. Then, voters can take the following strategy:

Counterattack 1:

In a given i th appeal round in which a given participant is a voter:

- Observe if the attacker can lock up enough funds to pay any required bribes.
 - If sufficient funds are locked up:
 - Vote Y to take the bribe
 - Pay R_i , the required appeal fees, so that the case progresses to the next round
 - If sufficient funds are not locked up:
 - Vote X
-

Namely, voters are said to be counterattackers if they vote to accept the bribe and then appeal in their vote round. [Counterattack 1](#) depends on the idea that if there is an appeal to a round where the attacker can no longer commit to enough money to pay required bribes, voters will revert back to the honest choice, X . Then the attacker's own bribes are turned against her as voters are willing to pay appeal fees in exchange for some percent chance of ultimately receiving the bribe. Consequently, as we will see below, this counterattack is particularly effective in redistributive Schelling games. See [Figure 3](#) for an illustration of this idea.

Figure 3: Example of multi-round counterattack in a redistributive Schelling game where voters can choose options favoring Alice or Bob, and Bob is executing a $p + \epsilon$ attack. Voters receive payouts from the coherence game and the attack contract, which are symbolized by the scale and the image of the attacker Bob respectively. Participants who execute Counterattack 1 have an amount deducted corresponding to paying to trigger an appeal. Participants are grouped horizontally by appeal round. In the first two rounds, counterattackers are shown on the left and non-counterattackers on the right; in the third round, we suppose Bob can no longer lock up enough funds to cover all of the bribes required of him, so voters default back to the “honest” choice of voting for Alice. Note that if either of the counterattackers in the first two rounds had voted for Bob without counterattacking, Bob would have won and all voters would have only received the coherence payout of p .

Concretely, if an attacker has a given budget b , this imposes limits on the number of appeal rounds for which she can commit to paying out $p + \epsilon$ bribes against a given

game. For our purposes, voters will not know the attacker's budget b exactly. However for some of our results, we will model voters as viewing this budget as a random variable B . Namely, a given attacker's budget is a sample from some fixed, known probability distribution f_B , $\text{Prob}(b_1 \leq B \leq b_2) = \int_{b_1}^{b_2} f_B(x)dx$, that captures the variability of budgets seen across the universe of all possible attackers.

Then, we define another random variable as a function of B by

$$L_* = \min \{l \in \mathbb{N} : B < \text{costs to continue attack until } l\text{th round}\}.$$

Namely, L_* reflects a voter's probabilistic assessment of the first round in which the attacker will not be capable of locking up the amounts required to continue the attack. These costs to continue the attack include the bribes considered in [Proposition 6.1](#) as well as potentially other costs (such as appeal fees depending on the appeal fee model).

Then, participants update their beliefs on this distribution as the attacker continues the attack in successive rounds. Specifically, if the attacker continues the attack through the i th round, voters will estimate her probability of continuing it further to the j th round for $j \geq i$ as $\text{Prob}(L_* > j | L_* > i)$.

Remark 7.1: If an attacker committed to a large attack budget at the beginning of an attack—for example, by placing it in a smart contract that could only be used to fund $p + \epsilon$ bribes—then we could consider this information as simply being included in the voters' prior beliefs on B (namely that voters imagine the attacker as being drawn from some corresponding subpopulation). For this work, we do not consider more sophisticated behavior in how the attacker provides information about her budget, such as by committing an additional amount, beyond what is required of her, in a smart contract during an intermediate round after the attack has begun, see [Section 5](#).

Remark 7.2: One could have potential complications from situations where the attacker has the budget required to pay partial, but not full bribes in a given round. In this section, we suppose that in any given appeal round the attacker credibly commits to paying all of the bribes for that round, or does not commit to paying any bribes, see [Section 5](#). Indeed, heuristically, one could imagine that each participant in Counterattack 1 places sufficient funds in a smart contract to automatically appeal the following *two* rounds, avoiding effects from rounds with partial bribes at the expense of increasing R_i . Attackers who spread out their budget further to pay partial bribes over a larger number of rounds are considered in [Section 9](#).

While voters observe L_* with uncertainty, there are sometimes natural upper bounds that they will have on it. Indeed, if the Schelling game permits a maximum total number of rounds of $l_{***} < \infty$, then $P(L_* > l_{***} + 1) = 0$. Moreover, we define

$$l_{**} = \min \{l \in \mathbb{N} \cup \{\infty\} : \text{Prob}(L_* > l) = 0\}.$$

Then $l_{**} \leq l_{***} + 1$. One sometimes has better natural upper bounds on l_{**} . Particularly, in Schelling games where bribes must be paid in terms of an asset with a limited supply, the quadratic growth of the required bribes seen in [Proposition 6.1](#) can cause them to exceed the total supply, as discussed in [Remark 6.3](#). In this case, there is a round where it is impossible to continue the attack, even as appeal is still possible, namely $l_{**} - 1 < l_{***}$.

As we will require the same set of assumptions on our parameters in several of our results, we will condense notation by saying:

Definition 7.3: $(M_i, R_i, c_1, c_2, c_3, d, p)$ are said to be well-designed if

- any user can trigger an appeal from the i th round to the $i + 1$ st round by paying R_i
- $M_{i+1} \leq c_1 M_i$ for all i , for some $c_1 \in \mathbb{R}_{>0}$.
- $M_i \geq 2$ for all i .
- $R_i \leq c_2 M_{i+1} p$ for some $c_2 \in \mathbb{R}_{>0}$.
- $d > c_3 p$, where $c_3 = \max \{2c_1 c_2 - 1, 0\}$.

Note that the constraints that make up the property of being “well-designed” are generally not very burdensome. Indeed, the designer of a Schelling game based system can ensure that the M_i grow reasonably and require sufficiently large deposits relative to payoffs so that these types of constraints hold before and independent of any eventual attacks.

Then we see:

Proposition 7.4: *Suppose that an attacker with a budget drawn from a distribution f_B launches a $p + \epsilon$ attack on a multi-round redistributive Schelling game such as in Definition 2.3 where round i voters can trigger an appeal from the i th round to the $i + 1$ st round by paying R_i . Suppose:*

- *Voters satisfy *. Moreover, f_B is common knowledge to voters.*
- *$(M_i, R_i, c_1, c_2, c_3, d, p)$ are well-designed.*
- *$l_{**} - 1 < l_{***}$, and in particular l_{**} is finite.*

Then there exists a pure strategy equilibrium where some (distinguished) voter executes [Counterattack 1](#) (i.e., takes the bribe and then appeals) in each round until the attacker is no longer capable of locking up sufficient resources to continue the bribe. Hence, in this equilibrium, the honest choice X eventually wins with probability 1

PROOF: We consider a situation where all voters vote Y in any round where the $p + \epsilon$ bribe is available, and one (distinguished) voter per round performs [Counterattack 1](#). Hence, this voter pays

$$R_i \leq c_2 M_{i+1} p \leq c_1 c_2 M_i p$$

to cover any required appeal fee in the i th round. We will show that this situation is an equilibrium.

Eventually, one arrives at a round (at latest l_{**}) where the attacker can no longer fund her attack. At this point, X wins. In any round where the $p + \epsilon$ attack bribes are available, voting X , whether subsequently appealing or not, is a dominated strategy.

Hence, we need only consider the strategies of voting Y and performing [Counterattack 1](#), or voting Y without performing [Counterattack 1](#). Then, following the payoffs shown in [Definition 2.3](#), participants receive $(M_i - 1)d + M_i p + \epsilon$ if Y wins from the coherence game and the bribe.

For the distinguished voter, if she deviates from the equilibrium by not paying the appeal fees, Y wins and her total payoff is p . On the other hand, if she pays the appeal fees and all of the other voters follow the candidate equilibrium, X wins and the distinguished voter receives a total payoff of $(M_i - 1)d + M_i p + \epsilon - R_i$. Hence, as $d > (2c_1 c_2 - 1)p$ and $M_i \geq 2$,

$$\begin{aligned} & (M_i - 1)d + M_i p + \epsilon - R_i \\ & \geq (M_i - 1)(2c_1 c_2 - 1)p + M_i p - c_1 c_2 M_i p \geq p \end{aligned}$$

so the distinguished voter does not have an incentive to deviate.

Similarly, we take the perspective of one of the non-distinguished voters. Then, as long as the other participants do not deviate from the candidate equilibrium, X wins regardless. Hence, if the non-distinguished voter does not counterattack, she receives

$(M_i - 1)d + M_i p + \epsilon$, and if she counterattacks she receives $(M_i - 1)d + M_i p + \epsilon - R_i$. Hence, the non-distinguished voters also do not have an incentive to deviate.

□

Nonetheless, the equilibrium of [Proposition 7.4](#) seems to be one that voters would be unlikely to fall into. Specifically, coordinating on which voters should be distinguished to pay the appeal fees seems difficult, as while each voter would prefer that *some* voter perform [Counterattack 1](#), they would prefer that it not be themselves. Here equilibrium formation is particularly difficult as one does not necessarily have repeated games where the same groups of voters are offered a $p + \epsilon$ bribe again. As users have the additional infrastructure of smart contracts at their disposal, one might propose that the voters in each round place a deposit in a smart contract that chooses one of them and obliges her to pay the appeal fees; however, this merely displaces the issue as then voters prefer that *some* voter(s) join this smart contract coordination while preferring it to not be themselves.

