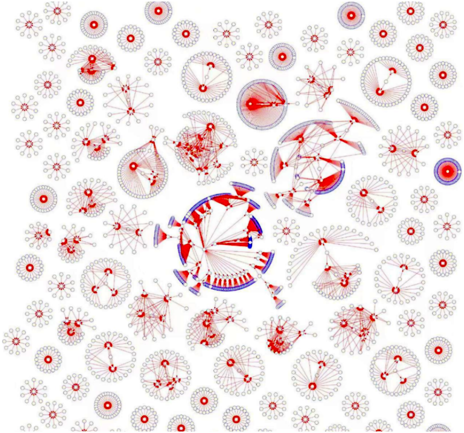
Sybil-proofness in Reputation-based Staking

Julien Prat, Yiyun Zheng

September 21, 2023

Initiating a network platform: in a nutshell

- Any power to the malicious parties can multiply in a large-scale peer-to-peer network, through self-replicating.
- These **Sybils**, when not distinguished from other genuine users, can benefit from an unfair weight and volume of voice to steer the big ship.



[a]

[a]image from Connext community report

Initiating a network platform: in a nutshell

- Who do we welcome when a new identity on the platform is just one click away?

Political voting

- **US election voting**:
- Real-name participation.
- Using an approved proof of identity.
- Cross-state voting is achievable due to poor communication between states,
- double-voting can have repercussions as being trialled for a felony in most states.

Political voting

- **Low Sybil concern.**
- **Highly centralised solution.**

1. Individuals have a non-pivotal position
2. costly on faking identities: certifying authority tends to have good security,
3. risky due to uncertain vote revision causing a possible sentence.

Social media

- **X premium**(previously, the Twitter blue):

  "An opt-in, paid monthly subscription that adds a blue checkmark to your account and offers early access to select new features, like Edit post." Verification requires a valid, active phone number.

A typical Web 2.0 approach that introduces some cash economics through pricing, not as restrictive as a fully-trusted certification that relies on a centralized authority.

Social media

- **Lower Sybil concern.**
- **Less centralised solution.**

1. Users give away credentials
2. Not too susceptible to the Sybil attack
3. Unlike the functionality of voting, being the majority does not possess all the advantages

Bio-metric proof of humanity

- **Worldcoin by Sam Altman**: Users are paid in their coins called 'WLD' if they proceed with a bio-metric authentication, i.e., to have their iris scanned.
- It does not rely on a central authority but the bio-metric aspect with a zero-knowledge promise failed to deliver to the audience.
- "The biggest issue is that little of what Worldcoin has done inspires confidence or trust, which should be the cornerstone of what such a company does."

High risk environment: Token airdrops

A short overview:

- **Uniswap**: 15% of the tokens were distributed among 1/4m users with proof of a one-time usage, ownership is earned by contributing liquidity and activity.

- **Ampleforth, Tornado.Cash etc.**: quickly followed suits with an airdrop value easily exceeding $5000.

- **Ethereum Name Service (ENS)** airdropped governance token when converted to a DAO. One of the largest airdrops with tokens valued between $250m and $500m.

High risk environment: Token airdrops

Users are getting the wrong education to start being strategic by adopting Sybils:

Airdrops Mining

Existing solutions:

- **Transaction patterns analysis**: reminiscent of the (almost) Sybil-proofness property of Bitcoin, getting a free lunch from airdrops will cost time and resources to bypass algorithmic screenings.
- **Witch-hunt program**: device a new reward to provoke internal auditing using the recovered tokens.

High risk environment: Token airdrops

Users are getting the wrong education to start being strategic by adopting Sybils:

Airdrops Mining

Existing solutions:

- **Transaction patterns analysis**: reminiscent of the (almost) Sybil-proofness property of Bitcoin, getting a free lunch from airdrops will cost time and resources to bypass algorithmic screenings. $\Rightarrow$ Not reliable. High interaction requirements may result in more active Sybils than more active authentic players.
- **Witch-hunt program**: device a new reward to provoke internal auditing using the recovered tokens. $\Rightarrow$ 'Hunters' need to identify themselves to redeem the prize, defying the decentralisation purpose. Hurting real users.

To summaries: a process of elimination

- Sybil attacks tend to be more prevalent in distributed systems.
- Centralised and/or privacy-divulging solutions are not favoured.
- No-cost participation with *patchworks* later is also not the best way to go at it.
  1. trust networks are prone to manipulation
  2. difficult to set a new mechanism to reward policing behaviours
- **We propose the use of staking to “monetise” participation.**

Core goal

- **Elicit an initial staking from participants who**:
- can but will not walk away from the platform easily, (induced commitment)
- willing to work (tolerates disutility) to maintain the identity,
- has a low budget. (decentralisation)

Referenced literature

- Contract design in the endogenous incomplete market: Endogenised by the one-sided commitment from the platform.
- Worker moral hazard, "backloading" payoff: Lazear(1981), Harris, Hölmstrom(1982)
- Reflecting reputation using payoff: Thomas, Worrall(1988)
- Efficient risk sharing without Commitment: Kochelakota(1996)
- Recursive insurance contract Ljungqvist, Sargent (2004)

Related literature

- The reputation is individually assigned and managed.
- Without concerning network effects and reputation from ranking.

