Stake-Based Consensus for Utility Scoring

Francois Luus [*1] **Jacob Steeves** [*1] **Ala Shaabana** [1] **Yuqian Hu** [1] **Sin Tai Liu** [1]

Abstract

We formulate a stake-based consensus problem in terms of a two-player game and propose a protagonist consensus policy to optimize a Nash equilibrium via a weight reduction algorithm with a guarantee of minority stake deterioration. We generalize this to a two-team game and propose a smooth density evolution algorithm that outperforms coarser estimates. We perform a full-scale Monte Carlo analysis and confirm the accuracy of our theoretical results, and show the possibility of a 40% stake + 25% utility attack. The result is a variable-expense consensus algorithm that can be fit to blockchain compute constraints to reach accurate consensus in adversarial settings.

1. Stake-based weight consensus

1.1. Problem definition

We consider a two-player game between (protagonist) honest stake ($0.5 < s_H \leq 1$) and (adversarial) cabal stake ($1-s_H$), competing for total fixed reward $e_H+e_C = 1$, with honest emission e_H and cabal emission e_C, respectively, followed by stake updates $s'_H = \frac{s_H+e_H}{2}$ and $s'_C = \frac{1-s_H+e_C}{2}$. The honest objective $s_H \leq e_H$ at least retains scoring power s_H over all action transitions in the game, otherwise when $e_H \leq s_H$ honest emission will erode to 0 over time, despite a starting condition of $0.5 < s_H$.

We assume honest stake sets objectively correct weights w_H on itself, and $1 - w_H$ on the cabal, where honest weight w_H represents an ongoing expense of the honest player, sustained throughout the game. However, cabal stake has an action policy that freely sets weight w_C on itself, and $1 - w_C$ on the honest player, at no cost to the cabal player, with the objective to maximize the required honest self-weight expense w_H via

$$w_C^* = \arg\max_{w_C} E[w_H \mid s_H = e_H(s_H, w_H, w_C)].$$

We then assume the honest majority $s_H > 0.5$ can counter with a consensus policy π allowed to modify all weights modulo player labels, so it is purely based on the anonymous weight distribution itself, optimizing the Nash equilibrium

$$\min_{\pi} \max_{w_C} E[w_H \mid s_H = e_H(s_H, \pi(\mathbf{w}))].$$

The majority stake enforces an independent and anonymous consensus policy π (e.g. through a blockchain solution) that modifies the weights to minimize the expense w_H, which has been maximized by the cabal applying an objectively incorrect gratis self-weight w_C. Consensus aims to produce $\pi(\mathbf{w}) \to (w'_H, w'_C)$ so that $w'_C = 1 - w'_H$, by correcting the error $\epsilon = w'_C + w'_H - 1 > 0$. Note that the input cost w_H remains fully expensed, and that w'_H merely modifies the reward distribution that follows, but not knowing which players are honest or cabal (anonymous property).

1.2. Reward emission

In the two-player characterization of the game, there are two bimodal weight distributions of $(w_H,\ 1 - w_C)$ and $(1 - w_H,\ w_C)$ on the honest and cabal players, respectively. The stake proportions behind the bimodal distributions are $(b_{HH},\ b_{CH}) = (w_H s_H,\ (1 - w_C)(1 - s_H))$ and $(b_{HC},\ b_{CC}) = ((1 - w_H)s_H,\ w_C(1 - s_H))$, respectively.

$$b_{HH} = w_H s_H \qquad b_{CH} = (1 - w_C)(1 - s_H)$$
$$b_{HC} = (1 - w_H)s_H \qquad b_{CC} = w_C(1 - s_H)$$

Primary incentive i is the normalized sum of stake proportions, where honest rank $r_H = w_H s_H + (1 - w_C)(1 - s_H)$ and cabal rank $r_C = (1 - w_H)s_H + w_C(1 - s_H)$ are normalized to give $i_H = \frac{r_H}{r_H+r_C}$ and $i_C = \frac{r_C}{r_H+r_C}$. An additional reward d is the scoring share of incentive, i.e. $d_H = \frac{w_H s_H}{r_H} i_H + \frac{(1-w_H)s_H}{r_C} i_C$ and $d_C = \frac{(1-w_C)s_C}{r_H} i_H + \frac{w_C s_C}{r_C} i_C$. Finally, the complete reward emissions are $e_H = \frac{i_H+d_H}{2}$ and $e_C = \frac{i_C+d_C}{2}$, such that $e_H + e_C = 1$.

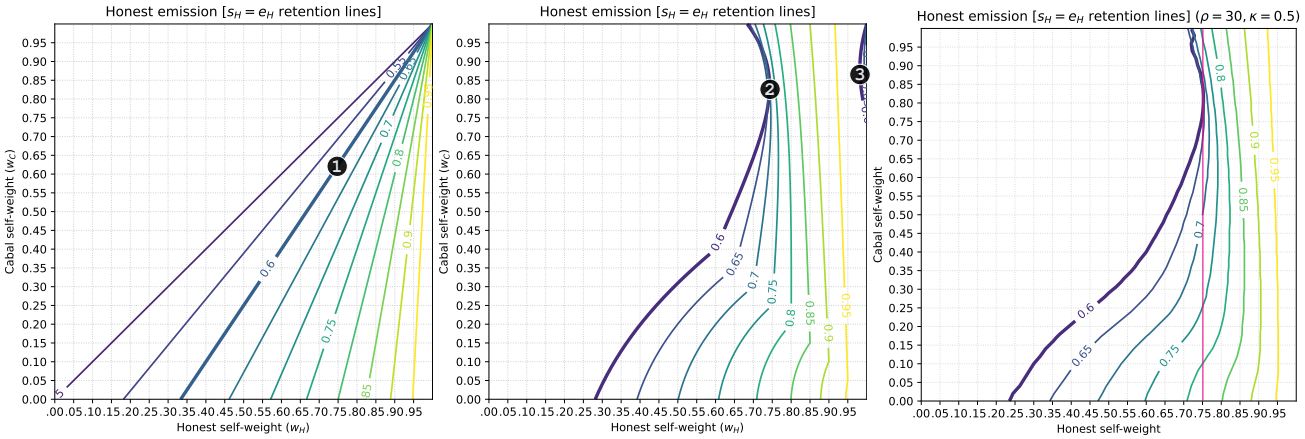
1.3. Consensus deviation

The weight consensus is the stake-proportion weight average $\overline{w_j} = \sum_i (s_i w_{ij}) w_{ij} \,/ \sum_k (s_k w_{kj})$, and accordingly the consensus weights for the honest and cabal players are

$$\overline{w_H} = \frac{s_H w_H^2 + (1 - s_H)(1 - w_C)^2}{s_H w_H + (1 - s_H)(1 - w_C)}, \text{ and}$$