In fact, in any given appeal round, our situation is similar to a multiplayer version of the anti-coordination game of “Chicken” [\[31\]](#); [Proposition 7.4](#) essentially corresponds to the pure strategy equilibria observed in that game. In the following result we see that there can also exist a mixed strategy equilibrium where voters (of a given appeal round) act symmetrically, each paying the required appeal fees with some positive probability, mirroring the mixed equilibria observed in “Chicken.” Due to its symmetry, this equilibrium seems somewhat more natural and more likely to arise in practice than that of [Proposition 7.4](#). However, in order to show that voters will estimate the probability of the counterattack continuing a sufficient number of rounds *beyond* their own as being high enough that the counterattack is likely to eventually succeed and that it is worthwhile for them to counterattack themselves, we will require additional assumptions on the budget of the attacker.

Theorem 7.5: *Suppose that an attacker with a budget drawn from a distribution f_B launches a $p + \epsilon$ attack on a multi-round redistributive Schelling game such as in [Definition 2.3](#) where round i voters can trigger an appeal from the i th round to the $i + 1$ st round by paying R_i . Suppose:*

- *Voters satisfy *. Moreover, f_B is common knowledge to voters.*
- *$(M_i, R_i, c_1, c_2, c_3, d, p)$ are well-designed.*
- *$l_{**} - 1 < l_{***}$, and in particular l_{**} is finite.*
- *The distribution f_B is such that for all $i, j \in \mathbb{N}$ such that $i \leq j$ and $i \leq l_{**} - 2$,*

$$\text{Prob}(L_* > j + 1 | L_* > i + 1) \leq \text{Prob}(L_* > j | L_* > i). \quad (2)$$

Then there exists a (potentially mixed strategy) equilibrium where each voter USR_k in round i funds appeals following [Counterattack 1](#) with some probability $y_{i,k} \in [0, 1]$, and

$$\text{Prob}(X \text{ wins}) \geq \sup_{l_0 \in \mathbb{N}_{\geq 3}} \{T_{l_0}\}, \quad (3)$$

where

$$T_{l_0} = 1 - \text{Prob}(L_* > l_0 | L_* > 2) - \frac{2c_1 c_2 p(l_0 - 2)}{(d + p)(1 - \text{Prob}(L_* > l_0 | L_* > 2))}.$$

Here if $\text{Prob}(L_* > l_0 | L_* > 2) = 1$, for the sake of computing the supremum, we take the convention that $T_{l_0} = -\infty$. However, as $l_{**} < \infty$, there is some value of l_0 for which this supremum is finite. Indeed, for d sufficiently large relative to p , c_1 , c_2 , and the distribution f_B , [Inequality 3](#) is positive. We will explore this bound in the case of specific distributions that satisfy [Inequality 2](#) in [Corollary 7.7](#).

Remark 7.6: While this counterattack requires no altruism on the part of its participants, there are several practical limitations to this approach. The amounts that the counterattackers must pay in appeal fees become quite large (though much smaller than the budget required by the attacker as in [Proposition 6.1](#)), and may exceed the liquidity or risk thresholds of the counterattackers. Also, an important element of this scheme is that counterattackers are incentivized to appeal 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 appeal) 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 counterattack.

Before we continue with the proof of this result, we see that the condition of [Theorem 7.5](#) given by [Inequality 2](#) is satisfied by reasonable models of the attacker’s budget.

Indeed, one expects that the size of the budget for an attack roughly corresponds to the attacker's overall wealth. Then, standard, classical models for distributions of wealth are given by Pareto distributions [32]. More recent work [33] suggests that wealth is distributed exponentially for less wealthy segments of the population and then distributed by a Pareto distribution in the tail of wealthier individuals. As the $p + \epsilon$ attack requires large budgets (see [Proposition 6.1](#)), we would expect that attackers would already be a wealthier subpopulation of the general population of users; hence, in the following corollary and example, we will model attacker wealth by a Pareto distribution for simplicity. Nonetheless, as we have assumed a maximum appeal round, $l_{**} - 1$, at which the attacker can possibly commit to paying the bribe, a model for the attacker budget with an infinite tail is inappropriate. However, as such upper bounds often arise in practical applications, truncated versions of the Pareto distribution have also been widely applied [34]. Then, we consider truncated Pareto distributions:

$$\text{Prob}(B \geq x) = \begin{cases} \frac{\alpha L^\alpha x^{-\alpha-1}}{1 - \left(\frac{L}{H}\right)^\alpha}, & \text{for } L \leq x \leq H \\ 0, & \text{for } x > H \end{cases},$$

for $0 < L < H$. We see that such distributions satisfy Inequality 2. Note that the lower bound we obtain on $\text{Prob}(X \text{ wins})$ does not depend on l_{**} , so one can approximate a standard Pareto distribution arbitrarily well.

Corollary 7.7: *Let f_B be a truncated Pareto distribution with support between L and H and $\alpha > 0$. Then suppose that an attacker with a budget drawn from the distribution f_B launches a $p + \epsilon$ attack on a multi-round redistributive Schelling game such as in [Definition 2.3](#) where round i voters can trigger an appeal from the i th round to the $i + 1$ st round by paying R_i . Suppose:*

- *Voters satisfy *. Moreover, f_B is common knowledge to voters.*
- *$(M_i, R_i, c_1, c_2, c_3, d, p)$ are well-designed.*
- *$c_1 \in \mathbb{N}_{>1}$, and $M_{i+1} = c_1 M_i$ for all i (i.e., the number of voters grows exactly exponentially).*
- *$l_{**} - 1 < l_{***}$, and in particular l_{**} is finite.*
- *There exist constants $d_1, d_2 \in \mathbb{R}_{\geq 0}$ such that the additional budget to continue the attack, from the $i - 1$ st round to the i th round, is exactly $d_1 M_i^2 + d_2 M_i$ for all i .*

Then there exists a (potentially mixed strategy) equilibrium where each voter USR_k in round i funds appeals following [Counterattack 1](#) with some probability $y_{i,k} \in [0, 1]$, and

$$\text{Prob}(\text{X wins}) \geq \sup_{l_0 \in \mathbb{N}_{\geq 3}} \left\{ 1 - q_{l_0}^{\alpha+1} - \frac{2c_1 c_2 p(l_0 - 2)}{(d + p)(1 - q_{l_0}^{\alpha+1})} \right\},$$

where we compute the explicit quantity

$$q_{l_0} = \frac{d_1 \frac{M_1}{c_1^2 - 1} c_1^4 + d_2 \frac{1}{c_1 - 1} c_1^2 - d_1 \frac{M_1}{c_1^2} \left(1 + \frac{1}{c_1^2 - 1}\right) - d_2 \frac{1}{c_1} \left(1 + \frac{1}{c_1 - 1}\right)}{d_1 \frac{M_1}{c_1^2 - 1} c_1^{2l_0} + d_2 \frac{1}{c_1 - 1} c_1^{l_0} - d_1 \frac{M_1}{c_1^2} \left(1 + \frac{1}{c_1^2 - 1}\right) - d_2 \frac{1}{c_1} \left(1 + \frac{1}{c_1 - 1}\right)}.$$

Moreover, if $d_1 = d + p$ and $d_2 = \epsilon$ (namely the additional budget of continuing the attack in the i th round corresponds exactly to the amounts required for the bribes seen in [Proposition 6.1](#)), by minimizing the bound over all choices of ϵ , we see

$$\text{Prob}(\text{X wins}) \geq \sup_{l_0 \in \mathbb{N}_{\geq 3}} \left\{ 1 - q_{*,l_0}^{\alpha+1} - \frac{2c_1 c_2 p(l_0 - 2)}{(d + p)(1 - q_{*,l_0}^{\alpha+1})} \right\},$$

where we compute the explicit quantity

$$q_{*,l_0} = \max \left\{ \frac{\frac{c_1^4}{c_1^2 - 1} - \frac{1}{c_1^2} \left(1 + \frac{1}{c_1^2 - 1}\right)}{\frac{c_1^{2l_0}}{c_1^2 - 1} - \frac{1}{c_1^2} \left(1 + \frac{1}{c_1^2 - 1}\right)}, \frac{\frac{1}{c_1 - 1} c_1^2 - \frac{1}{c_1} \left(1 + \frac{1}{c_1 - 1}\right)}{\frac{1}{c_1 - 1} c_1^{l_0} - \frac{1}{c_1} \left(1 + \frac{1}{c_1 - 1}\right)} \right\}.$$

The proof of this corollary depends on showing that the tail of the distribution decreases quickly enough that [Inequality 2](#) holds in this case. Then one computes the explicit lower bound given by [Inequality 3](#). This is a rather straightforward calculus argument, and is included in [Appendix A](#) for completeness.

Example 7.8: Take the attacker budget to be distributed by a truncated Pareto distribution with $\alpha = 2.1$ - the value observed for UK wealth distribution in [\[33\]](#), the cost of continuing the attack from the $i - 1$ st to the i th round to be exactly the $M_i^2(d + p) + M_i \epsilon$ that must be locked up in bribes according to Proposition 6.1,

$c_1 = c_2 = 2$, and $d \geq 15p$. Then, one can use [Corollary 7.7](#) with $l_0 = 3$ to see that, in the mixed strategy equilibrium, X wins and the attack fails at least 39.1% of the time. This implies a corresponding non-zero lower bound on the expected cost of the attack. Note that this lower bound holds for any ϵ that the attacker might choose. Generally, there is a trade-off that can be made between decreasing the size of the deposit d relative to p and increasing the lower bound on the rate at which $p + \epsilon$ attacks fail, see [Figure 4](#).

Figure 4: We show lower bounds on the rate of failure of $p + \epsilon$ attacks against redistributive Schelling games in the mixed strategy equilibrium of [Corollary 7.7](#) when $\alpha = 2.1$, $c_2 = 2$, and for varying choices of c_1 . In general there is a trade-off between increasing the required deposits (relative to payoffs) and attack resistance. We see that for $d/p = 10$ we already have meaningful lower bounds on the rate of attack failure.

