USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

Voting-Bloc Entropy: A New Metric for DAO Decentralization

Andres Fabrega, *Cornell University;* Amy Zhao, *IC3;* Jay Yu, *Stanford University;* James Austgen, *Cornell Tech;* Sarah Allen, *IC3 and Flashbots;* Kushal Babel, *Cornell Tech and IC3;* Mahimna Kelkar, *Cornell Tech;* Ari Juels, *Cornell Tech and IC3*

https://www.usenix.org/conference/usenixsecurity25/presentation/fabrega-entropy

This paper is included in the Proceedings of the 34th USENIX Security Symposium.

August 13–15, 2025 • Seattle, WA, USA

978-1-939133-52-6

Open access to the Proceedings of the 34th USENIX Security Symposium is sponsored by USENIX.

Voting-Bloc Entropy: A New Metric for DAO Decentralization

Andrés Fábrega[1], Amy Zhao[3], Jay Yu[5], James Austgen[2], Sarah Allen[3,4], Kushal Babel[2,3], Mahimna Kelkar[2], Ari Juels[2,3]

[1] *Cornell University* [2] *Cornell Tech* [3] *IC3* [4] *Flashbots* [5] *Stanford University*

Abstract

Decentralized Autonomous Organizations (DAOs) use smart contracts to foster communities working toward common goals. Existing definitions of decentralization, however—the 'D' in DAO—fall short of capturing the key properties characteristic of diverse and equitable participation.

This work proposes a new framework for measuring DAO decentralization called **V**oting-**B**loc **E**ntropy (VBE, pronounced "vibe"). VBE is based on the idea that voters with closely aligned interests act as a centralizing force and should be modeled as such. VBE formalizes this notion by measuring the similarity of participants' utility functions across a set of voting rounds. Unlike prior, ad hoc definitions of decentralization, VBE derives from first principles: We introduce a simple (yet powerful) reinforcement learning-based conceptual model for voting, that in turn implies VBE.

We first show VBE's utility as a theoretical tool. We prove a number of results about the (de)centralizing effects of vote delegation, proposal bundling, bribery, etc. that are overlooked in previous notions of DAO decentralization. Our results lead to practical suggestions for enhancing DAO decentralization.

We also show how VBE can be used empirically by presenting measurement studies and VBE-based governance experiments. We make the tools we developed for these results available to the community in the form of open-source artifacts in order to facilitate future study of DAO decentralization.

1 Introduction

A Decentralized Autonomous Organization (DAO) is an entity or community that operates based on rules encoded and executed on a public blockchain [19, 43]. DAOs can serve many goals, including investment [49], grant distribution [8], gaming-guild organization [6, 7], and ecosystem governance [10, 11]. DAOs play a prominent role in blockchain ecosystems, and are rising rapidly in popularity: at the time of writing (Aug. 2024), the aggregate value across all DAO treasuries exceeds $22 billion [5].

As the name suggests, a DAO's governance is decentralized, meaning that decision-making does not rely on a single individual or highly concentrated authority—in contrast to, e.g., a corporation, where a CEO and board of directors make major decisions. Instead, decisions in a DAO are typically made through community votes on proposals, the outcomes of which are enforced automatically by the blockchain in which the DAO's rules, i.e., its *smart contract*, reside.

Decentralization is a core property that DAOs strive for, as diverse and equitable participation are fundamental ideals in these communities (and the blockchain ecosystem more broadly) [1]. Most DAOs have their own associated crypto assets (or "tokens") and weigh voting power by token holdings. However, it is common for vote outcomes to be determined by a small set of "whales"—a colloquial term used to denote the largest token holders. Such centralization, as well as low voting participation, are a pervasive source of concern in DAO communities.

Decentralization is of critical importance and concern not just when it comes to successful governance of DAOs but also DAO *security*. As noted in [32], "if a large (delegated) token supply is held only by a few addresses or entities, many attack vectors become more likely to succeed." Poorly decentralized DAOs have historically been at higher risk of various governance attacks, including cabals pushing through adversarial proposals, e.g., [83], attackers plundering treasuries [32, 62], and systemic voter bribery [20]—for which active marketplaces exist today [59].

Academic works and DAO community participants have studied DAOs [27, 32, 33, 41, 73] and recommended ways to improve their decentralization [48]. To do so effectively, though, requires an ability to *model* and *measure* decentralization in a way that is reflective of a broad set of real-world concerns. These requirements motivate our work in this paper.

DAO decentralization: previous attempts. The most common basis for evaluating decentralization in DAOs and other blockchain settings is *token ownership*, specifically the distribution of assets and consequently voting rights among participants [41, 73]. Informally, concentration of a large fraction

of tokens in a small number of hands—and thus the ability of a small group to determine voting outcomes—is indicative of strong centralization. More widespread distribution, conversely, suggests decentralization. Prior decentralization measures generally formalize this intuition by computing a particular function over the distribution of tokens among individual members of the DAO, such as entropy,[1] the Gini coefficient [35] or the Nakamoto coefficient [9].

Token ownership distribution across individual accounts has serious shortcomings as a framework for decentralization, however. To begin with, it is visible on chain only in terms of per-address holdings, not control by real-world individuals. For instance, an individual who holds 51% of tokens in a DAO, but spreads them among a large number of addresses, could create an appearance of decentralization while having majority control. Even if tokens are held by distinct entities, a notion put forward in, e.g., [51], those entities may have *aligned interests and act in concert*—a form of centralization. The following examples illustrate cases in which a DAO may be strongly centralized, *even if token ownership appears to imply strong decentralization*.

Example 1 (Low participation / apathy)**.** Lack of participation in DAO governance votes is widespread in practice [23] and induces a form of centralization. Consider, for example, a DAO proposal where half of voters do not vote and voters other than whales vote "yes" by a 2:1 margin. Whales with just 12.6% of all tokens can swing the vote and force a "no" vote—an example of centralized power.

Example 2 (Herding)**.** Interviews with DAO participants reveal a tendency to vote in alignment with influential community members to preserve reputation [73], as individual votes are today usually publicly observable. This effect—often called *herding* [14]—has a centralizing effect. It aligns votes around the choices of a few participants. (This problem is similar to "herding" in classical voting theory [42].)

Example 3 (Bribery / vote-buying)**.** Bribery—specifically, *vote-buying*—has been a longstanding concern of DAO organizers [20, 31]. It has a centralizing effect, as it aligns voters around a choice dictated by the briber.

Recognizing that token-ownership alone doesn't give a full picture of decentralization, researchers have explored broader notions. Most notably, Sharma et al. [73] have considered entropy measures limited to those voters who participate in votes, and have also explored graph-based representations of voting patterns (degree centralization, degree assortativity, etc.). Token-ownership distribution among voting participants overlooks important issues, such as those in Examples 2 and 3, however, and it is unclear how to interpret graph-based empirical models. With no consensus in the community about how to measure DAO decentralization today, there is a lack of principled guidance on ways to improve DAO decentralization and to combat threats to decentralization, such as vote-buying.

Voting-Bloc Entropy (VBE). The primary contribution of our work is to introduce ***V**oting-**B**loc **E**ntropy* (VBE, pronounced "vibe"), a decentralization measure tailored to DAO governance. VBE is based on a core principle: individual voters with closely aligned interests across elections are a centralizing force—who operate in concert as single, abstract voting entities—and should be modelled as such. Expressed differently, the key idea in VBE is to *define centralization as the existence of large voting blocs*.

Formally, we express this principle in terms of the *utility functions* [39] of DAO participants, i.e., quantification of the gain or loss associated with voting outcomes. For a given set of elections, a voting bloc is a cluster of voters whose utility functions are similar over outcomes. VBE then, measures entropy over voting blocs based on utility functions—rather than over individual token holdings. The result is a broad concept that captures the centralization embodied in all of our examples above. Thus, VBE is a more accurate reflection of decentralization than prior metrics, as it accounts for subtle, centralizing forces which are not apparent from token ownership alone, but which affect a DAO's *true* degree of decentralization. These centralizing effects are reflected in the *alignment of voters' utility functions*, which VBE's clustering step is sensitive to. Note that VBE is in fact a framework: It allows different notions of clustering and entropy to be plugged in, and thus can be tailored to particular applications.

Unlike prior decentralization metrics, we arrive at VBE from first principles. Toward this end, we introduce the *Learning Hypothesis of DAOs (LHD)*, a reinforcement learning (RL) conceptual model of voting in DAOs. The key insight in the LHD is that the purpose of voting is to collectively *explore a policy space*, with the goal of maximizing some global reward function, (e.g., the DAO's treasury), as well as individuals' local reward functions (e.g., their monetary returns).

We show how this process can be naturally modeled as an RL problem, specifically a *multi-agent* RL or *MARL* problem. This modeling—the basis for the LHD—underpins our reasoning about DAO decentralization. Specifically, existing work on MARL stresses the importance of *agent diversity*, an aspect of MARL that motivates our formulation of VBE.

VBE in theory. VBE is an abstract metric: It cannot be measured *directly*, since users do not typically express (or even know) their utility functions explicitly. That is, utility functions are so-called *latent variables* [36], and thus VBE is as well. However, VBE provides an important basis for *reasoning about the impact of policy choices on decentralization*.

We use VBE to prove a number of simple theorems about how various practices might increase or decrease DAO decentralization. In some cases, our theorems capture intuitive or folklore notions of decentralization expressed by the community (e.g., the three examples mentioned above). In other cases,

[1] Entropy is typically defined over a random variable. A token ownership distribution may be viewed as a random variable for an experiment where a token is selected uniformly at random and its owner is output.

they offer *new* insights about decentralization. For example, we show that *as the decentralization of a DAO rises, so does the risk of systemic bribery—and vice versa*. This result—alongside our theorem showing that the act of bribery decreases decentralization—sheds new light on the connection between bribery in DAOs and decentralization.

Most importantly, several of our theorems lead to actionable recommendations regarding DAO governance. Further, our theorems serve as examples of a higher-level contribution: VBE's utility as a *theoretical tool* to derive formal results about decentralization.

VBE in practice. While latent variables are not directly measurable, they can be estimated via measurable quantities called *observable variables* [36]. As such, we introduce *observable VBE (oVBE)*, an observable variable that can be used to estimate VBE. oVBE estimates voters' utility functions in terms of observable, on-chain data—such as voting history—and clusters voters using this data. Like VBE, oVBE is a framework that can be instantiated with various clustering and entropy notions. By preserving VBE's structure, oVBE inherits the benefits of our framework, particularly with an accurate estimate of utility functions (we discuss this extensively in Section 5).

Since oVBE *is* directly measurable, it opens the door for VBE-based measurement studies and decentralization experiments. We report on two example use cases of oVBE.

First, we perform a measurement study of the historical oVBE across a number of popular DAOs, which we make available in the form of a public dashboard. Our dashboard contains a variety of per-DAO oVBE metrics, and updates weekly to reflect the evolving oVBE landscape.

Second, we show how oVBE can serve as a *metric in governance experiments* to understand the effect of a particular mechanism on a DAO's decentralization, by presenting an example experiment taking place in an ongoing collaboration with the Optimism Collective, who are using oVBE as a decentralization signal in an upcoming governance experiment in RetroPGF [69].

To perform these studies, we developed a suite of tools to process governance data and compute oVBE in a variety of settings. As an additional contribution, we make these artifacts available to the community in the form of a comprehensive, open-source *oVBE toolkit* for further decentralization studies. Our toolkit—which can be run as a standalone program or integrated as a library—supports various instantiations of oVBE, and is structured in a modular way so that users can easily tweak the various parameters to fit their use case.