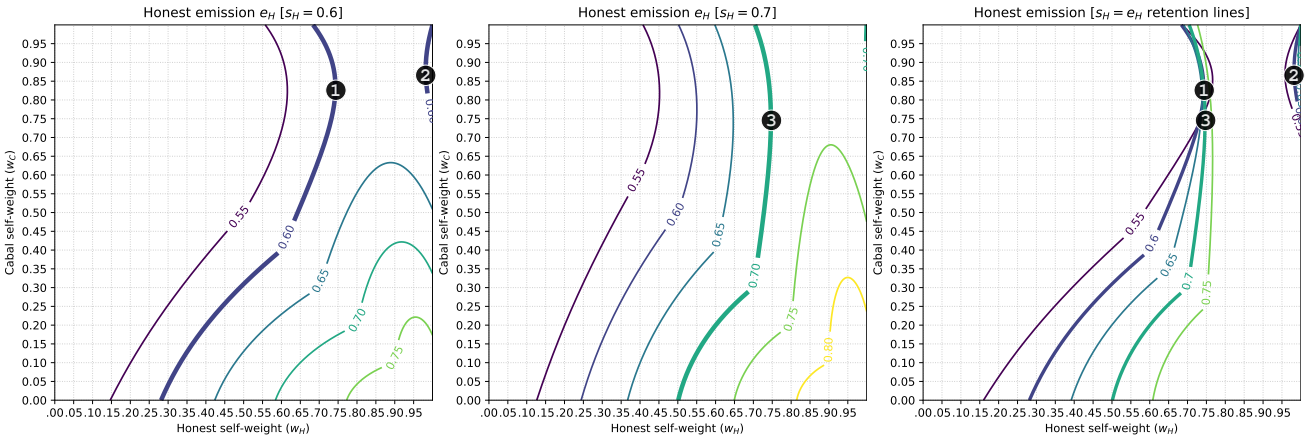
[*]Equal contribution [1]Opentensor Foundation. Correspondence to: Francois Luus <francois@opentensor.ai>.

Style courtesy of *International Conference on Machine Learning*.



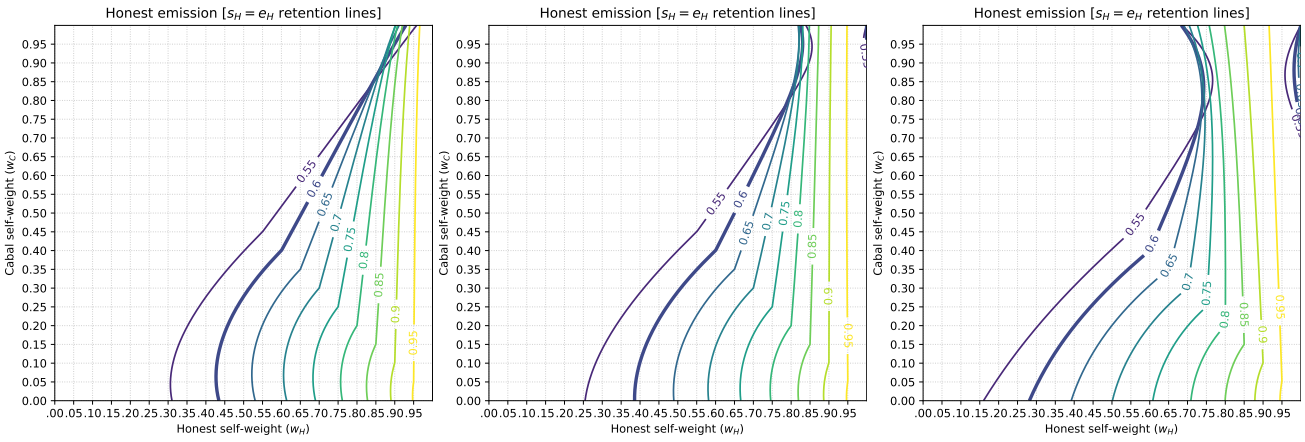
(a) Unprotected consensus. (b) Stake-based consensus policy. (c) Consensus policy with weight trust.

Figure 1. (a) **Selfish weighting problem**: Minority cabal sets $w_C = 1$ self-weight to maximally grow its relative stake, e.g. at ❶ honest majority stake of $s_H = 0.6$ and honest utility of $w_H = 0.75$ would require cabal to report self-weight $w_C < 0.62$ for honest stake to be retained. (b) **Consensus solution**: Stake-based consensus ($\eta = 3$) corrects excessive self-weight of minority stake, e.g. at ❷ $s_H = 0.6, w_H = 0.75$ no selfish cabal weight can prevent honest stake retention, even $w_C = 1$ results in honest stake ratio gain. **Zero-weight problem**: Minority cabal is virtually the only scoring incentive recipient of the cabal utility reward when its actual utility is near-zero, e.g. at ❸ where honest stake deteriorates. (c) **Weight trust solution**: Require the majority stake to agree that a weight is non-zero, otherwise smoothly nullify the associated reward to the degree of mistrust, which then removes the honest stake deterioration region when $w_H > 0.95$. **Consensus guarantee**: Honest majority stake is retained when $s_H \geq 0.6$ and $w_H \geq 0.75$, despite strategic cabal weight setting.



(a) Honest emission contours for $s_H = 0.6$. (b) Honest emission contours for $s_H = 0.7$. (c) Honest $s_H = e_H$ retention lines.