We will now develop the tools necessary to prove [Theorem 7.5](#), fixing the following notation that we will use towards this proof over the remainder of this section. We will show the existence of $y_{i,k} \in [0, 1]$ such that each voter in round i , $\mathcal{USR}_k$, choosing to trigger an appeal with probability $y_{i,k}$ gives rise to the equilibrium discussed in the statement of [Theorem 7.5](#). For given choices of $y_{i,k}$, we denote by

$$K_i \sim \sum_k \text{Bernoulli}(y_{i,k})$$

the number of participants who choose to appeal in the i th round. Similarly, fixing a distinguished participant $USR = USR_{k_0}$, we denote by

$$K'_i \sim \sum_{k \neq k_0} \text{Bernoulli}(y_{i,k})$$

the number of participants other than USR who choose to appeal in the i th appeal round.

Moreover, again from the perspective of USR who considers counterattacking, namely triggering an appeal, in round i , we denote

$$P_i(X|C) = \text{Prob} \left(\begin{array}{c|c} X & \text{in } i\text{th round and} \\ \text{wins} & USR \text{ counterattacks} \end{array} \right)$$

$$P_i(X|N) = \text{Prob} \left(\begin{array}{c|c} X & \text{in } i\text{th round and } USR \\ \text{wins} & \text{doesn't counterattack} \end{array} \right),$$

etc.

In any round where the $p + \epsilon$ attack bribes are available, voting X , whether followed by counterattacking or not, is a dominated strategy, so it can be excluded. Hence, we can calculate the payoff table for USR given by appealing in round i versus not counterattacking under the assumption that there are zero votes for X , see [Figure 5](#).

	X wins	Y wins
Counterattack	$(M_i - 1)d + M_i p + \epsilon - R_i^*$	$p - R_i^*$
Don't Counterattack	$(M_i - 1)d + M_i p + \epsilon$	p

Figure 5: Payoffs for executing [Counterattack 1](#) versus not counterattacking and voting Y , once the possibility of voters voting X , a dominated strategy, has been eliminated.

Note the R_i^* in the payoff table, which are given by the partial appeal payments of Equation 1. These values implicitly depend on the $y_{i,k}$; however, by our assumptions on our appeal fee models this dependence is continuous. Also note that we have $0 \leq R_i^* \leq R_i$.

We begin with a lemma that explores the probability that at least one voter appeals in a given round. In all of the following lemmas we assume that voters satisfy $*$ and that f_B

is common knowledge to voters.

Lemma 7.9: *Let $i \in \mathbb{N}$, $i \leq l_{**} - 1$. Suppose that $(M_i, R_i, c_1, c_2, c_3, d, p)$ are well-designed, and*

$$\text{Prob} \left(\begin{array}{c|c} X & \text{in } j\text{th} \\ \hline \text{wins} & \text{round} \end{array} \right) > \frac{R_{j-1}}{(M_{j-1} - 1)(d + p) + \epsilon}$$

holds for all $j \in \mathbb{N}$ such that $i < j \leq l_{**}$. Then, $P_i(X|C) > 0$, and there exists an equilibrium where each round i voter USR_k triggers an appeal independently with probability $y_{i,k}$ for some $y_{i,k} \in [0, 1]$, with

$$\text{Prob}(K_i \geq 1) \geq 1 - \frac{R_i}{((M_i - 1)(d + p) + \epsilon) P_i(X|C)}.$$

PROOF: Take USR a voter in round i . We calculate the expected returns from USR 's perspective using the payoffs of [Figure 5](#):

$$\begin{aligned} & E \left[\begin{array}{c} \text{Do not counter-} \\ \text{attack in } i\text{th round} \end{array} \right] \\ &= P_i(X|N)((M_i - 1)d + M_i p + \epsilon) + P_i(Y|N)p \end{aligned}$$

and

$$\begin{aligned} & E \left[\begin{array}{c} \text{Counterattack} \\ \text{in } i\text{th round} \end{array} \right] \\ &= -R_i^* + P_i(X|C)((M_i - 1)d + M_i p + \epsilon) + P_i(Y|C)p. \end{aligned}$$

The system is structured so that a single counterattacker in the i th round is sufficient for the game to progress to the $i + 1$ st round. Then

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

In particular, our assumed lower bound for $j = i + 1$ gives us $P_i(X|C) > 0$.

Moreover,

$$\begin{aligned}
 P_i(X|N) &= \text{Prob} \left(\begin{array}{c|c} \text{progress to} & \mathcal{USR} \text{ does not} \\ i + 1\text{st round} & \text{counterattack} \end{array} \right) \\
 &\quad \cdot \text{Prob} \left(\begin{array}{c|c} X & \text{in } i + 1\text{st} \\ \text{wins} & \text{round} \end{array} \right) \\
 &= \text{Prob}(K'_i \geq 1) P_i(X|C)
 \end{aligned}$$

We define

$$\begin{aligned}
 &f(y_{i,1}, \dots, y_{i,N}) \\
 &= E \left[\begin{array}{c} \text{Counterattack} \\ \text{in } i\text{th round} \end{array} \right] - E \left[\begin{array}{c} \text{Do not counter-} \\ \text{attack in } i\text{th round} \end{array} \right].
 \end{aligned}$$

Then, if $\mathcal{USR}$ believes that no other participant will counterattack, she estimates

$$\begin{aligned}
 f(0, \dots, 0) &= E \left[\begin{array}{c} \text{Counterattack} \\ \text{in } i\text{th round} \end{array} \right] - E \left[\begin{array}{c} \text{Do not counter-} \\ \text{attack in } i\text{th round} \end{array} \right] \\
 &= P_i(X|C) [((M_i - 1)(d + p) + \epsilon)] - R_i^* \geq 0,
 \end{aligned}$$

where the last inequality uses

$$\begin{aligned}
 P_i(X|C) &= \text{Prob} \left(\begin{array}{c|c} X & \text{in } i + 1\text{st} \\ \text{wins} & \text{round} \end{array} \right) \\
 &\geq \frac{R_i}{[((M_i - 1)(d + p) + \epsilon)]} \geq \frac{R_i^*}{[((M_i - 1)(d + p) + \epsilon)]}.
 \end{aligned}$$

Note that $P_i(X|C)$ is independent of $y_{i,k}$ for all k (though it depends on the $y_{j,k}$ for $j > i$). On the other hand, we have that R_i^* is continuous in the $y_{i,k}$ by our assumption in [Section 2](#), and $\text{Prob}(K'_i \geq 1)$ is continuous in $y_{i,k}$ by properties of Bernoulli distributions. Hence, $f(y_{i,1}, \dots, y_{i,N})$ is continuous. Then there either exists some values

of $y_{i,1}, \dots, y_{i,N} \in [0, 1]$ such that $f(y_{i,1}, \dots, y_{i,N}) = 0$ or $f(y_{i,1}, \dots, y_{i,N}) > 0$ for all $y_{i,1}, \dots, y_{i,N} \in [0, 1]$. In the latter case, triggering an appeal is a dominant strategy for $\mathcal{USR}$, so we can take $y_{i,k} = 1$ for all k , which yields an equilibrium where $\text{Prob}(K_i \geq 1) = 1$, so the statement holds.

Hence, we can suppose that we have $y_{i,1}, \dots, y_{i,N} \in [0, 1]$ such that $f(y_{i,1}, \dots, y_{i,N}) = 0$, which also yields an equilibrium. Here voters are willing to randomize and,

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

$$\Rightarrow R_i \geq R_i^* = ((M_i - 1)(d + p) + \epsilon) P_i(X|C) \left(1 - \frac{P_i(X|N)}{P_i(X|C)} \right)$$

$$\Rightarrow \text{Prob}(K'_i \geq 1) = \frac{P_i(X|N)}{P_i(X|C)} \geq 1 - \frac{R_i}{((M_i - 1)(d + p) + \epsilon) P_i(X|C)}.$$

Then, we note that $K_i \geq K'_i$, to conclude the proof.

□

Now we consider a technical lemma concerning the growth properties of a quantity that is relevant to the lower bound of [Theorem 7.5](#).

Lemma 7.10: Let $d, p, c_1, c_2 \in \mathbb{R}_{>0}$. Define, for $i \geq 1$,

$$s_i = \sup_{l_0 \in \mathbb{N}_{>i}} \{T_{l_0,i}\}$$

where

$$T_{l_0,i} = 1 - \text{Prob}(L_* > l_0 | L_* > i) -$$

$$(l_0 - i) \frac{2c_1 c_2 p}{(d + p)(1 - \text{Prob}(L_* > l_0 | L_* > i))}$$

and where, for the sake of computing the supremum, if $\text{Prob}(L_* > l_0 | L_* > i) = 1$ then $T_{l_0, i}$ is considered to be $-\infty$.

Let $t \in \mathbb{N}$. Suppose that for all $i, j \in \mathbb{N}$ such that $i \leq \min\{j, t\}$

then

$$\text{Prob}(L_* > j + 1 | L_* > i + 1) \leq \text{Prob}(L_* > j | L_* > i).$$

Then $s_i \leq s_{i+1}$ for all $i \in \mathbb{N}$, $1 \leq i \leq t$.

PROOF: Let $i, l_0 \in \mathbb{N}$ such that $i < l_0$, $i \leq t$, and $\text{Prob}(L_* > l_0 | L_* > i) < 1$. Then

$$\text{Prob}(L_* > l_0 + 1 | L_* > i + 1) < 1$$

as well. By our assumptions on L_* ,

$$\begin{aligned} \text{Prob}(L_* > l_0 | L_* > i) - \text{Prob}(L_* > l_0 + 1 | L_* > i + 1) &\geq 0 \\ &\geq \frac{(l_0 - i) \frac{2c_1 c_2 p}{d+p}}{1 - \text{Prob}(L_* > l_0 + 1 | L_* > i + 1)} - \\ &\quad \frac{(l_0 - i) \frac{2c_1 c_2 p}{d+p}}{1 - \text{Prob}(L_* > l_0 | L_* > i)}. \end{aligned}$$

Hence, for every element l_0 considered in the supremum for s_i (that gives a lower bound other than $-\infty$), there is a corresponding element $l_0 + 1$ that gives a term least as large in the supremum of s_{i+1} .