Contributions. In brief, our contributions in this work are:

- *Learning Hypothesis of DAOs* (LHD): We introduce LHD, a simple multi-agent RL-based model of DAO voting (Section 2). This model is of independent interest, as it enables principled study of other questions about DAO governance.
- *Voting-Bloc Entropy (VBE)*: We use the LHD to derive VBE, a new framework for DAO decentralization that generalizes prior metrics and addresses a number of their shortcomings (Section 3).
- *Proving results about DAO governance*: We leverage VBE to prove results about the impact of DAO practices and designs on decentralization (Section 4), in some cases reaffirming some folklore notions and in others revealing new insights about DAO decentralization. More broadly, these theorems are examples of VBE's utility as a theoretical tool.
- *Empirical studies:* We introduce oVBE, a directly measurable variant of VBE. We perform a oVBE-based measurement study of popular DAOs, presenting a public dashboard with our results (Section 5.3). We also release an open-source toolkit for oVBE to stimulate future empirical work in this space, and report on real-world governance experiments using the toolkit (Section 5.4).
- *Practical guidance*: Based on our theoretical and experimental results, we present and summarize concrete points of practical guidance for DAO design and deployment (Section 6).

The main conceptual contribution of our work is a new *conceptual model* for DAO decentralization that captures voting entities—instead of individual accounts—characterized by their *alignment of incentives*. From this foundational idea, we derive our theoretical and experimental results. While these results are independently interesting (and indeed yield the actionable lessons we summarize in Section 6), they showcase a broader point, namely, VBE's flexibility and utility as a powerful tool to understand decentralization, which we hope motivates future work in assessing both the effectiveness of DAO governance. Furthermore, since decentralization is a critical security feature for DAOs, an accurate model and measure for decentralization are of central importance to DAO security—both in terms of monitoring propensity to risk, as well as the development of new techniques aimed at enhancing security by increasing decentralization.

2 Voting as a Learning Problem

The goal of our work is to introduce a decentralization framework for DAOs that addresses the gaps of existing metrics. To do so, we take a different approach to this problem than that of prior works: rather than starting from a candidate decentralization metric, we first take a step back, and start from a more foundational question instead: What is a *first principle* for voting in DAOs, from which we can then derive a decentralization metric? That is, what is DAO decentralization itself?

Towards this, we introduce the Learning Hypothesis of DAOs (LHD) in this section. The LHD is a multi-agent reinforcement learning (MARL)-based model that characterizes voting in DAOs, thus serving as a foundational principle from

which we can study DAO governance. We use the LHD to cast DAO decentralization as an analogue to *agent diversity* in MARL; this insight will serve as the starting point for VBE. This first-principles derivation thus supports VBE as a logical choice among possible metrics. We discuss the main conceptual ideas behind our model in this section—which are sufficient to understand the rest of our work—and refer readers to the extended version of our paper for more details (including additional background on MARL).

Brief background on MARL. Reinforcement learning (RL) [38] is a field of machine learning that studies the behavior of an agent that interacts with an environment, with the goal of maximizing some long-term expected reward. Multi-agent reinforcement learning (MARL) [37] is a particular type of RL, in which *multiple* agents are present in a shared environment.

MARL is a broad framework that can be used to model a multitude of problems. A MARL model has a few key components. There is an *environment*, and a set of *states* for this environment. There is a set of *agents* present in this environment, each with a set of *actions*, which defines the ways they can interact with the environment. The environment transitions from one state to the next as a result of the agents' collective actions; this transition is modeled via a *state transition function*, which denotes the probability of transitioning from one state to another given the agents' actions. Whenever the environment makes a state transition, each agent receives an immediate reward, which is modeled via a *reward function*.

The model proceeds in rounds. In each iteration, the agents observe the environment's current state and each receive a reward, from which each chooses its next action. This vector of actions is fed back to the environment, which transitions to a new state (as per the state transition function), resulting in new rewards for the agents (as per the reward functions). This process continues iteratively.

The actions taken by each agent are defined by a *policy*, which denotes the probability that the agent takes a particular action in a particular state. The quality of a "policy" can be expressed in terms of a *state-value function*, which represents the long-term expected return of executing the policy for many time-steps. The goal of each agent, then, is to find an *optimal policy* that maximizes the value function. Finding this optimal policy requires a balance between *exploration* and *exploitation*: the agent needs to search for new potential strategies while also taking advantage of its present understanding of the environment.

The relationship between the reward functions of the agents is a key property of MARL models. All agents can *cooperate*, which means that they have the same reward functions. Two agents can *compete* against each other, which means that their reward functions are the negative of each other. More generally, agents may have related but different reward functions, i.e., there are elements of both cooperation and competition (which is typically referred to as a "general-sum game").

Diversity in MARL. Diversity of viewpoints is well recognized as a key aspect of collective decision making, as the "wisdom of the crowds" often leads to better decisions than singular viewpoints [76]. Similarly, in the context of MARL, *diversity* between individual agents (particularly in the fully or partially cooperative setting) is an important property of a model that leads to more efficient learning in many contexts.

Agent diversity is a loose term that captures the degree of heterogeneity between agents, which can be analyzed from many angles (e.g., roles [77], actions [47], rewards [50], policies [65], etc.). Diversity has many benefits, including more efficient exploration [56], development of more advanced cooperation policies [56], discovery of niche skills [70], etc. As such, a number of diversity-boosting techniques have been put forth in the MARL literature, leading to experimental validation that diversity improves the collective performance of a multi-agent model [55, 56, 65]. Indeed, agent diversity has led to improved performance in many concrete applications of MARL, such as investment portfolio management [54], multi-robot systems [18], autonomous driving [86], etc.

To study the benefits of agent diversity for model performance, it is important to have a way to *measure* diversity to begin with. As such, there are many proposed diversity measures in the MARL literature. While an in-depth systematization of diversity metrics is outside the scope of this work, below we provide a high-level overview of diversity metrics.

A (short) taxonomy of MARL diversity metrics. We performed a literature review of diversity metrics in MARL to understand their similarities and differences. We discuss our main takeaways here, and refer interested readers to, e.g., [57, 58] for more technical details on diversity metrics, which are outside the scope of our work.

Conceptually, existing MARL diversity metrics have two main ingredients. First, some property of the agents is selected as the basis for gauging diversity (e.g., actions, reward functions, value functions, policies, etc). This is followed by computing some function over the the distribution of this property across agents (e.g., a statistical distance measure, correlation coefficient, entropy measure, etc.). Examples of mathematical objects that are used as the basis for agent diversity include policy functions [44, 60, 85], reward functions [70, 84], actions [66], and state or action-value functions [44, 82]. Then, examples of functions that are computed over these objects are entropy [85], Jensen-Shannon divergence [60], total variation distance [66], KL divergence [44, 85], and the Pearson correlation coefficient [84].

While diversity metrics vary a lot in their details—and there seems to be no consensus in the community for what is the best metric—our literature review revealed an important conceptual lesson: diversity metrics characterize differences in *behavior* (e.g., actions or policies) and *incentives* (e.g., reward or value functions) of the agents, instead of just individuality of the agents. Indeed, works in this space have emphasized the fact that "differences" is not equivalent to "diversity", and

how only optimizing for the former can lead to, for example, learning circular behaviors [81]. This insight, in addition to the two-point structure of agent diversity metrics, will form the basis for VBE.

Learning Hypothesis of DAOs (LHD). We now describe the *Learning Hypothesis of DAOs (LHD)*, a simple MARL model for voting in DAOs. This model is based on the insight that the fundamental goal of voting is to collectively *explore a policy space* in order to maximize a set of related objectives. While members of a DAO have individual objectives (e.g., maximizing their monetary holdings), these are related to each other by the fact that (by definition) DAO members have shared assets among themselves, e.g., the DAO's token. Thus, they share a global goal, such as maximizing the value of this token or the DAO's treasury. To pursue these goals, the members collectively perform iterative "actions" in some policy space, as determined by the proposals and the resulting votes. We can frame this process as a simple MARL problem, which is what we refer to as the Learning Hypothesis of DAOs (LHD). The LHD serves as a foundational model for DAO voting.

The environment corresponds to the blockchain, and each state corresponds to a state of the blockchain at the time when an election is taking place. The agents correspond to the members of the DAO. At each iteration, each agent has three available actions corresponding to voting choices for the election in question. The collective action of the multi-agent model consists of the vector of votes of all players, and the state transition function then moves to the next blockchain state as per the outcome of the vote, based on the DAO's voting system. After each state transition, the reward for each agent corresponds to their monetary utility for that election.

The goal of each agent is to maximize their long-term expected rewards. These rewards, however, are underpinned by some common objective, which broadly speaking represents the well-being of the DAO. As such, LHD represents a general-sum game, were players need some level of coordination in order to achieve individual benefit.

DAO decentralization as agent diversity. Based on the model of DAO members as MARL agents in LHD, we can *frame decentralization in DAOs as equivalent to agent diversity in MARL*. Diversity reflects the *true* heterogeneity of MARL agents, i.e., the fundamental differences between these that lead to meaningful improvements in exploration, contribute new perspectives, and lead to greater collective wisdom. These are the same set of principles that fundamentally represent decentralization in DAOs, thus providing a new lens through which to study decentralization. Based on this insight, we can leverage the literature on diversity metrics to inspire a construction for a DAO decentralization metric; while the specific diversity metrics in MARL are highly bespoke for RL and application-specific, we can base our construction on the core principles behind these. This leads us to VBE, which is the main contribution of our work.

3 Voting-Bloc Entropy (VBE)

We introduce *Voting-Bloc Entropy* (VBE) in this section, our new framework for decentralization that generalizes prior metrics and sidesteps their limitations. It does so by normalizing token holdings based on voters' utility functions.

VBE: core ideas. The starting point for VBE is the main conceptual takeaway we gleaned from our study of MARL diversity metrics: an accurate measure of decentralization should be reflective of the different *behaviors* and *incentives* among agents (voters), instead of merely identifying that the agents are different entities. This directly leads to the key idea behind VBE: instead of modeling decentralization in terms of the distribution of tokens across individual voters (as prior metrics do), we frame it instead in terms of the distribution of tokens across *groups of voters with aligned interests across elections*, which are functionally acting as a single entity. Similar to diversity in MARL, aggregating voters based on aligned interests allows us to capture interactions and relationships among players in the system, which determine the true degree of decentralization of a DAO.

We formalize the notion of "aligned interests" in terms of the DAO members' *utility functions* [39] across elections. In our setting, this is the natural mathematical object that represents the voters' incentives across elections (and, indeed, corresponds to reward functions in LHD, which is a common first-ingredient in MARL diversity metrics). VBE, then, computes some function over these utilities in order to partition the set of players, which ultimately serves as the basis upon which to gauge the distribution of tokens. Thus, conceptually, VBE generalizes existing decentralization metrics by bootstrapping the two-point structure of MARL diversity metrics as a *pre-processing* step to determine the groups—or *blocs*—of voters with aligned incentives, over which we can then compute the entropy (or other similar function) of tokens.

DAO abstraction. Before presenting the formal definition of our VBE framework, we introduce the notation that our definition and theorems rely on.

Let $\mathcal{P} = \{P_1, \ldots, P_n\}$ be the set of token holders in a system, and $\mathsf{tokens}\colon \mathcal{P} \to \mathbb{R}^+$ a mapping specifying the number of tokens held by each $P \in \mathcal{P}$. (We will often overload this notation, and input a *set* of accounts to tokens instead, by which we mean the total tokens held across all accounts in the set). These token holders participate in a set of (binary) elections $E = \{e_1, e_2, \ldots, e_m\}$, where we denote by $\mathsf{vote}_P\colon E \to \{\mathtt{true}, \mathtt{false}, \perp\}$ player P's vote in election e; $\perp$ indicates that P abstained from voting in e. We represent all of P's votes across E by $V_{E,P}$. We define $\mathsf{util}_P\colon E \times \{\mathtt{true}, \mathtt{false}\} \to \mathbb{R}$ to be the monetary utility of an outcome of $\mathtt{true}$ or $\mathtt{false}$ in e to player P, where we make the simplifying assumption that $\mathsf{util}_P(e, \mathtt{true}) = -\mathsf{util}_P(e, \mathtt{false})$. Player P's total utility across all elections E is represented by a vector $U_{E,P} \coloneqq (\mathsf{util}_P(e_i, \mathtt{true}))_{i \in [m]} \in \mathbb{R}^m$; we denote by $U_{E,\mathcal{P}}$ all players' utilities, i.e., $U_{E,\mathcal{P}} \coloneqq (U_{E,P})_{P \in \mathcal{P}}$.