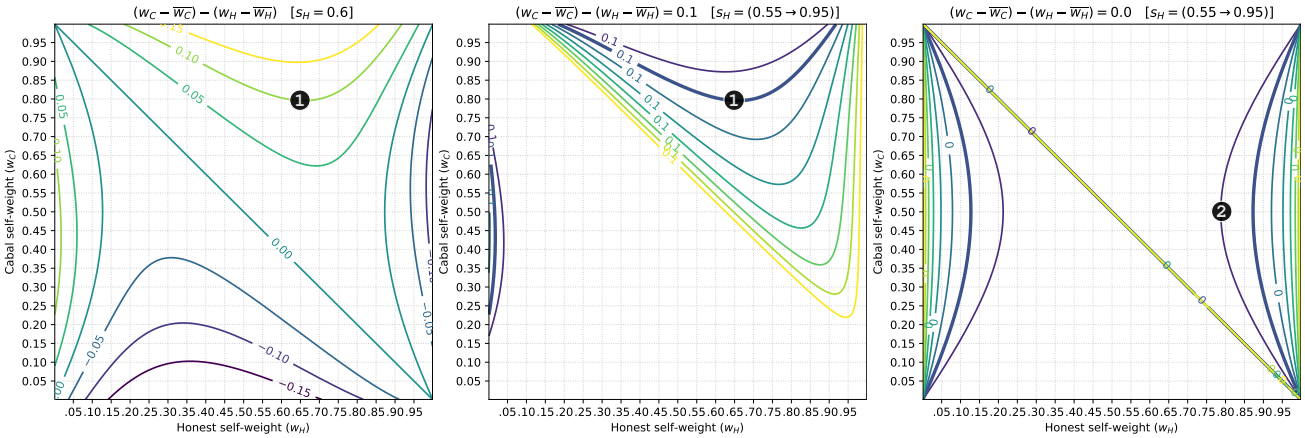
Figure 2. **Retention line interpretation**: (a) Honest incentive share contour plot for $s_H = 0.6$ only, highlighting where the emission is $e_H = 0.6$, e.g. at ❶. However, at ❷ the contour recedes again due to the *zero-weight problem*. (b) Similarly, the specific emission contour plot for $s_H = 0.7$, highlighting the contour where the emission is $e_H = 0.7$, which means with inflation the honest share ratio of $s_H = 0.7$ can be retained if honest utility is at least $w_H > 0.75$ like at ❸. (c) **Retention lines**: A compound plot combines all the highlighted $s_H = e_H$ contours from individual contour plots (e.g. $s_H = 0.6$ and $s_H = 0.7$), to show the overall retention profile. Generally, the higher the honest stake, the higher the honest utility requirement to retain stake proportion under adversarial weight setting.



(a) 0% scoring incentive. (b) 25% scoring incentive. (c) 50% scoring incentive.

(d) 75% scoring incentive. (e) 95% scoring incentive. (f) 100% scoring incentive.

Figure 3. **Scoring incentive**: A percentage share of the utility rewards equal to stake in a score, as incentive to encourage honest scoring. (a) No scoring incentive leads to extreme selfish weight setting, since the cabal does not share in the honest rewards. (b)-(f) Higher scoring incentive reduces selfishness evidenced by receding honest self-weight requirement for stake retention. However, the *zero-weight problem* at ❶ increases as well, since only the cabal can claim reward share from both honest and cabal subsets while honest sets zero weight on the cabal. The *weight trust solution* with smooth edge coverage $w_H > 0.95$ can only be extended so far before legitimate low-utility is also nullified, which practically limits scoring incentive around 50%.



(a) Excess difference ($s_H = 0.6$). (b) Fixed excess diff. ($s_H = 0.55 \to 0.95$). (c) Zero excess diff. ($s_H = 0.55 \to 0.95$).

Figure 4. **Larger minority excess**: The excess weight above consensus is larger for minority-stake when $w_H < w_C$. (a) Positive contours when $w_H < w_C$ at ❶ indicate regions of cabal error-correction potential. (b) Cabal error-correction region grows as majority stake increases $s_H = 0.55 \to 0.95$. (c) However at ❷, larger majority excess appears on the right-side when $w_C < w_H$ (to be avoided), which negatively impacts the majority weight more than the minority.

$$\overline{w_C} = \frac{s_H(1-w_H)^2 + (1-s_H)w_C^2}{s_H(1-w_H) + (1-s_H)w_C}, \text{ respectively.}$$

Under typical adversarial play with $1-w_H < w_C$, the upper modes $w_H > \overline{w_H}$ and $w_C > \overline{w_C}$ of the honest and cabal weight distributions, respectively, will exceed the consensus. Honest excess $\overline{w_H} < w_H$ is present when $1 - w_C < w_H$:

$$\begin{aligned}
\overline{w_H} &< w_H \\
\frac{s_H w_H^2 + (1-s_H)(1-w_C)^2}{s_H w_H + (1-s_H)(1-w_C)} &< w_H \\
s_H w_H^2 + (1-s_H)(1-w_C)^2 &< \\
s_H w_H^2 + (1-s_H)(1-w_C)w_H& \\
(1-s_H)(1-w_C)^2 &< (1-s_H)(1-w_C)w_H \\
1 - w_C &< w_H.
\end{aligned}$$

Similarly, $1 - w_H < w_C$ produces cabal excess $\overline{w_C} < w_C$:

$$\begin{aligned}
\overline{w_C} &< w_C \\
\frac{s_H(1-w_H)^2 + (1-s_H)w_C^2}{s_H(1-w_H) + (1-s_H)w_C} &< w_C \\
s_H(1-w_H)^2 + (1-s_H)w_C^2 &< \\
s_H(1-w_H)w_C + (1-s_H)w_C^2& \\
s_H(1-w_H)^2 &< s_H(1-w_H)w_C \\
1 - w_H &< w_C.
\end{aligned}$$

Lemma 1 (Larger Minority Excess). *Minority-stake excess weight is larger than majority-stake excess weight, i.e.* $w_C - \overline{w_C} > w_H - \overline{w_H}$*, when* $w_H < w_C$.