□

We now have another lemma where we show that the bounds given by [Lemma 7.9](#) on the chances that at least one party appeals are sufficient to have useful lower bounds on the probability that the attack ultimately fails.

Lemma 7.11: Suppose $(M_i, R_i, c_1, c_2, c_3, d, p)$ are well-designed, and the values of $y_{i,k}$ are such that $P_i(X|C) > 0$ and

$$\text{Prob}(K_i \geq 1) \geq 1 - \frac{R_i}{((M_i - 1)(d + p) + \epsilon) P_i(X|C)}.$$

Then,

1. if $i = 1, \dots, l_{**} - 1$, $l_0 > i$, and $\text{Prob}(L_* > l_0 | L_* > i) < 1$,

$$\begin{aligned} \text{Prob} \left(\begin{array}{c|c} X & \text{in } i\text{th} \\ \text{wins} & \text{round} \end{array} \right) &\geq 1 - \text{Prob}(L_* > l_0 | L_* > i) \\ &\quad - (l_0 - i - 1) \frac{2c_1 c_2 p}{(d + p)(1 - \text{Prob}(L_* > l_0 | L_* > i))}. \end{aligned}$$

2. if $i = 2, \dots, l_{**} - 1$ and $s_2 > 0$ (for s_i as in [Lemma 7.10](#)),

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

PROOF: Note that for all $l_0 \in \mathbb{N}$ such that $l_0 > i$, as X wins if there are appeals until round l_0 unless the attacker has the resources to continue the attack,

$$P_i(X|C) \geq \prod_{j=i+1}^{l_0-1} \text{Prob}(K_j \geq 1) \cdot (1 - \text{Prob}(L_* > l_0 | L_* > i)).$$

(Note here that there is the possibility that $l_0 - 1 < i + 1$ in which case the empty product here is taken to be 1.)

Suppose $\text{Prob}(L_* > l_0 | L_* > i) < 1$. Then,

$$\prod_{j=i}^{l_0-1} \text{Prob}(K_j \geq 1)$$

$$\begin{aligned}
&\geq \prod_{j=i+1}^{l_0-1} \text{Prob}(K_j \geq 1) \left(1 - \frac{R_i}{((M_i - 1)(d + p) + \epsilon)P_i(X|C)} \right) \\
&\geq \prod_{j=i+1}^{l_0-1} \text{Prod}(K_j \geq 1) \\
&\quad - \frac{R_i}{((M_i - 1)(d + p) + \epsilon) (1 - \text{Prob}(L_* > l_0 | L_* > i))} \\
&\geq 1 - \sum_{j=i}^{l_0-1} \frac{R_j}{((M_j - 1)(d + p) + \epsilon) (1 - \text{Prob}(L_* > l_0 | L_* > i))},
\end{aligned}$$

where the last step uses a simple induction argument.

Then, as X wins if there are appeals until round l_0 unless the attacker has the resources to continue the attack,

$$\begin{aligned}
&\text{Prob} \left(\begin{array}{c|c} X & \text{in } i\text{th} \\ \text{wins} & \text{round} \end{array} \right) \tag{4} \\
&\geq \prod_{j=i}^{l_0-1} \text{Prod}(K_j \geq 1) - \text{Prob}(L_* > l_0 | L_* > i) \\
&\geq 1 - \text{Prob}(L_* > l_0 | L_* > i) \\
&\quad - \sum_{j=i}^{l_0-1} \frac{R_j}{((M_j - 1)(d + p) + \epsilon) (1 - \text{Prob}(L_* > l_0 | L_* > i))} \\
&\geq 1 - \text{Prob}(L_* > l_0 | L_* > i)
\end{aligned}$$

$$-(l_0 - i - 1) \frac{2c_1 c_2 p}{(d + p) (1 - \text{Prob}(L_* > l_0 | L_* > i))},$$

where in the last inequality we use that $\epsilon > 0$, $M_j \geq 2$, and the bounds on R_j coming from the growth assumptions on M_j and on the appeal fees.

Now we consider the second part of the claim. As $s_2 > 0$, by [Lemma 7.10](#), $s_i > 0$ for all $2 \leq i \leq l_{**} - 1$. As these s_i are defined as suprema, this implies that for all such i , there exists $l_0 \in \mathbb{N}_{>i}$, such that $\text{Prob}(L_* > l_0 | L_* > i) < 1$ and

$$\begin{aligned} & 1 - \text{Prob}(L_* > l_0 | L_* > i) \\ & -(l_0 - i) \frac{2c_1 c_2 p}{(d + p) (1 - \text{Prob}(L_* > l_0 | L_* > i))} > 0 \\ & \Rightarrow 1 - \text{Prob}(L_* > l_0 | L_* > i) \\ & -(l_0 - i - 1) \frac{2c_1 c_2 p}{(d + p) (1 - \text{Prob}(L_* > l_0 | L_* > i))} - \frac{2c_1 c_2 p}{d + p} > 0 \\ & \Rightarrow \text{Prob} \left(\begin{array}{c|c} \text{X} & \text{in ith} \\ \text{wins} & \text{round} \end{array} \right) > \frac{2c_1 c_2 p}{d + p} > \frac{R_{i-1}}{(M_{i-1} - 1)(d + p) + \epsilon}, \end{aligned}$$

where the last line uses our previous lower bound.

□

Now we complete the proof of [Theorem 7.5](#). Essentially, we will use the assumption that $l_{**} - 1 < l_{***}$, namely there is a finite maximum round under which it is possible for the attacker to have the budget necessary for the attack where appeal is nonetheless possible, to perform a kind of backward induction. We see that at each round participants are sufficiently confident in the chance of the honest option eventually winning in appeal relative to the payoff they can receive from a potential

“lone voice of reason jackpot” that they are willing to perform the counterattack with some probability.

Proof of Theorem 7.5: As voting for X is a dominated strategy, with or without appealing, due to the $p + \epsilon$ bribe, we can consider users who choose between the strategy of executing [Counterattack 1](#), namely voting Y followed by triggering an appeal, and the strategy of voting Y without triggering an appeal. Then the payoffs for these strategies are shown in [Figure 5](#).

If $s_2 \leq 0$, following the notation of [Lemma 7.10](#), then the statement of the theorem holds trivially. Hence, throughout this argument we can assume that $s_2 > 0$.

We have by the definition of l_{**} ,

$$\text{Prob} \left(\begin{array}{c|c} X & \text{in } l_{**}\text{th} \\ \text{wins} & \text{round} \end{array} \right) = 1.$$

We will prove by induction that for all $i = 2, \dots, l_{**} - 1$ that: for all $j \in \mathbb{N}$ such that $i \leq j \leq l_{**} - 1$ and all $k \leq l_{**}$, there exist $y_{i,k} \in [0, 1]$ such that there is an equilibrium where each voter $\mathcal{USR}_k$ in the i th round triggers an appeal independently with probability $y_{i,k}$, which yields

$$\text{Prob} \left(\begin{array}{c|c} X & \text{in } i\text{th} \\ \text{wins} & \text{round} \end{array} \right) > \frac{R_{i-1}}{(M_{i-1} - 1)(d + p) + \epsilon}. \quad (5)$$

Note that $R_{l_{**}-1} \leq c_2 M_{l_{**}} p \leq c_1 c_2 M_{l_{**}-1} p$, as well as $M_{l_{**}-1} \geq 2$, and $\epsilon > 0$, so if $d > (2c_1 c_2 - 1)p$,

$$\text{Prob} \left(\begin{array}{c|c} X & \text{in } l_{**}\text{th} \\ \text{wins} & \text{round} \end{array} \right) = 1 > \frac{R_{l_{**}-1}}{(M_{l_{**}-1} - 1)(d + p) + \epsilon}. \quad (6)$$

Hence we can apply [Lemma 7.9](#) to obtain the required $y_{l_{**}-1,k} \in [0, 1]$ such that $P_{l_{**}-1}(X|C) > 0$ and

$$\text{Prob}(K_{l_{**}-1} \geq 1) \geq 1 - \frac{R_{l_{**}-1}}{(M_{l_{**}-1} - 1)(d + \epsilon)P_{l_{**}-1}(X|C)}.$$

Then we apply [Lemma 7.11](#) to show the induction hypothesis for $i = l_{**} - 1$.

Now assume that we have some $2 \leq i < l_{**} - 1$ such that the induction hypothesis holds for $i + 1$. We employ a similar argument. We apply [Lemma 7.9](#) (where we use the induction hypothesis and [Inequality 6](#) for the $j = l_{**}$ cases) to get $y_{i,k} \in [0, 1]$ such that

$$\text{Prob}(K_i \geq 1) \geq 1 - \frac{R_i}{((M_i - 1)(d + p) + \epsilon) P_i(X|C)}.$$

Then, using [Lemma 7.11](#), the induction hypothesis holds for i .