Token holders often have low stakes in the elections, resulting in lack of interest or abstaining from voting altogether. More formally, we say that player P is ε-*apathetic* in election e if and only if $|\mathsf{util}_P(e,\texttt{true})| \leq \varepsilon$. We denote this set of apathetic voters by $\mathcal{A}$. If the system supports vote delegation (for example, as a means to combat voter apathy), players may delegate their tokens to others, who cast a single vote on behalf of all the tokens they now hold.

3.1 Framework for VBE

We present VBE in this section. VBE is an abstract *framework* for decentralization, which is parameterized by: (1) a clustering function, and (2) an entropy measure, which are the two key ingredients that underpin our definition.

Clustering. We let $C\colon U_{E,\mathcal{P}} \times U_{E,\mathcal{P}} \to \{0,1\}$ be a clustering function[2] that outputs 1 if the utilities of two players are "aligned" across all elections E, and 0 otherwise. Our definition of VBE is agnostic to a specific clustering function, and instead only assumes that C specifies an equivalence relation $\sim_C$ on the set $U_{E,\mathcal{P}}$. Note that C additionally induces a partition on $\mathcal{P}$, whereby P_i and P_j are in an equivalence class if and only if $C(U_{E,P_i}, U_{E,P_j}) = 1$. That is, C partitions $\mathcal{P}$ into classes of players with aligned utility functions across elections. We will often overload notation and directly refer to C as a partition of $\mathcal{P}$. Following standard notation, we denote that two players are in the same class by $P_i \sim_C P_j$, the set of all classes by $\mathcal{P}/\sim_C$, and the class P belongs to by $[P]$.

Note that our model (and, thus, our definition of clustering) assumes that players vote on binary elections. This assumption is just for clarity of presentation. We could also consider clustering functions that partition players based on their utility functions over elections with more outcomes. (In fact, we explicitly do this in one of our experiments in Section 5.)

Entropy. The clustering step serves as a way to normalize token holdings based on the players' aligned incentives across a set of elections. After this, we can then, as is standard, compute some measure of the distribution of tokens, but this time across *the resulting equivalence classes*.

More formally, we let $F : \mathbb{P} \times \mathcal{P}^{\mathbb{R}^+} \to \mathbb{R}$ be a function from the distribution of tokens across *sets* of accounts to real numbers. In this notation, $\mathbb{P}$ denotes the set of all partitions of $\mathcal{P}$. So, the function F takes as input a partition of $\mathcal{P}$ and the function $\mathsf{tokens}\colon \mathcal{P} \to \mathbb{R}^+$ (which maps accounts to the tokens they own), and returns some real number. The purpose of F is to measure, in some sense, how "evenly distributed" tokens are across voting blocs. For instance, F can be any of the many variants of Rényi entropy[3] (e.g., min-entropy, Shannon entropy, or max entropy), or any of the token distribution functions found in prior decentralization metrics. (In particular, note that we can represent any prior decentralization metric as an instantiation of VBE, by using the vacuous clustering function wherein each bloc has one and only one account, over which we compute the distribution function used by the metric.) We stress, however, that in principle F can be any function, and our definition makes no assumptions about its structure.

We are now ready to define VBE. Intuitively, our definition says that a DAO is more decentralized if the distribution of tokens across the blocs specified by $\sim_C$ has high entropy according to F. More concretely:

Definition 1 (Voting-Bloc Entropy)**.** For a set of elections E, a set of players $\mathcal{P}$ with corresponding utilities $U_{E,\mathcal{P}}$, a mapping specifying token ownership across accounts tokens, a clustering metric C, and an entropy measure F, we define *Voting-Bloc Entropy (VBE)* to be:

$$\mathsf{VBE}_{C,F}(E,\mathcal{P},U_{E,\mathcal{P}},\mathsf{tokens}) \coloneqq F(\mathcal{P}/\sim_C,\mathsf{tokens}).$$

Note that, since VBE is a framework, it is (more accurately) a *family* of decentralization metrics, all under the umbrella of VBE's two-step structure. That is, concrete decentralization metrics are a result of *instantiations* of VBE with particular clustering functions and entropy measures. In particular, VBE can be instantiated with, for example, state-of-the-art clustering functions (e.g., K-means, hierarchical clustering, DBscan, etc.), and commonly-used entropy measures (e.g., Shannon or min-entropy). The specific instantiation of VBE to use is problem-specific, analogous to how different clustering functions are better suited for different types of inputs and applications. (We discuss this in more detail in Section 5).

An overview of VBE uses. As we emphasized in Section 1, (any instantiation of) VBE is a *latent metric*, i.e., it cannot be measured directly. This is due to the fact that utility functions are latent variables [36]—conceptually important, but not always directly measurable. Thus, a priori, VBE's primary use case is as a *theoretical tool* to formally derive results about the impact of policy choices or practices in decentralization, for which conceptual reasoning about utility functions is sufficient. We discuss this use case in more detail in Section 4.

The second application of VBE is as an *empirical tool* for VBE-based measurement studies and governance experiments. While we cannot directly measure the "true" VBE of a DAO, we can *estimate* it: utility functions (and, thus, VBE) can be approximated via measurable *observable variables* [36]. This variant of VBE, which we cover in depth in Section 5, inherits all the conceptual benefits of the VBE framework while additionally being tractable.

[2] Formally, a clustering function takes as input a set and returns a partition of it. For clarity of presentation, we represent C as the induced equivalence relation of the partition, as these are corresponding terms.

[3] Entropy is formally defined over a random variable, but we are overloading notation to think of the mapping between sets of accounts and their respective cumulative token balances as the probability mass function of a random variable.

3.2 VBE and DAO Security

While VBE is a decentralization metric, it has important connections to DAO *security* as well, which we discuss in this section. Uses of VBE (which are the focus of subsequent sections) therefore have direct applications to DAO security.

Decentralization and DAO security. As discussed in Section 1, decentralization is a critical *security* feature for DAOs: inadequate decentralization can lead to unfair extraction of value from DAOs by subgroups, and even outright theft in the form of *governance attacks*. Notable examples of governance attacks include treasury raids [83], rug pulls [46], systemic bribery [20], flash loan attacks [28], and more. DAOs that are more centralized are at higher risk for such attacks [32]. Adversaries often attack DAOs by acquiring sufficient voting power to perform a malicious action. Thus, more centralized DAOs—particularly those with high voter apathy—are especially vulnerable because of: (1) a low threshold to trigger an attack, (2) easy formation of adversarial coalitions, and (3) difficulty in reverting the impact of attacks. A healthy degree of decentralization is thus critical for mitigating security risks.

VBE and DAO security enhancements. The importance of decentralization for DAO security highlights the fact that an accurate decentralization *measure* is a critical security tool. VBE can thus help bolster DAO security in two ways:

- *Monitoring*: Continuous tracking of a DAO's decentralization helps monitor its security and propensity to risk. By offering a more accurate decentralization measure, VBE more accurately reflects a DAO's risk profile than previous measures. Importantly, many governance attacks are underpinned by dangerous *group alignment*, which VBE, unlike prior metrics, is sensitive to.
- *Design*: VBE offers a means of assessing the impact of governance design choices aimed at increasing decentralization to enhance security. VBE can be used to formally prove the directional impact of policy choices on decentralization (Section 4); and to experimentally test decentralization interventions aimed at increasing security (Section 5).

Now that we have introduced the VBE framework, and its relevance to DAO security, the rest of our work presents and exemplifies its use cases.

4 VBE in Theory

We present a variety of theoretical results implied by VBE in this section. In some cases, our results are simple, and show how VBE confirms intuitive notions about decentralization expressed by the community. While simple, these results show how VBE is able to capture many of the subtle issues that impact decentralization in a DAO. In other cases, VBE provides novel insights on decentralization, from which we can derive practical guidance (see Section 6). More broadly, this section showcases VBE's utility as a theoretical instrument, and how it can serve as formal groundwork for future results about decentralization. In particular, VBE can be used as a tool to assess the theoretical impact of governance design choices that are aimed at enhancing a DAO's security by increasing decentralization (Section 3.2).

Before presenting the implications of VBE, we first remark that for most "reasonable" instantiations of F (such as Shannon or min-entropy), the "trivial" clustering function which assigns each player to its own cluster *gives an upper bound on VBE*. Concretely, this fact holds for any F that increases whenever the tokens held by any players's voting bloc increase: if this is the case, for all players in the system, the number of tokens held by their bloc according to any clustering metric is necessarily greater than or equal to the number of tokens held by a "bloc" that only contains themselves. In particular, recall from Section 3 that prior decentralization metrics can be cast in the VBE framework precisely as instantiations that use this trivial clustering function. As such, VBE is, at worst, equivalent to the entropy-based notions introduced by prior work, which focus on account balances alone. (We also show this empirically in Section 5.)

4.1 Implications of VBE

We now explain the theoretical insights implied by VBE.

VBE Master Theorem. Our theoretical results all aim to show the impact of policy choices or system changes on DAO decentralization, in terms of VBE. They all have a similar structure: (1) we consider two systems such that the only difference between them is some "transformation" of interest, e.g., a portion of the voters become apathetic, votes are instead private, etc; (2) we reason about the impact of this transformation on the voting blocs of both systems; (3) based on this, we compute and compare the VBE of both systems. We now define a "master" theorem for VBE which captures this structure, and thus serves as a proof template that can be instantiated with concrete transformations of interest to prove different results. Our theorem is stated in terms of an arbitrary clustering function C, and *min-entropy* as the entropy measure. Min-entropy captures the amount of "information" in the largest voting bloc by token holdings, i.e., for a set of sets of addresses A with a total of T tokens held across all individual accounts,

$$F_{\min}(A, \mathsf{tokens}) := \log_2 \left(\frac{\max_{A' \in A} \mathsf{tokens}(A')}{T} \right).$$

We use min-entropy to state the VBE master theorem for clarity of presentation, but it can be refined to account for other entropy measures.

Theorem 4.1 (Voting-Bloc Entropy Master Theorem)**.** We define T to be a function that represents a *system transformation*, i.e., a change in the players, elections, utilities of

the players, and/or the distribution of tokens, which we denote by $(\mathcal{P}', E', U'_{E',\mathcal{P}}, \mathsf{tokens}') := T(\mathcal{P}, E, U_{E,\mathcal{P}}, \mathsf{tokens})$. The total number of tokens in the system stays constant, however. Let B and B' be the (not necessarily unique) largest clusters by token holdings according to clustering function C for $(E, U_{E,\mathcal{P}}, \mathsf{tokens})$ and $(E', U_{E',\mathcal{P}}, \mathsf{tokens}')$, respectively. Then, it follows that

$$\mathsf{tokens}'(B') \geq \mathsf{tokens}(B) \iff$$
$$\mathsf{VBE}_{C,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}) \geq \mathsf{VBE}_{C,\min}(E', \mathcal{P}', U'_{E',\mathcal{P}}, \mathsf{tokens}').$$

Proof. This follows directly from the definition of $\mathsf{VBE}_{C_\varepsilon,\min}$:

$$\begin{aligned}
&\mathsf{tokens}'(B') \geq \mathsf{tokens}(B) \\
\iff &\frac{\mathsf{tokens}'(B')}{\sum_{P\in\mathcal{P}'} \mathsf{tokens}'(P)} \geq \frac{\mathsf{tokens}(B)}{\sum_{P\in\mathcal{P}} \mathsf{tokens}(P)} \\
\iff &-\log_2\Big(\frac{\mathsf{tokens}'(B')}{\sum_{P\in\mathcal{P}'} \mathsf{tokens}'(P)}\Big) \leq -\log_2\Big(\frac{\mathsf{tokens}(B)}{\sum_{P\in\mathcal{P}} \mathsf{tokens}(P)}\Big) \\
\iff &\mathsf{VBE}_{C,\min}(E', \mathcal{P}', U'_{E',\mathcal{P}'}, \mathsf{tokens}') \\
&\leq \mathsf{VBE}_{C,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens})
\end{aligned}$$

□

Note that, if B' represents a (new) majority by token holdings, then VBE strictly decreases; equality follows when $\mathsf{tokens}'(B') = \mathsf{tokens}(B)$.