We use a symbolic solver (Wolfram) to show that the cabal excess is larger when $w_H < w_C$, i.e.

$$\frac{dw_C}{dw_H} = \frac{w_C - \overline{w_C}}{w_H - \overline{w_H}} > 1$$

The cabal excess is larger with majority honest stake when

$$\left\{\begin{matrix} 0 < w_C \le 0.5 \\ 0.5 < w_H \le 1 - w_C \\ 0.5 < s_H < 1 \end{matrix}\right\} \quad \text{or} \quad \left\{\begin{matrix} 0.5 < w_C \le 1 \\ 0.5 < w_H \le w_C \\ 0.5 < s_H < 1 \end{matrix}\right\}.$$

Otherwise, for the following conditions

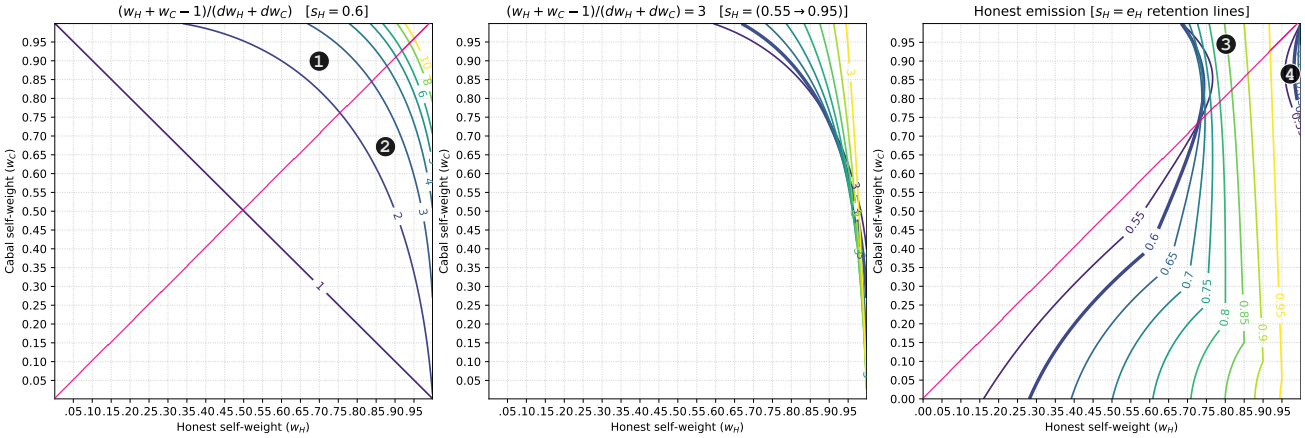
$$\left\{\begin{matrix} 0 < w_C \le 0.5 \\ 1 - w_C < w_H \end{matrix}\right\} \quad \text{or} \quad \left\{\begin{matrix} 0.5 < w_C \le 1 \\ w_C < w_H < 1 \end{matrix}\right\}$$

the honest stake criterion is

$$\begin{aligned}
&\frac{w_C^2 - w_C}{w_C^2 - w_C - w_H^2 + w_H} - \\
&\sqrt{\frac{w_C^2 w_H^2 - w_C^2 w_H - w_C w_H^2 + w_C w_H}{\left(w_C^2 - w_C - w_H^2 + w_H\right)^2}} < s_H < 1.
\end{aligned}$$

1.4. Excess weight reduction

Stake-proportional consensus advantages the honest player with $s_H > 0.5$, since it biases the consensus weight toward the honest vote and exposes the cabal excess self-weight $w_C > \overline{w_C}$ where $dw_C > dw_H$. Consequently, a consensus policy $\pi(\mathbf{w}) = \min(\overline{\mathbf{w}}, \mathbf{w})$ can reduce excess weight above the consensus $\overline{\mathbf{w}}$, where cabal weight should decrease more than honest weight. The weight reductions normally only happen in the upper modes w_H and w_C of the honest and cabal weights, respectively.



(a) Weight reduction steps η required. (b) $\eta = 3$ across honest stake range. (c) Honest stake retention for $\eta = 3$.

Figure 5. **Weight correction**: Iteratively recompute weight consensus and clip weight excess above new weight consensus for η iterations. (a) **Correction trade-off**: Choose η to maximize correction coverage ($s_H > 0.5$) for ❶ when $w_H < w_C$ (beneficial), while minimizing coverage for ❷ when $w_C < w_H$ (detrimental). $\eta = 2$ excessively exposes the detrimental region ❷ and wastes ❶ coverage when $s_H < 0.5$. Whereas $\eta = 3$ optimally trades off coverage of ❶ at expense of ❷ detriment. (b) As honest stake increases both correction regions ❶ and ❷ shrinks. (c) $\eta = 3$ error-correction protects against selfish weights in ❸, but exposes the *zero-weight* vulnerability at ❹.

The consensus policy $\pi(\mathbf{w}) \to (w'_H, w'_C)$ attempts to correct the error $\epsilon = w_H + w_C - 1$ so that

$$\begin{aligned} w'_C &= 1 - w'_H \\ w_C - \Delta w_C &= 1 - (w_H - \Delta w_H) \\ \Delta w_H + \Delta w_C &= w_H + w_C - 1 \\ \eta(dw_H + dw_C) &= w_H + w_C - 1 \\ \eta &= \frac{w_H + w_C - 1}{dw_H + dw_C}. \end{aligned}$$

The approximate number of weight reduction steps is η, and the consensus policy is thus converted to an iterated function $\pi = f^\eta$, where the function is repeated η times $f^3(\mathbf{w}) = f(f(f(\mathbf{w})))$. Note that $f(\mathbf{w}) = \min(\overline{\mathbf{w}}, \mathbf{w})$ recomputes the consensus weight $\overline{\mathbf{w}}$ each time.

We compute $\eta = \frac{w_H + w_C - 1}{dw_H + dw_C}$ and compare with the correction factor dw_C/dw_H to identify the optimal η avoiding over-correction in the detrimental $w_C < w_H$ region where $dw_C/dw_H < 1$.

We observe that higher $\eta > 3$ values extend the correction further into the detrimental $w_C < w_H$ region where $dw_C - dw_H < 0$, hence an optimal $\eta \approx 3$ is identified, which is large enough to provide sufficient correction when $w_H < w_C$.

Application of the consensus policy $\pi(\mathbf{w}) = f^{\eta \approx 3}(\mathbf{w})$ can partially correct the error $\epsilon = w'_C + w'_H - 1 > 0$, in particular the previous expense $w_H = 1$ is reduced to $w_H < 0.75$ for $s_H = 0.6$, even at Nash equilibrium with $w^*_C = 0.8$. Importantly, the consensus policy $\pi(\mathbf{w})$ operates on anonymized weights and do not assume the player identities, thus behaves impartially in terms of a stake-based consensus.