Then we consider a variation on this argument for $i = 1$. We use our above results on $i = 2$, [Lemma 7.9](#), and the first part of [Lemma 7.11](#) to obtain:

$$\text{Prob} \left(\begin{array}{c|c} X & \text{in 1st} \\ \text{wins} & \text{round} \end{array} \right) \geq 1 - \text{Prob}(L_* > l_0 | L_* > 1) - (l_0 - 2) \frac{2c_1 c_2 p}{(d+p)(1 - \text{Prob}(L_* > l_0 | L_* > 1))},$$

for all $l_0 \geq 2$ (for which $\text{Prob}(L_* > l_0 | L_* > 1) < 1$). So the lower bound in the statement of the theorem follows from noting that

$$\text{Prob}(L_* > l_0 | L_* > 1) \leq \text{Prob}(L_* > l_0 | L_* > 2)$$

for all $l_0 \geq 3$.

□

Remark 7.12: If a would-be attacker with a relatively small budget realizes that her $p + \epsilon$ attack has a high risk of failing due to [Counterattack 1](#) and [Theorem 7.5](#), then she can decide not to launch her attack. Then to the degree that many smaller budget would-be attackers do not participate, this shapes the distribution f_B that voters use to estimate the budgets of the attackers who *do* launch their attacks, and hence the f_B to which [Theorem 7.5](#) applies. In order to study potential equilibria for *which attackers* decide to launch their attacks based on their expected returns, one would need a joint distribution on the wealth of would-be attackers and how much they have to gain by a successful attack. This is a potentially interesting direction for future work, possibly via agent-based modeling.

8. Moral and cognitive costs

Even though, in the abstract, accepting a $p + \epsilon$ 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 [5]. 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 [35] for a summary of theoretical and experimental results along these lines.

Combining these two types of costs together, we assume a penalty of m if a participant votes for Y either due to moral or cognitive costs. Namely, m is subtracted from the payoffs given in Definition 2.1, Definition 2.2, and Definition 2.3 for any case where USR votes for Y , regardless of the eventual winner. For the simple Schelling games of Definition 2.1, this gives the payoff matrix of:

	X wins	Y wins
USR votes X	p	$-d$
USR votes Y	$p + \epsilon - m$	$p - m$

If $\epsilon - m$ 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). On the other hand, when $\epsilon - m$ and $p + d - m$ have different signs, we see that:

Proposition 8.1: *Let m be such that $\epsilon < m < p + d$. Suppose voters satisfy $*$ and m is common knowledge. Then, if there is a $p + \epsilon$ attack on a Schelling game where there is a cognitive/moral cost m that is constant for all voters, we have:*

- *For all simple, symbiotic, and redistributive games where the payoffs of Definition 2.1, Definition 2.2, and Definition 2.3, respectively, are modified to impose an additional moral/cognitive cost of m to any vote for Y , there is a mixed strategy equilibrium where voters randomize between voting for X and voting for Y , and the attack succeeds with a probability strictly between 0 and 1.*
- *Let $\delta > 0$, and consider a simple Schelling game where the payoffs of Definition 2.1 are modified impose an additional moral/cognitive cost of m to any vote for Y . Then there exists some $M_\delta \in \mathbb{N}$ such that if the number of independent voters M is greater than M_δ , there is a mixed strategy equilibrium where voters randomize between voting for X and voting for Y and*

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

Towards proving this result, we first consider the following lemma:

Lemma 8.2: *Consider a binary game consisting of a vote between options X and Y where the possible payoffs in each of the four cases: vote X, X wins; vote X, Y wins; vote Y, X wins; and vote Y, Y wins are functions of the number of voters other than $\mathcal{USR}$ who vote X, and which option wins depends only on the total number of votes that each of X and Y receive. Suppose voters take actions to maximize their expected returns. Consider the function*

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

where for each $\pi \in [0, 1]$ the expectations are calculated based on each of the voters other than $\mathcal{USR}$ voting X with probability π and voting Y with probability $1 - \pi$. Then if $f(0) \cdot f(1) < 0$, there exists a mixed Nash equilibrium where each voter is willing to randomize by independently voting for X with probability $\pi_0 \in (0, 1)$.

PROOF: We write $\text{payoff}_{\mathcal{USR} \text{ votes } X}(x)$ for the payoff when $\mathcal{USR}$ and exactly x other voters vote for X, $\text{payoff}_{\mathcal{USR} \text{ votes } Y}(x)$ for the payoff when $\mathcal{USR}$ votes for Y and exactly x voters vote for X, etc. These payoffs take into account which of X and Y wins for each vote total.

Then one can compute

$$\begin{aligned} f(\pi) &= \sum_i \text{payoff}_{\mathcal{USR} \text{ votes } X}(x) \cdot \text{Prob}(x = i) \\ &\quad - \sum_i \text{payoff}_{\mathcal{USR} \text{ votes } Y}(x) \cdot \text{Prob}(x = i). \end{aligned}$$

However, $\text{payoff}_{\mathcal{USR} \text{ votes } X}(x)$ and $\text{payoff}_{\mathcal{USR} \text{ votes } Y}(x)$ do not depend on π , and $\text{Prob}(x = i)$ is continuous in π for each i by properties of binomial distributions. Hence, f is continuous. Then, using the assumption that $f(0) \cdot f(1) < 0$, by the Intermediate Value Theorem, there exists some π_0 such that $f(\pi_0) = 0$. Hence, there exists an equilibrium where all participants are willing to randomize their votes, voting X with probability π_0 .

□

Now we can show:

Proof of Proposition 8.1: In all three cases, if all voters other than $\mathcal{USR}$ vote X, $\mathcal{USR}$'s payoff for voting X is p , whereas her payoff for voting Y is $p + \epsilon - m$. On the other hand, if all voters other than $\mathcal{USR}$ vote Y, $\mathcal{USR}$'s payoff for voting X is $-d$, whereas her payoff for voting Y is $p - m$. Then as

$$(-d - (p + m)) \cdot (p - (p + \epsilon - m)) = (-d - p + m) \cdot (m - \epsilon) < 0,$$

we can apply [Lemma 8.2](#), where the payoffs include the bribes as well as the cognitive/moral costs. Then 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 a mixed strategy equilibrium.

Now suppose that we restrict ourselves to simple Schelling games (with moral/cognitive costs). Then, if the number of independent voters M is sufficiently large, considering the (binomial) density function of how many votes X receives, one has

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

We denote $\text{Prob}(X \text{ wins} | \mathcal{USR} \text{ votes Y}) = P(X|Y)$, and $\text{Prob}(X \text{ wins} | \mathcal{USR} \text{ votes X}) = P(X|X)$. Then, using the fact that in equilibrium $\mathcal{USR}$ is indifferent between voting X and Y, we have

$$\begin{aligned} E[\text{vote X}] &= E[\text{vote Y}] \\ \Rightarrow pP(X|X) - d(1 - P(X|X)) \\ &= (p + \epsilon - m)P(X|Y) + (p - m)(1 - P(X|Y)) \\ \Rightarrow (p + d)P(X|X) - d &= \epsilon P(X|Y) + p - m \end{aligned}$$

$$\Rightarrow (p + d)P(X|Y) - d \leq \epsilon P(X|Y) + p - m, \text{ and}$$

$$\epsilon P(X|Y) + (p - m) < (p + d)(P(X|Y) + \delta) - d.$$

Then

$$\frac{p + d - m - (p + d)\delta}{p + d - \epsilon} < P(X|Y)$$

$$= \text{Prob}(X \text{ wins} | \mathcal{USR} \text{ votes } Y) \leq \frac{p + d - m}{p + d - \epsilon}, \text{ and}$$

$$\frac{p + d - m - (p + d)\delta}{p + d - \epsilon} < P(X|X)$$

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

Then 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 $p + \epsilon$ attack to be effective, further increasing the budget requirements we observed in [Section 6](#).

Remark 8.3: One could also evaluate [Counterattack 1](#) and build on our results from [Section 7](#) from the perspective of moral and cognitive costs. Here, one might expect that while participating in such a counterattack scheme may be considered by participants as more moral than accepting a bribe, this would likely require high cognitive costs.

9. Adaptations of $p + \epsilon$ attacks

In [Section 6](#) we saw that the budget required to perform a $p + \epsilon$ attack on a redistributive Schelling game across multiple appeal rounds grows quadratically in the number of voters and in [Section 7](#) we saw that counterattacking 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. Then, for the attacker to make voting for her a dominant strategy, she must offer a correspondingly high bribe.

We have so far identified two natural approaches for the attacker to preserve the spirit of a $p + \epsilon$ 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 H .

In this section we will consider the implications of these two approaches compared to “pure” $p + \epsilon$ attacks. While these adaptations can be applied to all of the structures of $p + \epsilon$ attack considered in [Section 2](#), considering our observations in [Section 6](#) and [Section 7](#), they are of particular relevance to redistributive Schelling games, so we will concentrate on this case.

9.1 Last-round only $p + \epsilon$ attacks

With a last round only bribe, participants are incentivized to vote for the attacker if:

- They believe they are in the last round; or
- If they think that the voters in the ultimate round will take the bribe; then they should vote for the attacker to vote in agreement 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. Note to the degree that voters attempt to evaluate whether they are one of these situations, this adds to the cognitive costs of the attack.

Example 9.1: Consider again the censorship attack on Serenity proof-of-stake discussed in [Example 6.2](#). The division of blocks into epochs in Serenity provides a natural appeal structure here; we saw the bribe required in the last epoch of this attack were upper bounded by $\delta \leq M \cdot (6B + \epsilon)$. Then, applying the values of parameters previously indicated implies that the last-round version of this attack only requires a total budget of $5.8 + 120000\epsilon$ ETH.

9.2 Capped $p + \epsilon$ attacks

In the alternative, where the bribe payout is capped at $H < Md$, 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 for the eventual winner (without losing her deposit of d). In this section, we study an equilibrium that arises.