This master theorem thus serves as a template that individual theorems can bootstrap from: simply specify a transformation T, explain how this modifies the largest voting bloc (if at all) according to the clustering function, and invoke Theorem 4.1. Armed with this formula, we now move on to concrete theoretical insights implied by VBE. Due to space constraints, we present the full theorem statement and proof for only the first of our results, to showcase the general structure of these. For the rest, we present the conceptual ideas behind the results here, and refer readers to Appendix A and the extended version of the paper for the full details,.

As explained in Section 3.1, the VBE framework is compatible with any clustering function that defines a partition on $\mathcal{P}$, even if the function is vacuous or contrived. Therefore, in order to derive meaningful conclusions, our results will need to assume basic properties about the clustering algorithm in use, as generic results for arbitrary partitions would lead to trivial conclusions. For the rest of this section, we assume that we are dealing with any clustering function C satisfying two simple properties. First, there exists some small constant ε such that, if P_i and P_j are such that $|\mathsf{util}_{P_i}(e, \mathtt{true})| \leq \varepsilon$ and $|\mathsf{util}_{P_j}(e, \mathtt{true})| \leq \varepsilon$ for all elections $e \in E$, then $C(U_{E,P_i}, U_{E,P_j}) = 1$. Second, if P_i and P_j are such that their utilities for every election have the same sign, then $C(U_{E,P_i}, U_{E,P_j}) = 1$. We note that our results can be modified in straightforward ways to accommodate for other "natural" assumptions of clustering functions, and thus our conceptual takeaways are general; we restrict ourselves to the aforementioned ones for clarity of presentation. More importantly, these assumptions are satisfied by all standard clustering functions if we focus our analysis on *ordinal* utility functions. Looking ahead, this will be the case in Section 5, where we empirically estimate VBE using voting history (there, we also discuss ordinal utilities, and their extensive use in economics, in more detail).

Result #1: Owning multiple accounts. As explained in Section 1, previous notions of entropy fail to capture the centralization that is present (but hidden) when a whale distributes tokens across multiple accounts / addresses. In such cases, it may appear that tokens are well diversified across accounts, while a large fraction are in fact under the control of one entity. Unlike prior notions, VBE captures this nuance: all these accounts would indeed be grouped together in a single voting bloc (we make the simplifying assumption that an individual's utility function is the same across all her accounts). We formalize this below.

Theorem 4.2 (Sybil Attacks and VBE)**.** Let $(\mathcal{P}', E, U'_{E,\mathcal{P}'}, \mathsf{tokens}') = T_{\mathtt{mult}}(\mathcal{P}, E, U_{E,\mathcal{P}}, \mathsf{tokens})$ be the transformation where some player $P \in \mathcal{P}$ divides her tokens across a new set of accounts $\hat{\mathcal{P}}$, i.e., $\mathcal{P}' = \mathcal{P} \cup \hat{\mathcal{P}}$, $\mathsf{tokens}'(\hat{\mathcal{P}}) = \mathsf{tokens}(P)$, and $\forall \hat{P} \in \hat{\mathcal{P}}$, $U'_{E,\hat{P}} = U_{E,P}$. The rest of the system remains unchanged. Then, it follows that

$$\mathsf{VBE}_{C,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}) = \mathsf{VBE}_{C,\min}(E, \mathcal{P}', U'_{E,\mathcal{P}'}, \mathsf{tokens}').$$

Proof. Let B be the largest voting bloc by token holdings before $T_{\mathtt{mult}}$, which may or may not include P. By assumption, all $\hat{P} \in \hat{\mathcal{P}}$ are such that $U'_{E,\hat{P}} = U_{E,P}$. Thus, all new accounts will be in the same voting bloc B' after $T_{\mathtt{mult}}$, namely, $B' = [P]$.

It follows then that, even though P's tokens are distributed between all individual accounts in $\hat{P}$, they are in fact still under the control of the same block, i.e., B'. As such, $\mathsf{tokens}'(B') = \mathsf{tokens}(B')$. So, since no blocs acquire any new tokens, B is still the largest voting bloc by token holdings after $T_{\mathtt{mult}}$. Then, from Theorem 4.1 it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}) = \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U'_{E,\mathcal{P}}, \mathsf{tokens})$$

as desired. □

This result shows that, according to VBE, the "true" decentralization of the system does not change when a whale splits her tokens into multiple accounts, as they are all still under the control of the same voting entity. Conversely, metrics that focus on account balances alone would mistakenly conclude that the decentralization of the system strictly increased, since a set of tokens is diversified across more accounts.

Result #2: Apathy. A system where voters are apathetic, i.e., not interested in the direction of the community, is not aligned with the goals of a DAO: distribution of tokens is irrelevant if individuals abstain from voting, as elections are narrowed squarely to the set of more invested stakeholders. Our definition captures this fact. Intuitively, *apathetic voters all have similar utility functions*, which reflects their lack

of stake in the elections. VBE groups all of these players within the same voting bloc, i.e., the cluster of voters with low utilities.

If the disinterested players are small stakeholders to begin with, apathy has a centralizing effect, as they now belong to a larger bloc of aligned voters. Indeed, in practice, it is common for the set of apathetic voters to represent a majority of token holdings [23, 41]. We refer to the bloc of apathetic voters in a DAO, i.e., non-voting token holders, as the *inactivity whale*. This term reflects the collective and potentially systemically important inactive behavior of this group.

Result #3: Delegation. Intuition would suggest that delegation leads to a more centralized system: tokens that are originally held by a large set of players are instead under the control of the (smaller) set of delegates. However, VBE shows how this situation is more nuanced, as delegation actually tends to make a DAO *more* decentralized: before delegation, the tokens are all held by a *single* voting bloc, namely, the inactivity whale. Delegation then diversifies the tokens held by this "whale", and distributes them amongst a set of voting blocs (the delegates). Assuming that the size of the inactivity whale is larger than each delegate's total tokens—which tends to be true in practice [23, 41]—the system is now more decentralized. That is, as long as the delegates are not "too big," delegation has a decentralizing effect. Conversely, if some delegate is a whale, or gets delegated an overwhelming majority of tokens, then the system may become more centralized. Thus, delegation is most useful in cases of high apathy.

Result #4: Herding. A core goal of DAOs—and any democratic system more broadly—is for token holders to vote according to their true preferences. In practice, however, many DAOs exhibit *herding* behavior: when votes are publicly observable, social dynamics lead to the formation of "coalitions" of voters. For example, token holders have reported feeling influenced to vote a certain way, often in alignment with influential community members, in order to thwart the reputational risks associated with opposing popular viewpoints [73]. Similarly, it has been observed and measured that token holders often vote in alignment with their peers [67], who now operate as a single, large entity. In both cases, the utility derived from the social impact of a player's vote skews the utility of her desired outcome in a vacuum. Herding leads to more centralization, as votes artificially converge on one outcome. Unlike token distribution across individual accounts, VBE does conclude that reputational risk lead to more centralization: it aligns the utilities of the players towards the socially preferred outcome, which results in a bloc of aligned voters.

An important conclusion of this theorem is that privacy instead *increases* the decentralization of a system, as it serves as a "mitigation" to herding. That is, if votes are private, token holders can vote for their true preferences, instead of being influenced by, e.g., social optics or the votes of their peers.

Result #5: Voting slates. Grouping together various elections into a lesser number of (more general) elections—so-called "voting slates"—is in opposition with decentralization: decision-making is more diluted, thus decreasing the relative impact of each voter in the underlying proposals. That is, voting slates "factor out" differences in the viewpoints of individuals, yielding more homogeneous utility functions. For example, two players may disagree in many of the individual proposals, but agree on a few of the more important ones, resulting in them casting the same overall vote. VBE reflects the fact that bundling proposals indeed decreased decentralization: by considering a narrower set of elections, which smoothens utility functions, different voting blocs are combined to form larger ones.

Result #6: Bribery. There is an intuitive relationship between decentralization and bribery, namely, that successful bribery poses a threat to decentralization: in such a case, the entity that acquires the votes of the other players now commands a higher proportion of the total tokens than before. While ownership of tokens has not changed, VBE is sensitive to this centralizing effect, as it groups all bribed voters into the briber's bloc, since all bribee's now have aligned utility functions in line with the bribers desired outcome.

A second, more nuanced observation is that successful bribery must be *systemic*, i.e., must involve a large number of tokens, if (and only if) a system is highly decentralized. Intuitively, if a DAO is highly centralized, a briber can directly coordinate with a few large players to guarantee an election outcome; or, if the briber is a whale herself, she only needs to bribe a few of the smaller players to accumulate enough tokens to mount a successful attack. Instead, in a more decentralized system, players are smaller, so a briber needs to widen the scale of their attack if they want to win an election. In this case, successful bribery requires large-scale coordination among various smallholders. Thus, as a DAO becomes more decentralized, a higher number of tokens need to be corrupted to guarantee an election outcome, since all players are small to begin with. Conversely, in a more centralized DAO, large whales only need to coordinate among themselves, or corrupt relatively few additional tokens to guarantee their desired election outcome.

Though a longstanding concern, systemic bribery is generally not considered realistic in secret ballot elections, due to logistical and economic challenges and criminal penalties—not to mention difficulty in verifying voters' compliance with bribers' demands. DAOs, however, are vulnerable: (1) Vote-buying is increasingly legal for proxy voting and thus may well be for DAOs [24]; and (2) Vote-buying in DAOs can be programmatically executed and verified by smart contracts, as shown by active marketplaces such as Votium [59]. It is even technically possible in principle for vote-buying to occur confidentially [27].

Result #7: Quadratic voting. Quadratic voting [53] is a voting mechanism that attempts to dilute the influence of whales on election outcomes. To do so, a vote from a player that

owns n tokens will only have an impact of $\sqrt{n}$ in the outcome election. At face value, quadratic voting seems to make a system more decentralized: the quadratic "tax" is directly proportional to the number of tokens a player owns, which thus shrinks the gap between smaller players and whales. However, quadratic voting is known to be vulnerable to Sybil attacks and other forms of malicious coordination [78], and thus may have a *centralizing* effect: players that are in large voting blocs implicitly subvert the quadratic tax due to the fact that their true token count is split among all bloc members.

5 VBE in Practice

As we have emphasized throughout this work, utility functions are latent variables, and consequently VBE is as well. Thus, even though VBE serves as a powerful conceptual tool, it is not directly measurable, which is an inherent limitation of any metric that depends on utility functions (including important results and models from voting theory, e.g., [29, 78]). However, latent variables, such as utility functions, can be measured *indirectly*, by inferring them via *observable variables*, which do lend themselves to direct measurement. We introduce "observable" VBE (oVBE) in this section, a variant of VBE based on estimating utility functions from observable data. Unlike its latent counterpart, oVBE *is* directly measurable, and thus provides an estimate for the "true" VBE of a DAO. oVBE represents the practical, measurable half of the VBE framework, which opens the door for exciting applications, many of which we highlight in this section.