1.5. Smoothed weight reduction

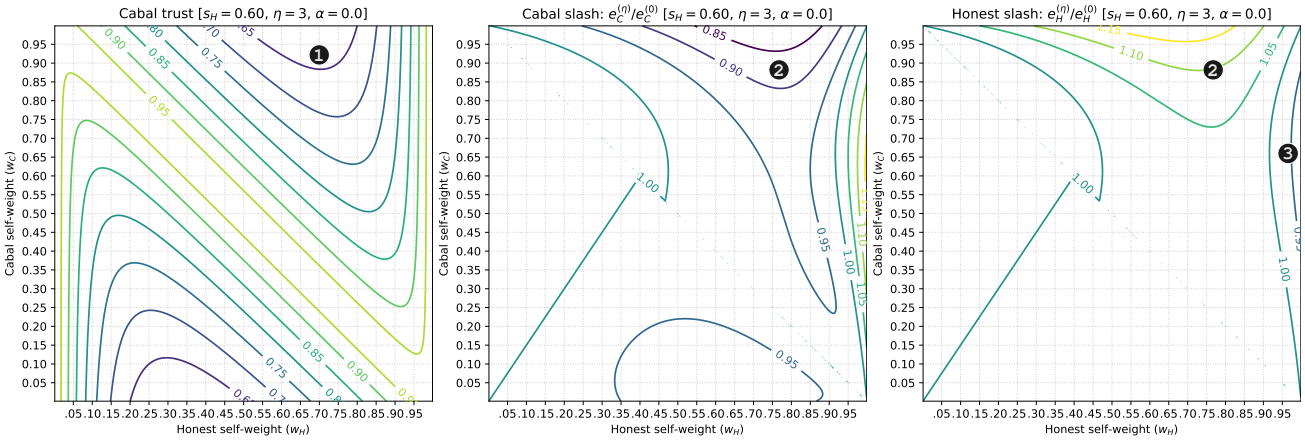
The correction function $f(\mathbf{w}) = \min(\overline{\mathbf{w}}, \mathbf{w})$ should be smoothed to ensure $\lim_{dw \to 0} f^\eta(w + dw) - f^\eta(w) < \varepsilon$ where adjacent weights are corrected to a similar degree. The correction factor should also depend on the magnitude of deviation from consensus, in terms of a standard deviation σ. We opt for a stake-weighted mean absolute deviation, since it does not make the normal assumption as strongly as mean square deviation, as follows

$$\sigma(\mathbf{w}) = \frac{\sum_i s_i w_{ij} |w_{ij} - \overline{\mathbf{w}}|}{\sum_k s_k w_{kj}}.$$

The standard correction $0 \leq \alpha < 1$ fully applies when $w - \overline{\mathbf{w}} = \sigma(\mathbf{w})$, and amplifies when $w - \overline{\mathbf{w}} > \sigma(\mathbf{w})$ to a maximum correction at $\overline{\mathbf{w}}$ with a proposed smoothed function iterate

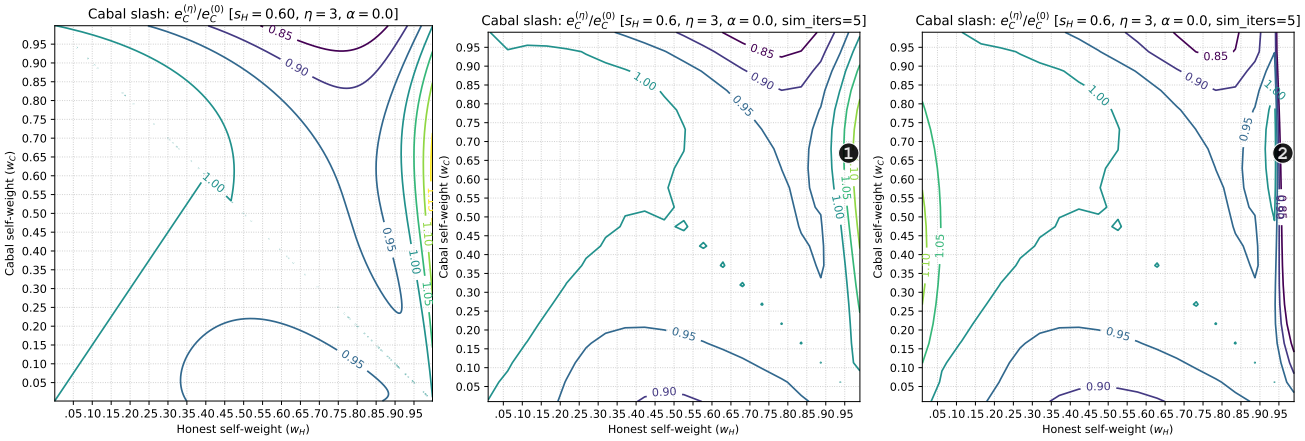
$$\begin{aligned} f(\mathbf{w} \mid w > \overline{\mathbf{w}}) &= \overline{\mathbf{w}} + (w - \overline{\mathbf{w}})\alpha^{\frac{w - \overline{\mathbf{w}}}{\sigma(\mathbf{w})}} \\ &= \overline{\mathbf{w}} + (w - \overline{\mathbf{w}})(1 - \delta) \\ &= \overline{\mathbf{w}} + w - \overline{\mathbf{w}} - \delta(w - \overline{\mathbf{w}}) \\ &= w - \delta(w - \overline{\mathbf{w}}) \\ &= w - \delta \cdot dw. \end{aligned}$$

The smoothed function iterate now requires more steps, where a larger $\alpha \approx 1 - \delta$ results in a larger η, such that