We denote by H 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 $H - d$ after losing her deposit. Appropriately adapting the payoffs described in [Definition 2.3](#) gives rise to the following payoff table:

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

We see:

Proposition 9.2: *Consider a redistributive Schelling game where bribes are capped by H . Suppose*

- $d + p < H < Md + Mp$,
- *voters satisfy * and the payoff matrix above, including H , is common knowledge.*
- *the voters all expect the attack to eventually fail, possibly after future appeals, with probability one.*

Let $\delta > 0$. Then there exists $M_0 \in \mathbb{N}$ and a mixed strategy equilibrium where voters randomize between voting for X and voting for Y . Furthermore, in this equilibrium, if $M \geq M_0$, then

$$\text{Prob}(\text{voter takes bribe}) \leq 1 - \frac{d + p}{H} + \delta.$$

PROOF: Consider voters adopting a mixed strategy where they vote for X with probability π_X . In order to apply [Lemma 8.2](#), we define $f(\pi_X) = E[\text{vote X}] - E[\text{vote Y}]$, and compute

$$\begin{aligned} f(0) &= (M - 1)d + Mp - \min \{ (M - 1)d + Mp + \epsilon, H - d \} \\ &= \max \{ -\epsilon, Md + Mp - H \}, \end{aligned}$$

$$f(1) = p - \min \{ p + \epsilon, H - d \} = \max \{ -\epsilon, d + p - H \}.$$

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 , H , M , and p , we have $f(0) = Md + Mp - H > 0$ and $f(1) = \max \{ -\epsilon, d + p - H \} < 0$, so by [Lemma 8.2](#), there exists some $\pi_X \in (0, 1)$ that yields a mixed strategy equilibrium.

Using the assumption that the final ruling is X with probability one, we compute

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

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

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

So taking $j = k + 1$ and using the fact that the sum over the binomial probability mass function is one,

$$\begin{aligned} E[\text{vote } X] &= p + \frac{1 - \Pi_X}{\Pi_X} \left(\sum_{j=1}^{M-1} (d + p) \binom{M-1}{j} \Pi_X^j (1 - \Pi_X)^{M-j-1} \right) \\ &= p + \frac{1 - \Pi_X}{\Pi_X} (d + p) (1 - (1 - \Pi_X)^{M-1}) . \end{aligned}$$

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

$$\frac{1 - \Pi_X}{\Pi_X} (d + p) (1 - (1 - \Pi_X)^{M-1}) \leq H - d - p.$$

Note the function $g(z) = \frac{(d+p)(1-z)}{H-(d+p)z}$ is continuous for $z \in (0, \frac{H}{d+p})$. Thus, there exists $0 < \delta_0 < \frac{H}{d+p}$ such that if $0 < z < \delta_0$,

$$\frac{d + p}{H} \leq \frac{(d + p)(1 - z)}{H - (d + p)z} + \delta.$$

Take some such $z_0 \in (0, \delta_0)$. Particularly, $z_0 < \frac{H}{d+p}$. Then, as $\Pi_X < 1$, we know that if M is sufficiently large, $(1 - \Pi_X)^{M-1} < z_0$. Then,

$$\begin{aligned} \frac{1 - \pi_X}{\pi_X} (d + p) (1 - z_0) &\leq \frac{1 - \pi_X}{\pi_X} (d + p) (1 - (1 - \pi_X)^{M-1}) \\ &\leq H - d - p \end{aligned}$$

$$\Rightarrow \frac{(d + p)(1 - z_0)}{H - (d + p)z_0} \leq \Pi_X.$$

Thus,

$$\frac{d + p}{H} \leq \frac{(d + p)(1 - z_0)}{H - (d + p)z_0} + \delta \leq \pi_X + \delta.$$

As $\text{Prob}(\text{voter takes bribe}) = 1 - \pi_X$, this gives the desired result.

□

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

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

Compared to [Proposition 6.1](#), 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. 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 8](#).

10. Experimental/stress test results

Finally, we mention a number of test $p + e$ attacks that we launched on the Kleros “Doges on Trial” pilot in August 2018 and September 2018. As an early implementation of the Kleros platform [\[17\]](#), this pilot was specifically intended to solicit attacks and provide a basic test of the incentives of the Schelling games underlying Kleros [\[36\]](#). This pilot maintained a curated list of “Doge” images (meme images of dogs, particularly Shiba Inus), hence the question posed to voters in Schelling games was, “is this image a Doge?” with options of “Yes” and “No.” This pilot used exclusively redistributive Schelling games, as in [Definition 2.3](#), with fixed choices of p and d . These values were roughly $p = \$4$ USD and $d = \$3.5$ USD at the time.

Moreover, a prize of 50 ETH (approximately \$9850 USD at the time) was offered to the first user who submitted an image of a cat that was accepted onto the list [36]. Consequently, a number of attacks were launched by various third parties.

We attached $p + \epsilon$ attacks to images of cats, hence providing voters with an unambiguous sense of what an “honest” vote was in this context. As the submission interface only allowed users to submit an image, we embedded a message to potential voters in the image, alerting them to the presence of a potential bribe and directing them to a forum post where the bribe was explained in further detail, see Figure 6. Moreover, these attacks were enforced with a smart contract; voter were provided with a link in the appropriate forum post directing them to where they could read these smart contracts, as deployed on the Ethereum blockchain, in advance of the vote.

Figure 6: Several images of cats were submitted to the Kleros “Doges on Trial” pilot, and in parallel with each image, a $p + \epsilon$ attack was launched to bribe participants to vote that that image instead depicts a “Doge”. Voters were selected via Kleros’ random selection mechanisms and then would see the image in a Kleros interface. The link embedded in the image directed these participants to more information about the attack including the smart contract enforcing it, as deployed on the Ethereum blockchain.

While p and d were fixed by the parameters of the underlying platform, we varied the values of ϵ . Moreover, 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. Specifically, we launched six attacks where bribes were guaranteed in the first two appeal rounds, but not afterward, of which the results are shown in [Figure 7](#). Additionally, we launched a single attack where the bribe was only guaranteed in the last round, whose results are shown in [Figure 8](#).

ϵ (approx. USD)	votes accepting bribe with multiplicity	votes rejecting bribe with multiplicity	unique voters accepting bribe	unique voters rejecting bribe
.02	0	9	0	5
.17	1	6	1	3
.87	1	4	1	3
1.7	2	4	2	4
5.2	3	3	2	2
17	1	7	1	6

Figure 7. Vote totals accepting or rejecting $p + \epsilon$ bribes at various values of ϵ when the bribe was guaranteed in non-terminal rounds.

ϵ (approx. USD)	votes accepting bribe with multiplicity	votes rejecting bribe with multiplicity	unique votes accepting bribe	unique votes rejecting bribe
5.2 (before last round)	30	69	2	7
5.2 (last round)	24	98	3	9

Figure 8: Vote totals accepting or rejecting $p + \epsilon$ 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.

In both cases, it was not known to participants in advance how many rounds there would be. Indeed, in one of the attacks of [Figure 7](#), a third-party community member triggered a third appeal round (in which the bribe was not offered and which is excluded from [Figure 7](#) for consistency). However, the attack of [Figure 7](#) was appealed

to the last possible round allowed by the smart contract of “Doges on Trial;” hence, voters in the last round knew that the bribe was guaranteed.

Participants were selected via the same random selection procedure that Kleros uses for all of its cases. On the Kleros platform, users, or “jurors”, are pseudo-anonymous. In order to select participants to vote in a given case, a random *token* is selected from a pool; then the user who possesses this token has the right to vote [17]. In particular, users that have large percentages of the total token pool may be drawn multiple times for a given case. The vote totals of Figure 7 and Figure 8 are given both with this multiplicity and by unique voters.

As a result of this relative concentration of votes to large token holders, it can be difficult to draw users who have small amounts of tokens for participation in experiments under this mechanism. Indeed, while our seven $p + \epsilon$ attacks resulted in 262 draws, each of which a priori could have selected a distinct individual for participation as a voter, only 17 distinct Ethereum addresses were drawn at least once.

However, as these 17 addresses controlled a large percentage of the token pool, we argue that our observations from this test nonetheless provide insight about the state of the platform at the time of the tests. Indeed, when considering the danger of these $p + \epsilon$ attacks, one is mostly concerned with whether 50% of the token holder pool from which the voters are drawn corresponds to people that would accept the bribe, as this determines whether the attack is likely to succeed or fail upon appeal and draws of larger panels of voters.

Then, it makes sense to take the model that a given token would vote one way or the other if presented with a given $p + \epsilon$ attack, namely that there is some percentage y of all tokens that would accept the bribe. As the voter selection mechanism’s random choices of tokens are independent in Kleros, the number of votes in a given round of that attack that would accept a given bribe should be distributed as $\text{Binomial}(\text{number of votes}, y)$. Hence, we can ask if the number of votes seen accepting the bribe can realistically occur by random chance under the null hypothesis that $y \geq .5$. Taking the last round of the “last round only attack” alone we see if $X \sim \text{Binomial}(122, .5)$, then $\text{Prob}(X \leq 24) < 0.000001$. Similarly, if we consider all of the guaranteed $p + \epsilon$ bribes together, we see if $X \sim \text{Binomial}(163, .5)$, then $\text{Prob}(X \leq 33) < 0.000001$.

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 necessarily 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 cognitive costs, and potentially moral costs to the degree to which participants feel a moral cost in accepting a “bribe” in a gameified pilot, as a barrier to $p + \epsilon$ attacks as in [Section 8](#).

Nevertheless, we have so far found no observable relationship between the percentage of votes accepting the bribe and the value of ϵ (*e.g.*, via logistic regression). Nor have we found a relationship between the percentage of votes accepting the bribe and whether the bribe was guaranteed versus not guaranteed in the last-round version of the attack. (In fact, a higher percentage of votes 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 questions might be further explored in future experiments that gather additional data.