5.1 Observable VBE (oVBE)

Recall that the VBE framework consists of two main steps: (1) clustering players based on utility functions, followed by (2) computing some entropy measure over the distribution of tokens over the resulting clusters. oVBE simply adds a preliminary step to this template: represent each player's utility function in terms of relevant observable data, and perform the clustering based on this representation instead. We now define oVBE formally, and then move on to study oVBE in practice, including an empirical validation of oVBE.

Defining oVBE. oVBE extends the VBE framework by introducing a *metric space* $M := (S,d)$ as a third parameter (alongside the clustering function and entropy measure) to the framework, where S is some set with $|\mathcal{P}|$ elements and d is a distance function on S. The space S represents a class of observable data that are being used to estimate utility functions, and the function d allows a clustering metric to partition S, since clustering functions require some similarity metric between their inputs in order to assign clusters. We thus get the following definition of oVBE:

Definition 2 (Observable Voting-Bloc Entropy)**.** For a set of elections E, a set of players $\mathcal{P}$, a mapping specifying the distribution of token ownership tokens, a clustering function C, an entropy measure F, and a metric space $M = (S,d)$ with $|\mathcal{P}|$ elements, we define *Observable Voting-Bloc Entropy (oVBE)* to be:

$$\mathsf{VBE}_{C,F,M}(E,\mathcal{P},\mathsf{tokens}) := F(\mathcal{P}/\sim_{C_M},\mathsf{tokens}).$$

In the notation above, $\mathcal{P}/\sim_{C_M}$ represents partitioning $\mathcal{P}$ based on how C partitions S using d. That is, P_i and P_j are in the same equivalence class if and only if $S_i \sim_{C_d} S_j$.

oVBE in practice. More intuitively, oVBE simply consists of (1) using some relevant dataset (e.g., on-chain metrics) that characterizes the DAO members' utility functions across elections, (2) defining a notion of "distance" between elements of this dataset, (3) computing a clustering function over this dataset (which uses the distance metric), and (4) computing the entropy metric over the tokens held by each cluster.

The observable data used for clustering is a key parameter for oVBE. There is a rich body of work dedicated to inference of latent variables (and of utility functions specifically) [68], and indeed oVBE is agnostic to the method used. The most natural starting point, which we use throughout the rest of this work, consists of using *voting history* as the observable data that estimate a player's utility function for a set of elections. Assuming that players are rational actors, their vote (or lack thereof) in an election is equivalent to the *direction* of their utility for that particular election. That is, for any player P and election e, it follows that if $\mathsf{vote}_P(e) \neq \perp$, then $\mathsf{util}_P(e,\mathsf{vote}_P(e)) > \varepsilon$; conversely, if $\mathsf{vote}_P(e) = \perp$, then $|\mathsf{util}_P(e,\mathtt{true})| < \varepsilon$. As such, P's vector of votes in the set of elections E, denoted by $V_{E,P}$, represents their *ordinal* utility function for E.

In this example, the dataset for clustering consists of a matrix $\mathcal{H}_{\mathcal{P},E} := (V_{E,P})_{P\in\mathcal{P}}$, where each row represents the list of votes for that particular player across elections. We can then use any number of distance metrics and clustering functions for vectors, such as cosine similarity or Euclidean distance for the former, and K-means or hierarchical clustering for the latter. As with other applications of clustering, the best variant to use depends on the nature of the input data. So, the choice of algorithm is a function of the observable data being used, and its characteristics. Standard considerations for choosing between algorithms apply to our setting (e.g., dataset size, dimensionality of points, sparsity of data, etc), and so ample literature [80] on clustering functions is directly applicable. For the example of $\mathcal{H}_{\mathcal{P},E}$ as input data, many of these clustering considerations translate to the nature of the DAO's elections, such as the number of proposals and voters, the degree of voter apathy, etc.

We note that, in some cases, a DAO's proposals and goals may be purposefully designed for broad acceptance, e.g., proposals for security patches or defeating attacks. oVBE is equipped to capture this nuance, by using a clustering function that is refined to take this into account. For example, we can

use a standard clustering function and apply less weight on these "popular" proposals, or remove these from the clustering altogether. Similarly, the duration of alignment for clustering can be adjusted, simply by modifying the size of each oVBE window, i.e., the number of proposals that are included in each computation of oVBE.

How closely oVBE estimates VBE will be primarily based on the observable data that are gathered, and how accurately it estimates utility functions. We refer readers to external sources such as [68] for more details on this relationship, which is outside the scope of our work. As mentioned above, for the rest of this work we will use voting histories, i.e., ordinal utilities, to compute oVBE. Ordinal utilities are widely used in the economics literature as a suitable estimate for utility functions [63] (and, in fact, are equivalent to cardinal utilities in some cases [40]), and thus our instantiation of oVBE serves as a good proxy for VBE. Other potential on-chain data that could indicate alignment include, for example, assets owned, membership in other DAOs, sponsored proposals, etc., and thus could potentially also be used as observable data. An interesting direction for future work is the use of off-chain data sources, such as social media interactions or low-cost straw polls, to refine estimates of utility functions.

oVBE open-source toolkit. Now that we have introduced oVBE, we are ready to perform measurement studies and derive VBE-based applications. For this purpose, we prepared a comprehensive oVBE toolkit that can be used for easily computing oVBE as part of broader applications. Our toolkit, which we make available to the community as an open-source artifact [2], contains a variety of popular distance metrics, clustering functions, and entropy metrics, which can be assembled together to get a myriad of oVBE instantiations. Our toolkit can be used as both a standalone program or integrated as a library in applications, and simply requires users to input a dataset and select a clustering function, distance function, and entropy measure. In the documentation of our toolkit, we include guidance for selecting specific oVBE instantiations based on the nature of the input data, to complement existing documentation on choices of algorithms.

Our toolkit is modular by design and can be easily extended to add support for more algorithms. Further, our toolkit provides a number of additional features besides just computing oVBE. For example, it has support for an intermediate report of the clusters and support for automatic detection of parameters. We refer readers to the documentation of our toolkit for more details on the features included.

While our toolkit is agnostic to the dataset used as input, we assume that voting history will often be used as observable data. As such, we include scripts for scraping the voting history of DAOs using open source APIs in a format that is directly compatible with the rest of our toolkit. (As described in Section 5.3, scraping voting history is surprisingly difficult).

In the rest of this section, we put this toolkit to the test by performing a variety of oVBE-based empirical studies. While these are independent contributions, they additional serve to showcase the flexibility and uses of our toolkit (and of oVBE more broadly).

5.2 Empirical Validation of oVBE

The starting point for our empirical work is an an experimental validation of oVBE. We first identify a *natural experiment* that can be used as the basis for testing a decentralization metric, which we then apply to oVBE.

A natural decentralization experiment. An ideal experiment to validate a decentralization metric consists of a single population of players who vote on two sets of proposals, such that the proposals in one set are purposefully crafted to yield more centralized results than the other. Then, one could independently compute a candidate decentralization metric on both sets of proposals, and see whether the outputs agree with the expected centralization.

It turns out that a *natural* example of this experiment takes place in some DAOs today, in the form of *two-round voting*. In these DAOs, proposals initially go through a preliminary round of voting—a so-called *temperature check*—during which the community discusses the proposal and agrees on its parameters. This is followed by a second round of voting, which decides whether the proposal is ultimately accepted. The goal of the first voting round is precisely to form consensus around proposals, thus *explicitly inducing a form of centralization for the second round*. As such, two-round voting is, by design, a clean, natural playground to test a decentralization metric: we can independently compute the metric on (1) all off-chain, temperature-check proposals, and (2) all on-chain, second round proposals, and confirm whether the metric reports higher centralization for the second type.

Of all DAOs with two-round voting, *Uniswap* serves as a particularly compelling test case for this experiment. First, as a mature DAO with years of governance history, Uniswap has one of the largest DAO treasuries and readily-available proposal history [5]. Second, Uniswap follows a straightfoward, unicameral legislative structure with UNI token holders serving as the sole decision-making body for all proposals [12]. This is unlike other major DAOs that follow a bicameral process, such as Arbitrum's Security Council [15] and Optimism's Citizen House [25].

Uniswap also explicitly defines the purpose of off-chain temperature checks to signal community sentiment prior to an on-chain vote, and the on-chain vote should incorporate feedback from prior voting rounds [4, 12]. Because of this governance mechanism design, we can reasonably infer that there would be more agreement in the second-round on-chain vote versus the off-chain vote.

oVBE validation. We performed this experiment on Uniswap using oVBE as a decentralization metric, and confirm that oVBE *reports a higher level of decentralization in the first*

round of voting. We computed oVBE using K-means for clustering and both min-entropy and Shannon entropy, in windows of 10 proposals across Uniswap data on Snapshot (round 1) and Tally (round 2). We then averaged these values to obtain the average oVBE for each voting round. The first round of voting resulted in an oVBE of 0.7804 (min-entropy) and 1.3149 (Shannon entropy), while round two resulted in 0.6601 (min-entropy) and 1.1527 (Shannon entropy).

Importantly, note that, unlike oVBE, prior decentralization metrics would *not* pass this empirical validation test, since the centralization induced by the first round of voting is orthogonal to account balances. In particular, if all token balances stay the same, these metrics would report that decentralization did not change.

5.3 The oVBE Landscape

After our empirical validation of oVBE, our next goal was to conduct a measurement study of DAOs, to understand how decentralization compares across the DAO ecosystem. We discuss the main steps of this process below. For a more extensive account of our methodology, we refer readers to the documentation in our open-source oVBE artifact [2].

Step #0: corpus of DAOs. We first assembled a corpus of target DAOs to include in our analysis. We defined qualifying DAOs for the study as those with the top 20 largest treasury sizes, at least 25 recorded proposals, over 5,000 unique voters, and data availability in open source APIs for Snapshot and Tally. Narrowing down this list was critical for resource management, data quality, and applicability of the study. The final list included 34 DAOs.

Step #1: data collection. Once we had a list of target DAOs, the next step consisted of scraping the voting history and token holdings of all members of each DAO, to use as inputs to oVBE. Acquiring this data turned out to be a surprisingly difficult task. As prior work also points out, "in practice it is not trivial to acquire all governance related information from raw blockchain data" [33]. The greatest challenge we faced in this step was navigating the heterogeneity of voting data formats across DAOs, which required us to create data extraction scripts to assemble a full dataset of voting history and incorporate manually added metadata; we make these scripts available as part of our oVBE toolkit. Collecting the governance data of all 34 DAOs from our study took approximately 8 hours using a typical laptop.

Step #2: computing oVBE. Once we had a dataset with the voting history of each DAO, we used our toolkit to compute oVBE across proposal windows. We used K-means with $k = 3$ as our clustering function, since align with the nature of voting choices which are normally "Yes," "No," and "Abstain." In addition, we used Euclidean distance as the distance metric for clustering, a rolling window of 10 proposals, and min-entropy and Shannon entropy as our entropy measures. For comparison, we additionally computed other popular decentralization metrics for DAOs, such as the Gini Nakamoto coefficients. Performing all these computations took approximately 3.5 hours using a typical laptop. In our study, as little as 10 proposals sufficed for clustering. Exploring the theoretically-minimum number of proposals required to compute oVBE is an interesting direction for future work.

Step #3: results and analysis. We display the results of our measurement study in a public dashboard, which can be found at [71]. We also show a summary of a few DAOs in Table 1. Our dashboard updates weekly to show the evolving VBE landscape. The focal point of our dashboard is the oVBE of each DAO. The dashboard first presents an overview of all DAOs, with high-level statistics such as proposal count, unique voters, voter participation, oVBE, Gini and Nakamoto coefficient. The DAO-specific view displays current, average, max, and min oVBE, as well as a graph of oVBE fluctuations. Finally, the proposal view specifies vote counts and voting power allocated to proposal choices, discussion URL, description, and outcomes by voter power and by vote count.