- In another stream of literature, EigenTrust (Kamvar et al.), Sybil behaviours arise because sub-graph creation yields a reputation boost. (similar to the airdrop mining that we observe today.)
- Sybil-proofness then hinges on an Asymmetric reputation resembling a federated environment: trusted and identified nodes. (Cheng, Friedman)

Set-up

- Discrete time $t = 0, 1, 2, \ldots$.
- Subject to an agent with an initial staking value of v_0 at time 0.
- Our control instrument is a pair of promises:
  - contract value, v_t
  - cash-out value, c_t
- The reputation is explicitly represented by the contract value.
-
$$c_{t+1} \leq v_{t+1}, \ \forall t \geq 1. \tag{LE}$$
- **Lock-in effect(LE)**: Always more beneficial to enjoy the contract rather than cashing out.

Primary constraints

- If we only focus on induced commitment.
- Action space of the agent : {walk away and collect c_t, stay}.
  - leave with c_t collectable.
  - stay and receive new sets of (v_{t+1}, c_{t+1}).
-
$$\beta\left[(1-\delta)v_{t+1} + \delta c_{t+1}\right] \geq v_t, \quad \text{(PK)}$$
  - β: discount factor of the agent.
- **Promise-keeping(PK)**: The new package always has a value exceeding the current contract value (reputation). Increasing sequence of v, because the longer they commit the more they are known to be "loyal".

Our problem

- We explore the optimal design for updating the rewards, facing a trade-off between Sybil-proofness and decentralisation.

Our problem

- Platform's problem: updating the reward over time to retain the agent.

$$P(v_t) = \max_{\{c_{t+1}, v_{t+1}\}} \rho \Big[(1-\delta) P(v_{t+1}) - \delta c_{t+1} \Big]. \quad (1)$$

  - ρ: discount factor of the platform.
  - δ: probability of the agent's heterogeneous liquidity shock, forcing a leave.
- Impatient platform 'back-loads' promises.
- $\rho < \beta \Rightarrow$ The objective function is strictly decreasing and concave.

Taking FOCs

We take the Lagrangian of the platform's problem:

$$\mathcal{L} = \rho\Big[(1-\delta)P(v_{t+1}) - \delta c_{t+1}\Big] + \lambda\Big\{\beta\Big[(1-\delta)v_{t+1} + \delta c_{t+1}\Big] - v_t\Big\} + \mu(v_{t+1} - c_{t+1}). \quad (2)$$

Then take the First Order Conditions on v_{t+1} and c_{t+1} respectively,
FOC:

$$\Rightarrow \mu = \delta(\lambda\beta - \rho),$$
$$\Rightarrow \mu = -(1-\delta)\Big[\rho P'(v_{t+1}) + \lambda\beta\Big],$$
$$\Rightarrow \lambda > 0.$$

- PK is always binding to curb the growing package value which chips away profit.

Some initial results

- The comparatively impatient platform gradually adjusts the promises in an optimal way to retain the agent.

	$s_1 : \triangle I = 0$	$s_2 : \triangle I > 0$
v'	$\tilde{v}' = \frac{v}{\beta}$	$\hat{v}' = \frac{v}{\beta} + \delta \triangle I(v)$
c'	$\tilde{c}' = \frac{v}{\beta}$	$\hat{c}' = \frac{v}{\beta} - (1-\delta) \triangle I(v)$

- $\triangle I$ is the gap between $(\hat{v}', \hat{c}')$.

Some initial results

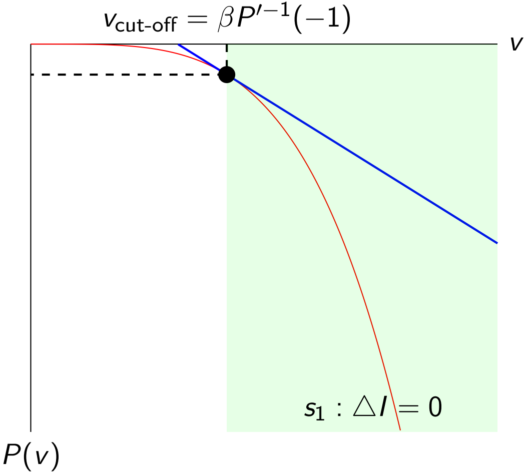
Theorem

In the absence of effort elicitation, when trying to produce a pair of promises to retain the agent for the next period. The profit-maximising platform is better off creating a gap between the pair (v_{t+1}, c_{t+1}) *whenever the current contract value is below* $\beta P'^{-1}(-1)$ *while conforming to a binding PK. Otherwise, the platform sets* $v_{t+1} = c_{t+1} = v_t/\beta$.