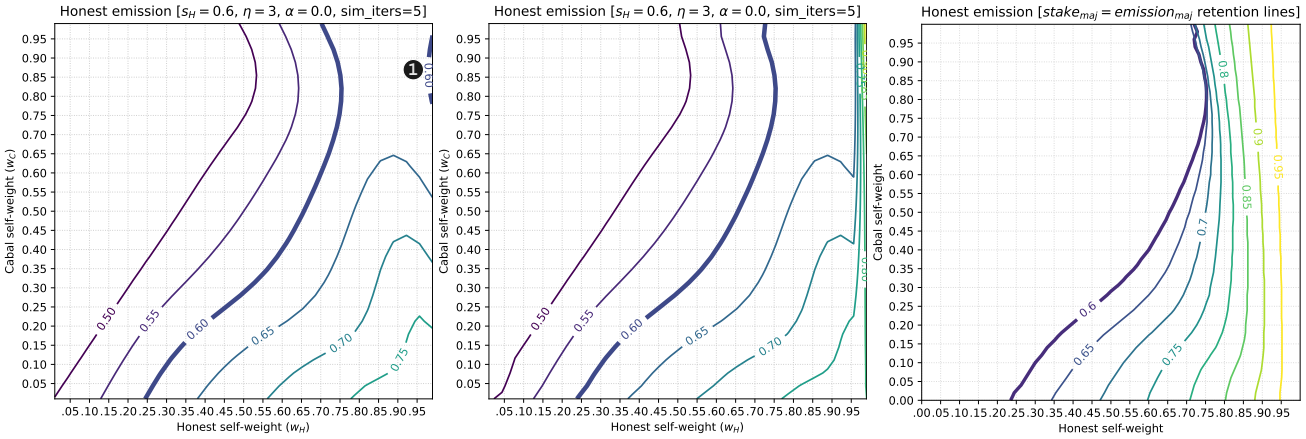
(a) Cabal trust (remaining scoring weight). (b) Cabal slash (emission after/before). (c) Honest slash (emission after/before).

Figure 6. **Emission slashing**: Iterative weight correction reduces effective scoring weight and incentive share. (a) Initial scoring weight is always 1, but weight correction reduces this whenever $\sum w \neq 1$, with the largest reduction seen at ❶. (b) Emission is the average of scoring incentive and utility reward and cabal emission is slashed, i.e. $e^{(\eta=3)} < e^{(\eta=0)}$ particularly in the $w_H < w_C$ and $0.5 < s_H$ region around ❷. (c) Consequently, honest emission is boosted in region ❷, but the *zero weight* vulnerability at ❸ slashes honest emission, although comparatively little.



(a) Two-player result. (b) Two-team (Monte Carlo analysis). (c) Two-team (with weight trust).

Figure 7. **Cabal slash** (two-team). (a) Statistical analysis simplifies the two-player result. (b) The two-team generalization fully enacts the weight distribution, and a Monte Carlo analysis reveals the worst-case cabal slash closely following the two-player result. (c) Cabal incentive is boosted at the *zero weight* vulnerability of ❶, but the *weight trust* solution ensures active cabal slash at ❷.



(a) Consensus policy. (b) Consensus policy with weight trust. (c) Honest retention lines.

Figure 8. **Weight trust**: Majority stake should set non-zero weight to a player, otherwise its reward is nullified. (a) Two-team honest emission with just the consensus policy, with *zero weight* vulnerability at ❶. (b) Applying weight trust protects the $0.95 < w_H$ region and removes the exploit. (c) Honest retention is now monotonically possible at increasing honest self-weight.

$\eta\delta \approx 3$, according to

$$\begin{aligned} w'_C &= 1 - w'_H \\ w_C - \Delta w_C &= 1 - (w_H - \Delta w_H) \\ \Delta w_H + \Delta w_C &= w_H + w_C - 1 \\ \eta\delta(dw_H + dw_C) &= w_H + w_C - 1 \\ \eta &= \frac{w_H + w_C - 1}{\delta(dw_H + dw_C)}. \end{aligned}$$

We compare the previous retention graph $(\eta = 3, \alpha = 0)$ against $(\eta = 3/(1-0.95), \alpha = 0.95)$ and observe a reduced cost $w_H = 0.7 < 0.75$ with the smoothed iterate with $\alpha > 0$.

We ensure monotonicity of the consensus policy by choosing the minimum $\varepsilon > 0$ added to the deviation $\sigma(\mathbf{w})$.

$$\begin{aligned} f(\mathbf{w} \mid w > \overline{\mathbf{w}}) &< f(\mathbf{w} + dw \mid w > \overline{\mathbf{w}}) \\ \overline{\mathbf{w}} + (w - \overline{\mathbf{w}})\alpha^{\frac{w-\overline{\mathbf{w}}}{\sigma(\mathbf{w})+\varepsilon}} &< \overline{\mathbf{w}} + (w + dw - \overline{\mathbf{w}})\alpha^{\frac{w+dw-\overline{\mathbf{w}}}{\sigma(\mathbf{w})+\varepsilon}} \\ (w - \overline{\mathbf{w}})\alpha^{\frac{w-\overline{\mathbf{w}}}{\sigma(\mathbf{w})+\varepsilon}} &< (w + dw - \overline{\mathbf{w}})\alpha^{\frac{w+dw-\overline{\mathbf{w}}}{\sigma(\mathbf{w})+\varepsilon}} \\ \frac{w - \overline{\mathbf{w}}}{w + dw - \overline{\mathbf{w}}} &< \alpha^{\frac{w+dw-\overline{\mathbf{w}}}{\sigma(\mathbf{w})+\varepsilon} - \frac{w-\overline{\mathbf{w}}}{\sigma(\mathbf{w})+\varepsilon}} \\ \log \frac{w - \overline{\mathbf{w}}}{w + dw - \overline{\mathbf{w}}} &< \log \alpha^{\frac{dw}{\sigma(\mathbf{w})+\varepsilon}} \\ \frac{dw \log \alpha}{\log \frac{w-\overline{\mathbf{w}}}{w+dw-\overline{\mathbf{w}}}} - \sigma(\mathbf{w}) &< \varepsilon \end{aligned}$$