Dashboard use cases. The main use case of our dashboard is to track and monitor the oVBE of a particular DAO and understand shifts in decentralization over time. This can help reason about the overall "health" of a DAO, and investigate potential forces in proposals or voter behavior that may change oVBE internally in an organization. For example, a stakeholder may want to understand the cause of a spike or drop in oVBE and whether it aligns with their expectations. These metrics and understanding of an organization can be used to improve a DAO's governance structure, build safeguards to protect against governance attacks, or serve as an alert function for risks or suspicious behavior. We note, however, that a limitation of oVBE is that, a priori, it does not provide direct insights on *cross-DAO* comparisons.

As a second example, our dashboard can also help practitioners refine delegation measures and governance mechanisms. Clustered voting blocs can allow voters to easily identify delegates that align with their preferences, reducing this barrier of entry to delegation and thus increase voter participation in DAOs. Furthermore, many DAOs today implement measures to increase decentralization by redistributing voting power, such as delegating large amounts of voting power to a council of independent delegates [3]. DAOs can dynamically change the amount of voting power they provide to these councils in order to stay within a target range of oVBE.

5.4 oVBE in Governance Experiments

We discuss another application of oVBE in this section: as a decentralization metric in governance experiments. oVBE can be a useful metric to, for example, empirically test hypotheses about the impact of mechanism choices on the decentralization of a DAO. In particular, such governance experiments can be used to, e.g., empirically verify the impact of mechanisms

DAO	Avg. VBE	Std. dev. of VBE	Treasury	Category	Unique Voters	Total Proposals	Avg Voter Participation	Nakamoto Coefficient	Gini Index
Optimism	0.9929	0.2461	$2,800.00M	Infrastructure	221,066	222	20.18%	20	0.997135224
Arbitrum	0.9254	0.2374	$2,100.00M	Infrastructure	283,300	544	19.18%	24	0.994582646
Nouns DAO	0.8806	0.2493	$13.30M	DAO Tool	1,059	563	2.97%	35	0.779029699
Gitcoin	0.8536	0.2417	$42.20M	Funding	11,903	141	9.87%	16	0.994281210
Uniswap	0.7435	0.2161	$2,500.00M	DeFi	68,240	328	9.56%	28	0.998941837
Decentraland	0.5995	0.2544	$93.80M	Gaming, NFTs	1,244	285	4.62%	6	0.972266008
Aave	0.5394	0.2320	$138.90M	DeFi	9,565	398	0.70%	9	0.995135576

Table 1: Summary of a subset of the DAOs included in our measurement study of oVBE calculated with min-entropy.

designed to increase the security of a DAO by increasing decentralization (Section 3.2). While governance experiments can use any decentralization metric as their outcome, oVBE has the potential to be particularly informative due to its sensitivity to social dynamics like herding and apathy. We describe this application through the lens of a real-world case study, in the form of an ongoing collaboration with the Optimism Foundation.

Example: collaboration with Optimism. Optimism [11] is a collective that runs the popular RetroPGF [69] program, an initiative based on the idea of *retroactively rewarding* projects that provide positive value to the community. Every funding round, a set of chosen voters determine the "impact" of each nominated project, by submitting ballots in the form of a distribution of some total dollar amount across the set of nominated projects. These ballots are then aggregated to determine the monetary reward for each project. To date, Optimism has had four rounds of funding, allocating over $100 million to various projects [26].

Maintaining a healthy level of decentralization is of central importance for RetroPGF and its security, as collusion between voters could lead to millions of dollars of misallocated funds. For this reason, Optimism is performing an experiment on a future round of RetroPGF, to understand whether changes to their governance structure would lead to higher decentralization. In particular, their goal is to test whether a *random selection of voters* (sampled from the Optimism ecosystem) leads to a more decentralized funding round than *web of trust*, which is their current mechanism to select voters. This hypothesis will be tested by comparing the casted ballots of two groups—a set of 100 voters selected at random, and a set of 100 voters selected via web of trust—using some decentralization metric to quantify the alignment of each group. This experiment would help inform whether random sampling could lead to a more heterogeneous group of voters.

Optimism has decided to use oVBE as their decentralization metric to use for this upcoming experiment. We worked with them to adapt oVBE to their particular governance structure, showcasing oVBE's ability to mold itself to a variety of governance structures. The main challenge in this process was selecting an instantiation of oVBE that matched their non-standard voting structure (i.e., where ballots are vectors of integers instead of a single, binary number). To address this, we adapted clustering functions used in other domains, where inputs are also vectors of higher dimensions, such as document clustering tasks [45]). This is based on the insight that, conceptually, we can think of a retroPGF ballot as analogous to a *document in the bag-of-words model*, and thus we can leverage document similarity metrics for the clustering of voters. At a high-level, these metrics tend to follow a typical template, where documents are first pre-processed to normalize for common words, followed by a clustering algorithm (such as K-means or hierarchical clustering) on the input vectors, which under-the-hood relies on a distance metric for the vectors, e.g., Euclidean distance or cosine similarity. In our case, the pre-processing step would correspond to normalizing for highly popular projects.

The result of Optimism's experiments are pending. xHowever, our process of working with them serves as a case-study of oVBE utility as a metric for governance experiments, and its flexibility to meet a wide range of requirements and unique settings.

6 Summary Guidance for DAOs

Our VBE framework and the theoretical implications we show in Section 4 suggest a number of forms of concrete guidance for DAOs seeking to enforce or improve meaningful decentralization. We summarize our guidance for practitioners in Table 2.

Apathy / inactivity whale and delegation. One way to diminish the size of the inactivity whale is through delegation. Intuitively, if tokens associated with the inactivity whale are distributed between at least two delegatees in distinct clusters, then they come to represent distinct utility functions—and thus contribute to decentralization. Our theorems also show that when the inactivity whale is large—with respect to delegatees—delegation increases decentralization. (Otherwise, delegation may or may not have this effect.)

Herding / voting privacy. Herding arises because votes are publicly visible. Voting privacy in principle alleviates such pressure and therefore has a decentralizing effect. Snapshot, a popular platform for DAO voting, has recently implemented

Topic	*General Guidance*	*Reason*	*Relevant result*
1. **Vote delegation**	Given a large inactivity whale, vote delegation tends to increase decentralization.	Delegation increases decentralization by diversifying tokens away from a big inactivity whale.	Thm. A.2
2. **Voting privacy**	Voting privacy increases decentralization.	Private voting eases herding, whose effects are centralizing.	Thm. A.3
3. **Voter bribery**	The scale of bribery increases with decentralization.	Low alignment of utility functions means systemic coordination is required to impose alignment.	Thms. A.5, A.6, and A.7
4. **Identity verification**	Weak identity verification increases centralization in quadratic voting.	A whale that can spread tokens across identities amplifies its voting power.	Analysis in Section 4.1
5. **Voting slates / proposal bundling**	Bundling choices into slates (like protocol upgrades that include many voting issues in one package) decreases decentralization.	Bundled choices artificially align otherwise heterogeneous utility functions and/or induce apathy by smoothing out utility functions.	Thm. A.4
6. **Data collection**	Careful voting-statistic collection facilitates decentralization measurement.	Lack of systematic collection and publication of detailed voting statistics makes decentralization measurement challenging today.	Discussion in Section 5.

Table 2: Guidance implied by this paper's results regarding DAO decentralization.

a form of privacy called *shielded voting* [74]. This form of privacy, however, is only ephemeral: Votes are private when submitted, but revealed at the end of the vote-casting period. So it is unclear that it can fully address the centralizing effects of herding. End-to-end verifiable voting systems have been proposed in the literature for decades that achieve both voting integrity and confidentiality [13]. How to implemented them with token-based weighting is, to the best of our knowledge, though, an open problem.

Voter bribery. DAOs today are largely centralized [20, 33, 73]. Bribery may not be especially useful, as whales generally exert strong control and require relatively little coordination to align utility functions into a favorable voting bloc. Voter bribery, however, is a problem in many settings, both in political voting [64] and in corporate governance (see, e.g., [72]). One implication of our results is that as DAO decentralization increases, in order for bribery to succeed, it will need to be systemic. DAO designers should therefore recognize large-scale bribery as a future risk.

Voting slates / bundling proposals. As the practice of bundling proposals / measures has the goal of aligning utility functions, from the standpoint of VBE, it generally has a centralizing effect. DAOs may therefore wish to consider limiting the practice and instead explore way to unbundle multi-component proposals.

Data collection. As discussed in Section 5, we found it challenging to collect full voting histories even for popular DAOs. A recommendation for the community is to establish and adhere to standards for archiving DAO voting data.

Future work: practical insights from LHD. In Section 2, we introduce the Learning Hypothesis of DAOs (LHD), a conceptual model for DAO decentralization, which we use to derive VBE. Yet, we believe LHD has profound (practical) applications to DAO decentralization beyond just VBE. For example, techniques to increase diversity in MARL could form the basis for the development of mechanisms aimed at increasing DAO decentralization. Similarly, LHD could be used to motivate existing approaches to decentralization that otherwise lack principled motivation. For example, existing proposals for rewarding the use of voting power (by voting or delegating) is akin to diversity-boosting in MARL via tailoring of reward functions. Exploring these (and other applications of LHD) is an interesting direction for future work.

7 Related Work

DAOs. Research literature on DAOs has been limited to date, but has included measurement studies [33, 73], retrospectives on the failure of The DAO (e.g., [30]) and ways of addressing related technical flaws in smart contracts such as dangerous reentrancy (e.g., [61]), DAO mechanism design (e.g., [16]), and exploration of DAOs from the standpoint of legal theory (e.g., [43]) and economics and governance (e.g., [17]). Works exploring measurement of DAOs' degree of decentralization most notably include Feichtinger et al. [33], who explore Gini and Nakamoto indices, as well as participation rates and the monetary cost of governance, Sharma et al. [73], who consider various notions of entropy, participation rate, and graph-based measures of decentralization, and [79], which taxonomizes DAOs by comparison with other autonomous systems. Sun et al. use clustering to identify voting blocs in a study of MakerDAO [75]. Also of note is the informal notion of "credible neutrality," a community standard articulated in, e.g., [21, 22].

Social choice and voting theory. A long line of work on social choice and voting theory investigates how best to aggregate preferences of individual voters (see, e.g, [34, 52] for overviews)—the same functionality that DAOs seek to provide in the decentralized setting. There are some major differences between the classical and DAO settings, however. For instance, the permissionless nature of DAOs—and use of token weighting—changes the nature and meaning of voter participation. Further, while the threat of large-scale voter bribery is typically safe to ignore in classical voting, both due to the high likelihood of detecting such an attack, as well as the the challenge in coordinating the attack itself, it is a realistic threat in DAOs, as we have explained.

8 Conclusion

We have introduced Voting-Bloc Entropy (VBE), a metric for DAO decentralization stemming from a new model of DAOs that derives from multi-agent reinforcement learning (MARL). VBE 's sensitivity to aligned behavior among voters makes it a powerful tool, yielding insights well beyond those of existing token-distribution-based metrics. VBE yields both

theoretical results and practical guidance for DAO communities, particularly through its observable variant oVBE. We envision that the open-source dashboard and analysis tools we introduce here will serve as springboards for principled future research and practices that enhance the governance efficacy and security of DAOs.

9 Extra: Ethics and Open-Science

In this section, we discuss the ethical considerations of our work, and our compliance with the open-science policy.

Ethical considerations. Our research starts with theoretical principles to derive a decentralization metric (VBE) with the goal of advancing a principle embraced by the DAO community at which our work is directed. We perceive no ethical dilemmas in that portion of our work. Additionally, based on public data, we have published a dashboard reporting on VBE in current DAOs. We believe that there is a compelling deontological motivation for such publication: researchers have a duty to surface information that sheds light on features of DAOs that reflect their security properties and conformance with publicly stated objectives.