The proof is immediate if we compare the two profits generated from strategies s_1 and s_2.

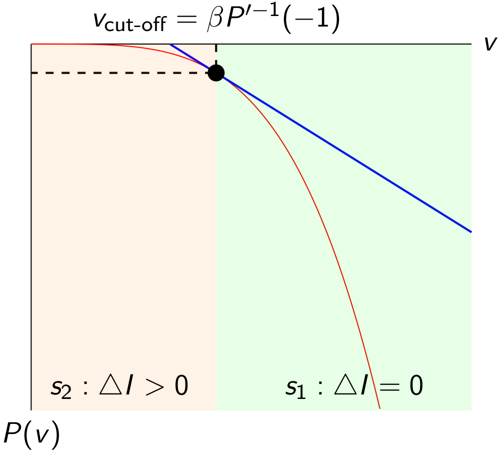
Sybil-proof during s_1

Linearly growing reputation does not encourage splitting: $v' = \frac{v}{\beta}$



Sybil-proof between s_1 and s_2

By adopting s_1 and s_2, the Sybil strategy is dominated when the agent splits causing a differential treatment that lowers the with-drawl value of the smaller accounts.



Require efforts to raise participation cost

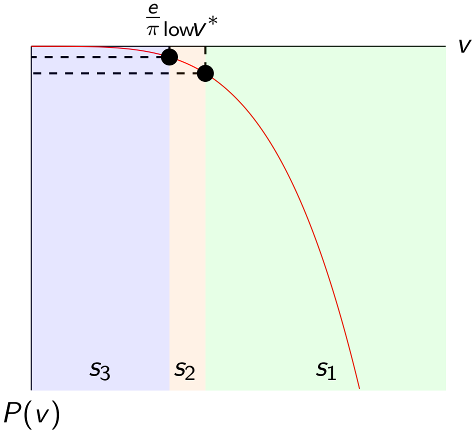
$$\beta\left[(1-\delta)v_{t+1}+\delta c_{t+1}\right]-e \geq \beta\left\{(1-\pi)\left[(1-\delta)v_{t+1}+\delta c_{t+1}\right]+\pi\left(c_{t+1}-\kappa(c_{t+1})\right)\right\}. \quad \text{(IC)}$$

- e is the mandated effort to maintain the identity in monetary terms.
- π is the success rate of a hardwired detection system.
- $\kappa(c_{t+1})$ is an automatic subtraction from c_{t+1} when the agent is detected for shirking.

Require efforts

- A contract of current value v is said to be *feasible* if the platform is able to issue a set of promises (v', c') that deters shirking.
- Whenever the PK constraint is binding, the set of feasible contracts requires the current contract value v to be at least e/π.
- To bring down the threshold for decentralisation purposes, the platform can sacrifice part of the profit and have an expanded strategy set.
- Set $\mu = 0$ s.t. $\beta\big[(1-\delta)v_{t+1} + \delta c_{t+1}\big] > v_t$.

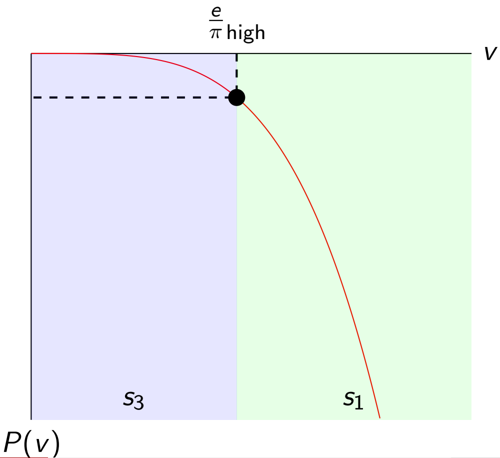
Case 1: low effort requirement/high detection rate



- Sybil-proof mechanism if there is no incentive to split for an s_2 user to disguise as many s_3 users.
- Sketch proof:
  1. Under liquidity shock δ.
  2. Restrictive withdrawal for all Sybils:
  3. When in s_3, $v'^{s_3} = \frac{v}{\alpha(v_0)} > \frac{v}{\beta}$.
  4. $c_t^{s_3} < c_t^{s_2}$ for $t \leq 1, \ldots k(\alpha(v_0))$, $v_k^{s_3} = \frac{e}{\pi}$.
  5. Agent bears multiple efforts.
  6. $\exists \alpha(v_0)$, $k(\alpha(v_0))$, s.t. splitting into more than 1 account is inferior to entering honestly as an s_2 user.

Case 2: high effort requirement/low detection rate

- Sybil-proof if there is no incentive to split for an s_1 user to disguise as many s_3 users.
- Proof: Similar



To conclude

- Online identities are only derived from their physical counterparts.
- Such Sybil attacks can have long-term structural effects to any distributed system.
- Stemming from wanting honest participation, our model captures some basic intuitions.
  1. Use two complementary promise instruments.
  2. Back-load promises.
  3. Different strategies needed to treat users with different commitment levels.