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. However, this is exactly the appropriate context to test the resistance of a given platform to $p + \epsilon$ attacks. Hence, these tests can serve as a model for stress tests on cryptoeconomic platforms that are based on Schelling games.

Nevertheless, 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 $p + \epsilon$ attacks. As a result, we believe that further experiments regarding peoples’ behavior when faced with a $p + \epsilon$ attack in varying conditions should be conducted. Then one can periodically gauge a given platforms resistance to $p + \epsilon$ attacks by repeating this stress test and seeing if the population of token holders is willing to accept these bribes or not in conditions given by the typical parameters on that platform.

11. Conclusion

We have observed that redistributive Schelling games are somewhat more resistant to $p + \epsilon$ attacks than other payoff structures. Particularly, we have seen that under certain conditions there exists a viable counterattack, that is stable in equilibrium, against redistributive $p + \epsilon$ attacks on systems where appeal is possible. Moreover, even in

contexts where counterattacking against redistributive Schelling games is not practical, we have seen that the budget required for $p + e$ 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 $p + e$ attacks, and they likely require higher cognitive costs to understand. However, we see that 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 $p + e$ attacks on a real platform, and while these tests are preliminary and mostly provide insight to the specific context of that platform, we observed that these attacks did not succeed.

Acknowledgements

We wish to acknowledge the contribution of Daniel Babbe towards writing the attack smart contract used in the tests discussed in [Section 10](#). For the code of this contract see [\[37\]](#).

Appendix

A. Calculation of explicit lower bounds on counterattacking when attacker budget follows a truncated Pareto distribution

In this appendix we present the proof of [Corollary 7.7](#). Namely, we will see that when attacker budgets are distributed according to by (truncated) Pareto distributions, [Inequality 2](#) holds. Then we compute the explicit lower bound given by [Inequality 3](#) in this case.

Proof of [Corollary 7.7](#): We note that, by our assumptions on exponential growth, we can write $M_i = M_0 c_1^i$ for some $M_0 \in \mathbb{R}_{>0}$. Then the total costs to fund an attack from its launch up through the i th round is

$$\begin{aligned} \sum_{j=1}^i d_1 M_j^2 + d_2 M_j &= \sum_{j=1}^i d_1 M_0^2 c_1^{2j} + d_2 M_0 c_1^j \\ &= d'_1 c_1^{2i} + d'_2 c_1^i + d'_3 \end{aligned}$$

for d'_1, d'_2, d'_3 defined as follows

$$d'_1 = \frac{d_1 M_0^2 c_1^2}{c_1^2 - 1},$$

$$d'_2 = \frac{d_2 M_0 c_1}{c_1 - 1},$$

$$d'_3 = -d_1 M_0^2 \left(1 + \frac{1}{c_1^2 - 1}\right) - d_2 M_0 \left(1 + \frac{1}{c_1 - 1}\right).$$

Here, we are using the sum formulas for geometric series.

Then let $i, j \in \mathbb{N}$ such that $i \leq j$ and $i \leq l_{**} - 2$. By the definition of L_* ,

$$\text{Prob}(L_* > j + 1 | L_* > i + 1) \leq \text{Prob}(L_* > j | L_* > i)$$

is equivalent to

$$\begin{aligned} &\text{Prob}(B \geq d'_1 c_1^{2(j+1)} + d'_2 c_1^{j+1} + d'_3 | B \geq d'_1 c_1^{2(i+1)} + d'_2 c_1^{i+1} + d'_3) \\ &\leq \text{Prob}(B \geq d'_1 c_1^{2j} + d'_2 c_1^j + d'_3 | B \geq d'_1 c_1^{2i} + d'_2 c_1^i + d'_3). \end{aligned}$$

As f_B is a truncated Pareto distribution, we have

$$\text{Prob}(B \geq x) = \begin{cases} \frac{\alpha L^\alpha x^{-\alpha-1}}{1 - \left(\frac{L}{H}\right)^\alpha}, & \text{for } L \leq x \leq H \\ 0, & \text{for } x > H \end{cases},$$

for $0 < L < H$.

Let $t \in [0, 1]$. Then, we calculate

$$\text{Prob}(B \geq d'_1 c_1^{2(j+t)} + d'_2 c_1^{j+t} + d'_3 | B \geq d'_1 c_1^{2(i+t)} + d'_2 c_1^{i+t} + d'_3)$$

$$= \frac{\text{Prob}(B \geq d'_1 c_1^{2(j+t)} + d'_2 c_1^{j+t} + d'_3)}{\text{Prob}(B \geq d'_1 c_1^{2(i+t)} + d'_2 c_1^{i+t} + d'_3)}.$$

Note that the denominator here is non-zero as $i \leq l_{**} - 2$. If $d'_1 c_1^{2(j+1)} + d'_2 c_1^{j+1} + d'_3 > H$, then one already has that

$$\begin{aligned} 0 &= \text{Prob}(B \geq d'_1 c_1^{2(j+1)} + d'_2 c_1^{j+1} + d'_3 | B \geq d'_1 c_1^{2(i+1)} + d'_2 c_1^{i+1} + d'_3) \\ &\leq \text{Prob}(B \geq d'_1 c_1^{2j} + d'_2 c_1^j + d'_3 | B > d'_1 c_1^{2i} + d'_2 c_1^i + d'_3). \end{aligned}$$

Otherwise, we can assume that all of the relevant values for the budgets are between L and H .

Hence,

$$\frac{\text{Prob}(B \geq d'_1 c_1^{2(j+t)} + d'_2 c_1^{j+t} + d'_3)}{\text{Prob}(B \geq d'_1 c_1^{2(i+t)} + d'_2 c_1^{i+t} + d'_3)} = \frac{(d'_1 c_1^{2(j+t)} + d'_2 c_1^{j+t} + d'_3)^{-\alpha-1}}{(d'_1 c_1^{2(i+t)} + d'_2 c_1^{i+t} + d'_3)^{-\alpha-1}}.$$

However, a standard first derivative argument shows that, as $j \geq i$, this quantity is decreasing in t . Hence, we conclude $\text{Prob}(L_* > j+1 | L_* > i+1) \leq \text{Prob}(L_* > j | L_* > i)$, and that this situation satisfies the assumptions of Theorem 7.5. Then to compute the lower bound, for $l_0 \geq 3$, we note

$$\text{Prob}(L_* > l_0 | L_* > 2) = \left(\frac{d'_1 c_1^4 + d'_2 c_1^2 + d'_3}{d'_1 c_1^{2l_0} + d'_2 c_1^{l_0} + d'_3} \right)^{\alpha+1} = q_{l_0}^{\alpha+1},$$

where we have plugged in the explicit values for d'_1 , d'_2 , and d'_3 that we found above.

Now we consider the case where $d_1 = d + p$, $d_2 = \epsilon$. We note that, as the lower bound is monotonic in q_{l_0} , it is sufficient to show $q_{l_0} \leq q_{*,l_0}$. However, we have that

$$q_{l_0} =$$

$$\frac{\frac{M_1(d+p)}{c_1^2-1}c_1^4 + \epsilon\frac{1}{c_1-1}c_1^2 - \frac{M_1}{c_1^2}(d+p)\left(1 + \frac{1}{c_1^2-1}\right) - \epsilon\frac{1}{c_1}\left(1 + \frac{1}{c_1-1}\right)}{\frac{M_1(d+p)}{c_1^2-1}c_1^{2l_0} + \epsilon\frac{1}{c_1-1}c_1^{l_0} - \frac{M_1}{c_1^2}(d+p)\left(1 + \frac{1}{c_1^2-1}\right) - \epsilon\frac{1}{c_1}\left(1 + \frac{1}{c_1-1}\right)}.$$

After grouping terms, this is of the form $\frac{K_1+K_2\epsilon}{K_3+K_4\epsilon}$, where K_1, K_2, K_3, K_4 are independent of ϵ . Moreover, as $c_1 > 1$, we see that K_1, K_2, K_3, K_4 are all positive. Then, by a first derivative test, one sees that q_{l_0} is bounded by the maximum of the value it takes when $\epsilon = 0$ and its limit when $\epsilon \rightarrow \infty$.

□

Footnotes

1. This also corresponds to different models for how the resources required to fund the appeal are collected—they 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. The examples we will discuss in [Section 4](#) make a variety of different choices on this point. However, it will generally not be necessary to specify the exact model for how appeals are financed for this work. [↵](#)

2.

In “The $p + \epsilon$ Attack” (Buterin, 2015) the idea that “use[s] round $N + 1$ to determine who should be rewarded during round N ,” is considered.

Buterin, V. (2015). The $p + \epsilon$ Attack. *Ethereum Blog*. [↵](#)

3.

Specifically, in major proof-of-work chains such as Bitcoin (Nakamoto 2008) and the existing version of Ethereum (Wood 2014), no financial deposits in cryptocurrency are required of miners. One could impose the requirement that, in order to mine on a proof-of-work chain, a miner would be required to lock up an amount of cryptocurrency associated to a given public key, signing new blocks with their corresponding private key. Then to the degree that the miner commits actions that are detectable on-chain as malicious, such as signing invalid blocks, their deposit could be slashed. Nonetheless, while this type of structure is integral to proof-of-stake chains (Buterin 2019), it requires an additional layer of complexity when added on top of a proof-of-work structure.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*.

Buterin, V. (2019). Serenity design rationale. [↗](#)

4.

Compare this to the counter-coordination considered in [Buterin 2015] for simple Schelling games, which are seen to not provide equilibria, in pure or mixed strategies, except in limited situations corresponding to “tipping points”. Nor is it required in our setting that the victim party be willing and able to finance a second $p + \epsilon$ attack as in the example seen in [Abramowicz 2019].