The consequentialist perspective raises the question or concern of whether reputational harm might arise for DAOs that score poorly on this metric. With this consideration in mind, we offer a clear statement of the limitations of our metric and also offer a number of alternative metrics and statistics on our dashboard. Conversely, however—and more importantly given the potential scope of impact and the irremediable consequences—users that engage unknowingly with DAOs exhibiting poor decentralization can incur any of a number of real harms that have already occurred in practice. These harms range from disenfranchisement to a loss of funds resulting from a treasury raid. We conclude that there is a strong ethical argument in favor of the transparency our dashboard will provide the community.

Compliance with the open-science policy. We have released the tools we used for our results in the form of an open-source toolkit, accompanied by detailed documentation and an in-depth report of our methodology [2].

References

[1] Decentralized autonomous organizations (DAOs). https://ethereum.org/en/dao/.

[2] Voting Block Entropy (VBE) Repository. https://zenodo.org/records/14675832.

[3] Anticapture commission. https://gov.optimism.io/t/anticapture-commission/6889, 2023.

[4] Uniswap community governance process update [Jan 2023]. https://gov.uniswap.org/t/community-governance-process-update-jan-2023/19976, 2023.

[5] DeepDAO. https://deepdao.io/, Referenced Aug. 2024.

[6] AvocadoDAO. https://www.avocadodao.io, Referenced Oct. 2023.

[7] GuildFi. https://guildfi.com, Referenced Oct. 2023.

[8] MolochDAO: The original grant giving DAO. https://molochdao.com, Referenced Oct. 2023.

[9] Nakamoto coefficient: An accurate indicator for blockchain decentralization? Bybit Learn, Referenced Oct. 2023.

[10] Uniswap: Governance. https://uniswap.org/governance, Referenced Oct. 2023.

[11] Optimism collective. https://www.optimism.io/, Referenced Sept. 2024.

[12] Uniswap governance process. https://docs.uniswap.org/concepts/governance/process, Referenced Sept. 2024.

[13] Syed Taha Ali and Judy Murray. An overview of end-to-end verifiable voting systems. *Real-World Electronic Voting*, 2016.

[14] Noga Alon, Moshe Babaioff, Ron Karidi, Ron Lavi, and Moshe Tennenholtz. Sequential voting with externalities: herding in social networks. In *EC*, page 36, 2012.

[15] Arbitrum DAO: A conceptual overview. https://docs.arbitrum.foundation/concepts/arbitrum-dao, Referenced Oct. 2023.

[16] Maryam Bahrani, Pranav Garimidi, and Tim Roughgarden. When bidders are DAOs. *arXiv preprint arXiv:2306.17099*, 2023.

[17] Roman Beck, Christoph Müller-Bloch, and John Leslie King. Governance in the blockchain economy: A framework and research agenda. *Journal of the association for information systems*, 2018.

[18] Matteo Bettini, Ajay Shankar, and Amanda Prorok. Heterogeneous multi-robot reinforcement learning. *arXiv preprint arXiv:2301.07137*, 2023.

[19] Vitalik Buterin. Bootstrapping a decentralized autonomous corporation: part I. *Bitcoin Magazine*, 2013.

[20] Vitalik Buterin. Moving beyond coin voting governance, Aug. 2016.

[21] Vitalik Buterin. Credible neutrality as a guiding principle. https://nakamoto.com/credible-neutrality/, Jan. 2020.

[22] Vitalik Buterin. What do I think about Community Notes? https://vitalik.ca/general/2023/08/16/communitynotes.html, Oct. 2023.

[23] Saudu Clement. A list of possible solutions to DAO voter apathy. DAO Times, Dec. 2022.

[24] Douglas R Cole. E-proxies for sale–corporate vote-buying in the internet age. *Wash. L. Rev.*, 2001.

[25] Optimism Collective. Welcome to the optimism collective. https://community.optimism.io/welcome/welcome-overview, Referenced Sept. 2024.

[26] Cryptoknowmics. Optimism's retropgf round 3 distributes over $100 million in public goods funding. https://www.cryptoknowmics.com/news/optimisms-retropgf-round-3-distributes-over-100-million-in-public-goods-funding, Referenced Sept. 2024.

[27] Phil Daian. Vote buying, on-chain governance, and quadratic plutocracy. https://web.archive.org/web/20210122070951/https://pdaian.com/blog/vote-buying-on-chain-governance-and-quadratic-plutocracy/, June 2018.

[28] Maya Dotan, Aviv Yaish, Hsin-Chu Yin, Eytan Tsytkin, and Aviv Zohar. The vulnerable nature of decentralized governance in defi. In *Proceedings of the 2023 Workshop on Decentralized Finance and Security*, 2023.

[29] John Duffy and Margit Tavits. Beliefs and voting decisions: A test of the pivotal voter model. *American Journal of Political Science*, 2008.

[30] Quinn DuPont. Experiments in algorithmic governance: A history and ethnography of "the DAO," a failed decentralized autonomous organization. In *Bitcoin and beyond*, pages 157–177. Routledge, 2017.

[31] Victor Eluke. DeFi bribes: The battle for liquidity supremacy. Blocverse Blog, 25 Sept. 2023.

[32] Rainer Feichtinger, Robin Fritsch, Lioba Heimbach, Yann Vonlanthen, and Roger Wattenhofer. SoK: Attacks on DAOs. In *AFT*, 2024.

[33] Rainer Feichtinger, Robin Fritsch, Yann Vonlanthen, and Roger Wattenhofer. The hidden shortcomings of (D)AOs–an empirical study of on-chain governance. *arXiv preprint arXiv:2302.12125*, 2023.

[34] Peter C Fishburn. *The theory of social choice*. Princeton University Press, 2015.

[35] Wikimedia Foundation. Gini coefficient. https://en.wikipedia.org/wiki/Gini_coefficient, Referenced Sept. 2024.

[36] Wikimedia Foundation. Latent and observable variables. https://en.wikipedia.org/wiki/Latent_and_observable_variables, Referenced Sept. 2024.

[37] Wikimedia Foundation. Multi-agent reinforcement learning. https://en.wikipedia.org/wiki/Multi-agent_reinforcement_learning, Referenced Sept. 2024.

[38] Wikimedia Foundation. Reinforcement learning. https://en.wikipedia.org/wiki/Reinforcement_learning, Referenced Sept. 2024.

[39] Wikimedia Foundation. Utility. https://en.wikipedia.org/wiki/Utility, Referenced Sept. 2024.

[40] Wikimedia Foundation. Von neumann–morgenstern utility theorem. https://en.wikipedia.org/wiki/Von_Neumann-Morgenstern_utility_theorem, Referenced Sept. 2024.

[41] Robin Fritsch, Marino Müller, and Roger Wattenhofer. Analyzing voting power in decentralized governance: Who controls DAOs? *arXiv preprint arXiv:2204.01176*, 2022.

[42] Maximiliano González, Renato Modernell, and Elisa París. Herding behaviour inside the board: An experimental approach. *Corporate Governance: An International Review*, 2006.

[43] Samer Hassan and Primavera De Filippi. Decentralized autonomous organization. *Internet Policy Review*, 10(2):1–10, 2021.

[44] Siyi Hu, Chuanlong Xie, Xiaodan Liang, and Xiaojun Chang. Policy diagnosis via measuring role diversity in cooperative multi-agent rl. In *ICML*. PMLR.

[45] Anna Huang et al. Similarity measures for text document clustering. In *NZCSRSC*, 2008.

[46] What is a rug pull and how to avoid it? Coinbase, https://www.coinbase.com/learn/tips-and-tutorials/what-is-a-rug-pull-and-how-to-avoid-it.

[47] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *ICML*, 2019.

[48] Miles Jennings. Machiavelli for DAOs: principles for fixing decentralized governance. a16z blog post, https://a16zcrypto.com/posts/article/machiavelli

-principles-dao-decentralized-governance/, 21 Sept. 2023.

[49] Christoph Jentzsch. Decentralized autonomous organization to automate governance. *White paper, November*, 2016.

[50] Jiechuan Jiang and Zongqing Lu. The emergence of individuality. In *ICML*, 2021.

[51] Dimitris Karakostas, Aggelos Kiayias, and Christina Ovezik. Sok: A stratified approach to blockchain decentralization. *arXiv preprint arXiv:2211.01291*, 2022.

[52] Jerry S Kelly. *Social choice theory: An introduction*. Springer Science & Business Media, 2013.

[53] Steven P Lalley and E Glen Weyl. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, pages 33–37. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2018.

[54] Jinho Lee, Raehyun Kim, Seok-Won Yi, and Jaewoo Kang. Maps: Multi-agent reinforcement learning-based portfolio management system. *arXiv preprint arXiv:2007.05402*, 2020.

[55] Youngwoon Lee, Jingyun Yang, and Joseph J Lim. Learning to coordinate manipulation skills via skill behavior diversification. In *ICLR*, 2019.

[56] Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. *NeurIPS*, 2021.

[57] Xiangyu Liu, Hangtian Jia, Ying Wen, Yujing Hu, Yingfeng Chen, Changjie Fan, Zhipeng Hu, and Yaodong Yang. Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. *NeurIPS*, 2021.

[58] Zongkai Liu, Chao Yu, Yaodong Yang, Zifan Wu, Yuan Li, et al. A unified diversity measure for multiagent reinforcement learning. *NeurIPS*, 2022.

[59] Thomas Lloyd, Daire O'Broin, and Martin Harrigan. Emergent outcomes of the vetoken model. In *2023 IEEE international conference on omni-layer intelligent systems (COINS)*, pages 1–6. IEEE, 2023.

[60] Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *ICML*, 2021.

[61] Loi Luu, Duc-Hiep Chu, Hrishi Olickel, Prateek Saxena, and Aquinas Hobor. Making smart contracts smarter. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 254–269, 2016.

[62] Shaurya Malwa. Attacker takes over Tornado Cash DAO with vote fraud, token slumps 40%. *CoinDesk*, 21 May 2023.

[63] Charles F Manski. Ordinal utility models of decision making under uncertainty. *Theory and Decision*, 1988.

[64] Robert W McGee and Yanira Petrides. How often are voters bribed? a ranking of 82 countries. In *The Ethics of Bribery: Theoretical and Empirical Studies*, pages 367–384. Springer, 2023.

[65] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325*, 2020.

[66] Kevin R McKee, Joel Z Leibo, Charlie Beattie, and Richard Everett. Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 2022.

[67] Johnnatan Messias, Vabuk Pahari, Balakrishnan Chandrasekaran, Krishna P Gummadi, and Patrick Loiseau. Understanding blockchain governance: Analyzing decentralized voting to amend defi smart contracts. *arXiv preprint arXiv:2305.17655*, 2023.

[68] David M Messick and Keith P Sentis. Estimating social and nonsocial utility functions from ordinal data. *European Journal of Social Psychology*, 1985.

[69] Optimism. Retroactive public goods funcding. https://app.optimism.io/retropgf, Referenced Sept. 2024.

[70] Nicolas Perez-Nieves, Yaodong Yang, Oliver Slumbers, David H Mguni, Ying Wen, and Jun Wang. Modelling behavioural diversity for learning in open-ended games. In *ICML*, 2021.

[71] Tableau Public. https://public.tableau.com/app/profile/dao.vbe/viz/DAOVBE/DAOOverview, 2024.

[72] Jack Schickler. Creditors accuse genesis of ballot-stuffing over $175m FTX deal. *CoinDesk*, 1 Sept. 2023.

[73] Tanusree Sharma, Yujin Potter, Kornrapat Pongmala, Henry Wang, Andrew Miller, Dawn Song, and Yang Wang. Unpacking how decentralized autonomous organizations (daos) work in practice. In *ICBC*, 2024.

[74] Snapshot Labs. Shielded voting is live! https://snapshot.mirror.xyz/yGz91njKbw-sXsnAT6RkoMzPwvuddZritz37h1OWO8o, 13 Oct. 2022.