1.6. Weight trust

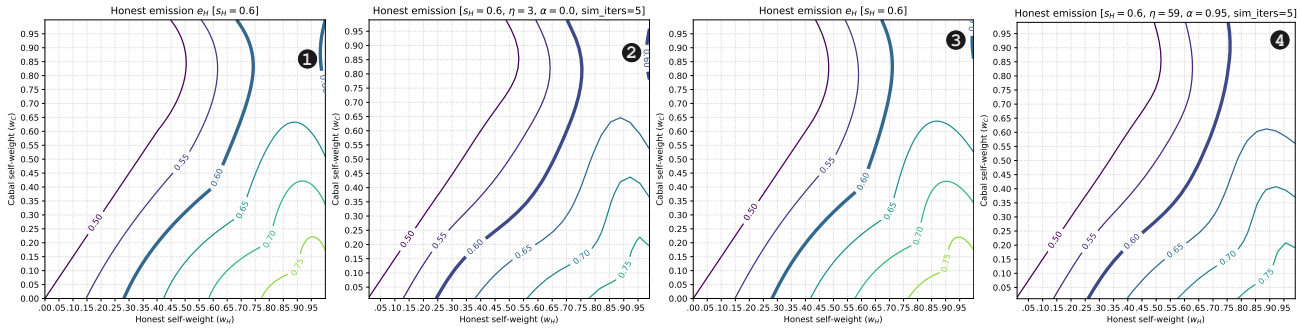
Weight trust $T = (W > 0)S$ is the sum of stake assigning a non-zero weight to a player, and a consensus $C = (1 + \exp(-\rho(T-\kappa)))^{-1}$ provides a smooth threshold at κ where exceeding κ ratio of stake quickly allows for high trust. A modified rank $r' = rc$ multiplies rank with the weight trust consensus, which influences the emission so that zero cabal weight $w'_C = 1 - w_H \approx 0$ receives low consensus thereby penalizing cabal emissions.

The vulnerable region of $w_H = 1$ and $0.8 < w_C < 0.95$ allows for cabal stake gain when $s_H = 0.6$, but the weight trust consensus smoothly pads the region around $w_H = 1$ and removes the vulnerability. The cabal can thus not claim reward when the honest majority deems cabal utility to be zero, despite the non-zero self-weight reported by the minority cabal.

2. Density generalization

2.1. Overview

We generalize the two-player game to a two-team game with $|H|$ honest and $|C|$ cabal players, that have $\sum_{i\in H} s_i = s_H$ honest stake and $\sum_{i\in C} s_i = 1 - s_H$ cabal stake. Honest players $i \in H$ set $\sum_{j\in H} w_{ij} = w_H$ self-weight and $\sum_{j\in C} w_{ij} = 1 - w_H$ weight on cabal players, while cabal players $i \in C$ set $\sum_{j\in C} w_{ij} = w_C$ self-weight and $\sum_{j\in H} w_{ij} = 1 - w_C$ weight on honest players. The rank components result in the same aggregates in the two-player



(a) Two-player ($\eta = 3$). (b) Two-team ($\eta = 3$). (c) Two-player ($\eta = 59$). (d) Two-team ($\eta = 59$).

Figure 9. **Evolution smoothness**: Evolving through more smaller steps with $\eta = 59$ reduces the *zero weight* exploit at ❸, compared to $\eta = 3$ at ❶, since fine-grained correction steps more accurately track changes in consensus. The two-team honest emissions (Monte Carlo worst-case analysis) tend to further reduce the *zero weight* exploits at ❷, ❹ compared to the two-player case at ❶, ❸.

game.

$$
\begin{aligned}
b_{HH} &= \sum_{j\in H} w_{ij} \sum_{i\in H} s_i = w_H s_H \\
b_{CH} &= \sum_{j\in H} w_{ij} \sum_{i\in C} s_i = (1-w_C)(1-s_H) \\
b_{HC} &= \sum_{j\in C} w_{ij} \sum_{i\in H} s_i = (1-w_H) s_H \\
b_{CC} &= \sum_{j\in C} w_{ij} \sum_{i\in C} s_i = w_C (1-s_H)
\end{aligned}
$$

In particular, the weight consensus of an individual honest player is shown to be $\overline{w_h} = \frac{\overline{w_H}}{|H|}$ as follows (and similarly for cabal players $\overline{w_c} = \frac{\overline{w_C}}{|C|}$)

$$
\begin{aligned}
\overline{w_h} &= \sum_i s_i w_{ih}^2 \Big/ \sum_k s_k w_{kh} \\
&= \frac{\sum_{i\in H} s_i w_{ih}^2 + \sum_{i\in C} s_i w_{ih}^2}{\sum_{k\in H} s_k w_{kh} + \sum_{k\in C} s_k w_{kh}} \\
&\approx \frac{\sum_{i\in H} s_i w_H^2/|H|^2 + \sum_{i\in C} s_i (1-w_C)^2/|H|^2}{\sum_{k\in H} s_k w_H/|H| + \sum_{k\in C} s_k (1-w_C)/|H|} \\
&= \frac{1}{|H|} \frac{s_H w_H^2 + (1-s_H)(1-w_C)^2}{s_H w_H + (1-s_H)(1-w_C)} = \frac{\overline{w_H}}{|H|}
\end{aligned}
$$

Under the minimal assumption that the average self-weights set on honest and cabal players are $\frac{w_H}{|H|}$ and $\frac{w_C}{|C|}$ we can construct weight densities $p_h(w) = p_{hh}(w) + p_{ch}(w)$ and $p_c(w) = p_{hc}(w) + p_{cc}(w)$, here according to the normal assumption (other densities with a similar first moment could possibly also be valid)

$$
\begin{aligned}
p_{hh}(w) &= s_H w \mathcal{N}\left(\frac{w_H}{|H|}, \frac{w_H}{|H|}\sigma\right) \\
p_{ch}(w) &= (1-s_H) w \mathcal{N}\left(\frac{1-w_C}{|H|}, \frac{1-w_C}{|H|}\sigma\right) \\
p_{hc}(w) &= s_H w \mathcal{N}\left(\frac{1-w_H}{|C|}, \frac{1-w_H}{|C|}\sigma\right) \\
p_{cc}(w) &= (1-s_H) w \mathcal{N}\left(\frac{w_C}{|C|}, \frac{w_C}{|C|}\sigma\right).
\end{aligned}
$$

The consensus and mean absolute deviations of a weight density function $p(w)$ are

$$
\overline{p} = \int w p(w) dw, \quad \text{and} \quad \sigma(p) = \int |w - \overline{p}| p(w) dw.
$$