Buterin, V. (2015). The $p + \epsilon$ Attack. *Ethereum Blog*.

Abramowicz, M. (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*. [↗](#)

5. 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 9](#). Note, however, that which round will be the last, definitive round is generally not known in advance, and that the attacker typically must launch a contract in advance that commits her to paying out bribes according to clear, programmable conditions. [↗](#)

6.

Where this assumption is taken, it is used to simplify the behavior that must be studied by allowing us to assume that various participants share the same payoff matrix. Often, the degree to which this assumption is a justifiable approximation is related to the degree to which systems are decentralized or not, allowing large miners or token holders to be selected to vote more often than others over relevant time frames. Future work might explore the degree to which this assumption can be reduced or eliminated. Indeed, heuristically, one might expect that an unequal distribution of votes might make $p + \epsilon$ attacks more difficult and/or expensive as fewer large actors would be necessary to successfully “counter-coordinate” (in the sense of [Buterin 2015]).

Buterin, V. (2015). The $p + \epsilon$ Attack. *Ethereum Blog*. [↵](#)

7. This includes knowledge of the various parameters: M, p, d, ϵ featured in the payoff tables of the different structures of Schelling game in [Definition 2.1](#), [Definition 2.2](#), and [Definition 2.3](#). To the degree that other parameters are also common knowledge to voters, such as partial information about the attacker’s budget, etc, we will note this in the assumptions of relevant results. [↵](#)

8.

Validators are also rewarded based on how many blocks it takes for their attestation to be included [Buterin 2019]. We assume that whether a validator accepts a $p + \epsilon$ bribe or not does not impact which block her attestation is included in. Furthermore, there is also a notion of “slashing conditions” [Buterin and Griffith 2017]. Note that as validators whose attestations violate a slashing conditions are penalized on both chains, and are hence penalized regardless of the outcome of the vote, the decision to violate a slashing condition or not does not have the structure of a Schelling game. Thus, we assume that the attacker does not provide an incentive to validators to violate the slashing conditions and consequently that no slashable actions occur during the period of the attack.

Buterin, V. (2019). Serenity design rationale.

Buterin, V., & Griffith, V. (2017). Casper the friendly finality gadget. *Computing Research Repository (CoRR)*, arxiv.org/abs/1710.09437v4. [↵](#)

9.

In the current Phase 0 release of Serenity, such a censorship attack could be used, for example, to delay finalization of certain rewards and penalties. In the future, such an attack on the beacon chain would also delay finalization of transactions occurring on the various shards whose deployment is planned for future phases of release [Buterin 2019]. Note, however, that an attacker may have additional mechanisms to try to delay the finalization of such a transaction, for example, by offering a $p + \epsilon$ attack on the votes for the “crosslinks” coming from the shard where the transaction originated.

Buterin, V. (2019). Serenity design rationale. [↗](#)

10.

Note that this censorship has a structure similar to the censorship attacks that we have considered on proof-of-work in that, even if the censorship attack succeeds in preventing $\mathcal{TX}$ from being in a finalized block by time $\mathcal{T}$, to the degree that $\mathcal{Y}$ does not finalize it would still be possible for $\mathcal{X}$ to finalize later, resulting in $\mathcal{TX}$ being seen as having occurred in epoch i_0 according to the internal clock of the beacon chain. However, in the setting of Serenity proof-of-stake, one could also consider variants on this censorship attack where the attacker attempts to finalize $\mathcal{Y}$ in exchange for paying bribes to voters in additional epochs. Such an attack could prevent $\mathcal{TX}$ from ever appearing in a finalized block in epoch i_0 . This could be used to block actions which have a deadline expressed in terms of (beacon chain) blocknumber, for example, challenge transactions to malicious state transitions in optimistic rollups [Konstantopoulos 2021] or votes in systems such as Kleros [Lesaegre et al. 2020].

Lesaegre, C., George, W., & Ast, F. (2020). Kleros: Long paper v1.0.0. [↗](#)

11.

We take a few approximations on the reward and penalty structure here, justifying our use of approximation symbols in [Figure 2](#). First we assume that all of the validators begin the attack period with equal weight corresponding to them all having deposits of 32 ETH. Furthermore, we do not take into account the discrete adjustments to effective balance [McDonald 2020], particularly in calculating inactivity leak penalties. Instead we use the continuous approximation given in [Edgington 2021].

Edgington, B. (2021). Ethereum 2.0 phase 0 - the beacon chain. [↗](#)

12. Our bound here is presumably an overestimate as it corresponds to all validators taking the bribe even as X finalizes; however, it allows for the possibility that the threshold for finalization will have changed due to the inactivity leak reducing the balances of participants who took bribes in previous rounds. [↵](#)

13. An attacker could offer bribes denominated in some other unit of value that is simply equal to the value of these tokens. However, this requires participants who are offered the bribe to take on some uncertainty about the price evolution of the two units of values, which in some cases could result in taking the bribe no longer being a dominant strategy. [↵](#)

Citations

1. Schelling, T. C. (1980). *The Strategy of Conflict*. Harvard University Press. [↵](#)
2. Huyck, van J., Battalio, R., & Beil, R. O. (1991). Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games. *The Quarterly Journal of Economics*, 106(3), 885–910. [↵](#)
3. Devetag, G., & Ortmann, A. (2007). When and Why? A Critical Survey on Coordination Failure in the Laboratory. *Experimental Economics*, 10(3), 331–344. [↵](#)
4. Cooper, R. (1999). *Coordination games*. Cambridge University Press. [↵](#)
5. Buterin, V. (2015). The $p + \epsilon$ attack. *Ethereum Blog*. [↵](#)
6. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. [↵](#)
7. Hardin, G. (1968). The tragedy of the commons. *Science*, 162(3859), 1243–1248. [↵](#)
8. Buffett, W. (2000). The billionaire’s buyout plan. *New York Times*. [↵](#)
9. Abramowicz, M. (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. [↵](#)
10. Deirmentzoglou, E., Papakyriakopoulos, G., & Patsakis, C. (2019). A survey on long-range attacks for proof of stake protocols. *IEEE Access*, PP. [↵](#)
11. Wang, K. (2019). Cryptoeconomics: Paving the future of blockchain technology. *Hackernoon*. [↵](#)

12. Sayeed, S., & Marco Gisbert, H. (2018). On the effectiveness of blockchain against cryptocurrency attacks. In *The twelfth international conference on mobile ubiquitous computing, systems, services and technologies (ubicomm 2018)* (pp. 9–14). International Academy, Research,; Industry Association. [↵](#)
13. Goldin, M. (2017). Token-curated registries 1.0. *Medium*. [↵](#)
14. Goldin, M., Soleimani, A., & Young, J. (2017). The AdChain registry. [↵](#)
15. Civil. (2017). The Civil cryptoeconomic whitepaper. *Medium*. [↵](#)
16. Sztorc, P. (2015). Truthcoin: Peer-to-peer oracle system and prediction marketplace. [↵](#)
17. Lesaege, C., George, W., & Ast, F. (2020). Kleros: Long paper v1.0.0. [↵](#)
18. Buterin, V. (2019). Serenity design rationale. [↵](#)
19. Buterin, V. (2018). Discouragement attacks. *Github Repository*. [↵](#)
20. Buterin, V. (2016). On settlement finality. *Ethereum Blog*. [↵](#)
21. Bitinfocharts. (2021). [↵](#)
22. Support, K. (2021). Cryptocurrency deposit processing times. [↵](#)
23. Edgington, B. (2021). Ethereum 2.0 phase 0 - the beacon chain. [↵](#)
24. Buterin, V., & Griffith, V. (2017). Casper the friendly finality gadget. *Computing Research Repository (CoRR)*, arxiv.org/abs/1710.09437v4. [↵](#)
25. Bitfly. (2021). Open source ethereum 2.0 beacon chain explorer. [↵](#)
26. Judmayer, A., Stifter, N., Zamyatin, A., Tsabary, I., Eyal, I., Gazi, P., ... Weippl, E. (2019). Pay-to-win: Incentive attacks on proof-of-work cryptocurrencies. *Cryptology EPrint Archive, Report 2019/775*. [↵](#)
27. McCorry, P., Hicks, A., & Meiklejohn, S. (2019). Smart contracts for bribing miners. In *22nd international conference on financial cryptography and data security (fc 2018 international workshops)* (pp. 3–18). Springer Berlin Heidelberg. [↵](#)
28. Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*. [↵](#)

29. Peterson, J., & Krug, J. (2018). Augur: A decentralized, open-source platform for prediction markets. [↵](#)
30. Klerosboard. (2021). [↵](#)
31. Rapoport, A., & Chammah, A. M. (1966). The game of chicken. *American Behavioral Scientist*, 10(3), 10–28. Retrieved from [↵](#)
32. Pareto, V. (1896). *Écrits sur la courbe de la répartition de la richesse*. Republished in *Oeuvres Complètes*, Droz., 1967. [↵](#)
33. Drăgulescu, A., & Yakovenko, V. (2001). Exponential and power-law probability distributions of wealth and income in the united kingdom and the united states. *Physica A: Statistical Mechanics and Its Applications*, 299, 213–221. [↵](#)
34. Aban, I. B., Meerschaert, M. M., & Panorska, A. K. (2006). Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473), 270–277. Retrieved from [↵](#)
35. Burguet, R., Ganuza, J.-J., & Montalvo, J. G. (2016). *The microeconomics of corruption: A review of thirty years of research*. [↵](#)
36. James, S. (2018). Doges on trial - the largest decentralized curated list ever. *Kleros Blog*. [↵](#)
37. Babbey, D. (2018). P-epsilon.sol. *Github Repository*. [↵](#)