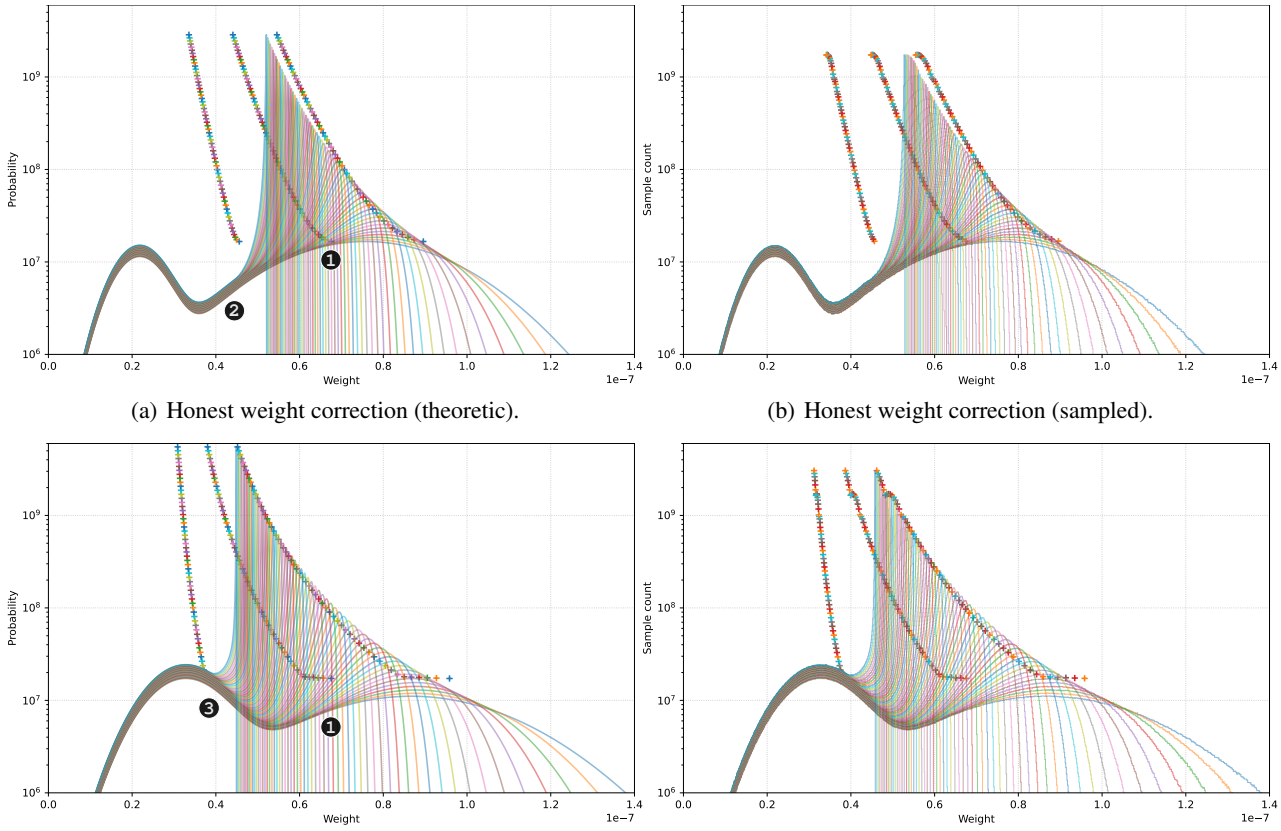
We overload the iterated function f as a density evolution function $f(p(w))$ that contracts a density $p(w)$ above consensus $\overline{p}$ by a nominal degree of α at a single deviation $\frac{w-\overline{p}}{\sigma(p)}$, in order to correct the error $\epsilon = w_H + w_C - 1$. The density is contracted via $g(w) = f^{-1}(w)$ involving the original iterated function f.

$$
f(p(w)) = p(w \mid w \le \overline{p}) + p(g(w) \mid \overline{p} < w) \frac{w}{g(w)} \frac{dg(w)}{dw}
$$

The final rank after applying the consensus policy $\pi = f^\eta$ is $r_i = \int f^\eta(p_i(w)) dw$, where a single function iteration contracts the consensus by $\overline{p}' - \overline{p}$, which is equal to

$$
\int w p(w) dw - \int w p(g(w) \mid \mu_p < w) \frac{w}{g(w)} \frac{dg(w)}{dw} dw.
$$

Simulating the consensus policy $(\eta = 3/(1-0.95), \alpha = 0.95)$ on weight densities set on honest and cabal players



(a) Honest weight correction (theoretic).

(b) Honest weight correction (sampled).

(c) Cabal weight correction (theoretic).

(d) Cabal weight correction (sampled).

Figure 10. **Density evolution**: Weight correction through density evolution reduces cabal weight consensus more than honest reduction ($s_H = 0.6, w_H = 0.7, w_C = 0.8$). The honest weights and cabal weights have an equal starting consensus at ❶, but through density evolution the cabal reduces more to ❸, versus a higher end honest consensus at ❷. Density evolution thus succeeds in penalizing the minority cabal and allows for honest stake retention. The theoretical probability densities in (a), (c) closely match the stochastically sampled results in (b), (d). The crosshair markers indicate the consensus flanked by a standard deviation above and below.

where $s_H = 0.6, w_H = 0.7, w_C = 0.8$, we see an equal starting consensus weight reduce further for the cabal players $(6.76 \rightarrow 3.8)$ vs honest players $(6.76 \rightarrow 4.41)$. The consensus policy acts as an upper-mode resiliency test where cabal self-weight with minority stake fails comparatively to honest self-weight with majority stake.

2.2. Stochastic sampling

We move from theoretical density analysis to a stochastic sampling analysis, where the original $\pi(\mathbf{w}) = f^{\eta}(\mathbf{w})$ can be applied directly to a weight sample $\mathbf{w}$ for a player, gradually contracting excess weight toward the consensus until an optimal contraction volume is reached. We observe very similar density evolution results as with the theoretical density analysis.

2.3. Two-team game

We perform a worst-case Monte Carlo analysis of a full-scale two-team game by sampling from normal densities, primarily to confirm the accuracy of the preceding aggregate analysis. We run a number of Monte Carlo iterations and record the worse-case results. A blockchain-based con-

sensus algorithm has space and compute limitations, which would favor a smaller η number of density evolution operations, each of which requires $O(n^2)$ operations. A small $\eta = 3$ with $\alpha = 0$ produce a full-scale result very close to the aggregate result.

Increasing the number of density evolution steps to $\eta = 59$ with $\alpha = 0.95$ manages to remove the zero-utility exploit at $w_H > 0.98$ seen at $\eta = 3$. However, the aggregate result in the theoretical honest retention deviates slightly, likely due to deviation of the upper mode density below consensus not accounted for in the aggregate.