[75] Xiaotong Sun, Xi Chen, Charalampos Stasinakis, and Georgios Sermpinis. Multiparty democracy in decentralized autonomous organization (DAO): Evidence from MakerDAO. *arXiv preprint arXiv:2210.11203*, 2022.

[76] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[77] Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. Roma: Multi-agent reinforcement learning with emergent roles. *ICML*, 2020.

[78] E Glen Weyl. The robustness of quadratic voting. *Public choice*, 2017.

[79] Steven A Wright. Measuring DAO autonomy: Lessons from other autonomous systems. *IEEE Transactions on Technology and Society*, 2(1):43–53, 2021.

[80] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of data science*, 2015.

[81] Yaodong Yang, Jun Luo, Ying Wen, Oliver Slumbers, Daniel Graves, Haitham Bou Ammar, Jun Wang, and Matthew E Taylor. Diverse auto-curriculum is critical for successful real-world multiagent learning systems. *arXiv preprint arXiv:2102.07659*, 2021.

[82] Yaodong Yang, Ying Wen, Jun Wang, Liheng Chen, Kun Shao, David Mguni, and Weinan Zhang. Multi-agent determinantal q-learning. In *ICML*, 2020.

[83] Martin Young. 'golden boys' behind Compound 'governance attack' agree to rescind proposal. *CoinTelegraph*, 30 July 2024.

[84] Yang Zhang, Qingyu Yang, Dou An, and Chengwei Zhang. Coordination between individual agents in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[85] Rui Zhao, Jinming Song, Yufeng Yuan, Haifeng Hu, Yang Gao, Yi Wu, Zhongqian Sun, and Wei Yang. Maximum entropy population-based training for zero-shot human-ai coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[86] et al. Zhou, Ming. Smarts: An open-source scalable multi-agent rl training school for autonomous driving. In *Proceedings of the 2020 Conference on Robot Learning*. PMLR.

A Additional Theorems and Proofs

In this section, we show the rest of our theorems and proofs from the theoretical implications of VBE (Section 4). Due to space constraints, we defer some of our proofs to the extended version of the paper.

Theorem A.1 (Apathy and VBE)**.** Let $(E, U'_{E,\mathcal{P}}, \mathsf{tokens}) = T_{\mathtt{apath}}(\mathcal{P}, E, U_{E,\mathcal{P}}, \mathsf{tokens})$ be the transformation where players $\hat{\mathcal{P}} \subseteq \mathcal{P}$ become ε-apathetic, i.e., $\forall P \in \hat{\mathcal{P}}\ \forall e \in E$, $|\mathsf{util}'_P(e, \mathtt{true})| \leq \varepsilon$. The rest of the system remains unchanged. Then, if $\forall P \in \hat{\mathcal{P}}$, $\mathsf{tokens}(\mathcal{A}) \geq \mathsf{tokens}([P])$, it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}) \geq \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U'_{E,\mathcal{P}}, \mathsf{tokens}).$$

Proof. Let B be the largest voting bloc by token holdings before $T_{\mathtt{apath}}$. We first note that all apathetic voters belong to the same voting bloc B', according to C: by the definition of ε-apathetic, it follows that, for all $P_i, P_j \in \hat{\mathcal{P}}$ and $e \in E$,

$$|\mathsf{util}'_{P_i}(e, \mathtt{true})|, |\mathsf{util}'_{P_j}(e, \mathtt{true})| \leq \varepsilon,$$

which corresponds precisely to the bloc of apathetic voters in C, containing all players in $\mathcal{A}$. Then, by assumption, $\mathsf{tokens}(B') = \mathsf{tokens}(\mathcal{A}) \geq \mathsf{tokens}([P])$, $\forall P \in \hat{\mathcal{P}}$. So, since no other blocs decrease in size, it follows that $\mathsf{tokens}(B') \geq \mathsf{tokens}(B)$: either the bloc that aggregates all apathetic voters is now the largest bloc, or the same bloc is the largest in both instances. Thus, from Theorem 4.1, it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}) \geq \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U'_{E,\mathcal{P}}, \mathsf{tokens})$$

as desired. □

Theorem A.2 (Delegation and VBE)**.** Let $(E, U'_{E,\mathcal{P}}, \mathsf{tokens}') = T_{\mathtt{deleg}}(\mathcal{P}, E, U_{E,\mathcal{P}}, \mathsf{tokens})$ be the transformation where players $\hat{\mathcal{P}} \subseteq \mathcal{P}$, who are ε-apathetic, instead delegate their votes to a set of delegates $D \subset \mathcal{P}$, i.e., $\mathsf{tokens}'(D) = \mathsf{tokens}(D) + \mathsf{tokens}(\hat{P})$ and $\mathsf{tokens}'(\hat{\mathcal{P}}) = 0$. The rest of the system remains unchanged. Then, if $\forall d \in D$, $\mathsf{tokens}(\mathcal{A}) \geq \mathsf{tokens}'([d])$, it follows that,

$$\mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U'_{E,\mathcal{P}}, \mathsf{tokens}') \geq \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}).$$

Proof. Let B by the largest voting bloc by token holdings before $T_{\mathtt{deleg}}$. As discussed in the proof of Theorem A.1, all players in $\hat{P}$ belong to the same voting bloc for all elections in E—the inactivity whale—since they are all part of the set of apathetic voters $\mathcal{A}$. Let B' be the largest voting bloc by token holdings after $T_{\mathtt{deleg}}$; note that it may be the case that $B' = [d]$ for some $d \in D$.

We first note that B' is equal to either (1) B itself, (2) the second largest voting bloc after B before delegation, or (3) $[d]$, for some $d \in D$. That is, since the only blocs that change after $T_{\mathtt{deleg}}$ are all the $[d]$ and the inactivity whale (which lost $\mathsf{tokens}(\hat{P})$ tokens), it must the be case that the new largest voting bloc is either the same one as before delegation, the second largest voting bloc before delegation (i.e., B was the

inactivity whale, which got fractionated by delegation), or one of the $[d]$ which increased in size.

For (1) and (2), it is clearly the case that $\mathsf{tokens}(B) \geq \mathsf{tokens}'(B')$. Then, for (3), note that, by assumption, $\mathsf{tokens}(\mathcal{A}) \geq \mathsf{tokens}'([d])$, for all $d \in D$. So, $\mathsf{tokens}(B) \geq \mathsf{tokens}(\mathcal{A}) \implies \mathsf{tokens}(B) \geq \mathsf{tokens}'([d]) = \mathsf{tokens}'(B')$.

It follows then that, in all cases, $\mathsf{tokens}(B) \geq \mathsf{tokens}'(B')$. Thus, from Theorem 4.1, we get that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U'_{E,\mathcal{P}},\mathsf{tokens}') \geq \mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U_{E,\mathcal{P}},\mathsf{tokens})$$

as desired. □

Theorem A.3 (Herding and VBE)**.** Let $(E,U'_{E,\mathcal{P}},\mathsf{tokens}) = T_{\texttt{herd}}(\mathcal{P},E,U_{E,\mathcal{P}},\mathsf{tokens})$ be the transformation where players $\hat{\mathcal{P}} \subseteq \mathcal{P}$ exhibit herding toward, without loss of generality, `true`. That is, for all $P \in \hat{P}$ and $e \in E$, the monetary reputational cost of voting for `false` is greater than or equal to $\max(2 \cdot \mathsf{util}_P(e,\texttt{false}]) + \varepsilon, 0)$ for some constant ε. The rest of the system remains unchanged. Then, it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U_{E,\mathcal{P}},\mathsf{tokens}) \geq \mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U'_{E,\mathcal{P}},\mathsf{tokens}).$$

Proof. Let B be the largest voting bloc by token holdings before $T_{\texttt{herd}}$. Note that, after $T_{\texttt{herd}}$, all voters in $\hat{P}$ belong to the same voting bloc B': for every $P \in \hat{\mathcal{P}}$, $U'_{E,P}$ will consist of only positive values: either P preferred an outcome of `true` in e to begin with, or their monetary utility of `true` is now $|\mathsf{util}_P(e,\texttt{false})| + \varepsilon$. Thus, since $\mathrm{sgn}(\mathsf{util}_P(e,\texttt{true})) = 1$ for all $e \in E$, all of $\hat{P}$ consists of a single voting bloc B' according to C.

It follows then that $\mathsf{tokens}(B') \geq \mathsf{tokens}(B)$, as either the "new" voting bloc B' is now the largest bloc, or the same bloc is the largest before and after $T_{\texttt{mirr}}$. Then, from Theorem 4.1, it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U_{E,\mathcal{P}},\mathsf{tokens}) \geq \mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U'_{E,\mathcal{P}},\mathsf{tokens})$$

as desired. □

Note that the corollary that privacy *increases* decentralization follows directly via a proof by contradiction of Theorem A.3).

Result #5: voting slates. We model a player's utility for a slate of elections simply by adding the utilities of the underlying proposals. That is, for all $P \in \mathcal{P}$ and some election $\mathcal{E}$ comprised of some subset of elections of E, the utility of P in $\mathcal{E}$ is:

$$\mathsf{util}_P(\mathcal{E},\texttt{true}) = \sum_{e \in \mathcal{E}} \mathsf{util}_P(e,\texttt{true}).$$

Voting slates are generally used to "hide" unpopular proposals among a larger set of benign, popular proposals, and thus increase their chances of passing. We model this by saying that if two P_i, P_j have aligned utilities (according to C) on all proposals underlying $\mathcal{E}$, then they will agree on $\mathcal{E}$ itself, i.e., $C_\varepsilon(U_{E,P_i},U_{E,P_j}) = 1 \implies \mathrm{sgn}(\sum_{e \in \mathcal{E}} \mathsf{util}_{P_i}(e,\texttt{true})) = \mathrm{sgn}(\sum_{e \in \mathcal{E}} \mathsf{util}_{P_j}(e,\texttt{true}))$.

Theorem A.4 (Voting Slates and VBE)**.** Let $(E',U'_{E',\mathcal{P}},\mathsf{tokens}) = T_{\texttt{slates}}(\mathcal{P},E,U_{E,\mathcal{P}},\mathsf{tokens})$ be the transformation where all elections E are bundled together into slates to form a smaller set of elections E'. The rest of the system remains unchanged. Then, it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U_{E,\mathcal{P}},\mathsf{tokens}) \geq \mathsf{VBE}_{C_\varepsilon,\min}(E',\mathcal{P},U'_{E',\mathcal{P}},\mathsf{tokens}).$$

Proof. Let B be the largest voting bloc by token holdings before $T_{\texttt{slates}}$. Then, note that all players in B are still in the same voting bloc B' after $T_{\texttt{slates}}$: since $C_\varepsilon(U_{E,P_i},U_{E,P_j}) = 1$ for every pair of players in B, by assumption, it follows that

$$\forall \mathcal{E} \in E',\ \mathrm{sgn}(\sum_{e \in \mathcal{E}} \mathsf{util}_{P_i}(e,\texttt{true})) = \mathrm{sgn}(\sum_{e \in \mathcal{E}} \mathsf{util}_{P_j}(e,\texttt{true})).$$

Conversely, players who did not belong to B may, in fact, join B' after $T_{\texttt{slates}}$: even if the players disagree in some of the underlying proposals for a particular slate $\mathcal{E}$, they may cast the same overall vote for the entire slate. As such, B' contains strictly more players than B, which implies that $\mathsf{tokens}(B') \geq \mathsf{tokens}(B)$. Then, from Theorem 4.1, it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U_{E,\mathcal{P}},\mathsf{tokens}) \geq \mathsf{VBE}_{C_\varepsilon,\min}(E',\mathcal{P},U'_{E',\mathcal{P}},\mathsf{tokens})$$

□

Result #7: bribery. Before presenting our results, we first introduce some additional notation. We define $\mathsf{bribe}_P : E \times \{\texttt{true},\texttt{false}\} \to \mathbb{R}$ to be such that it is possible for player P to achieve an outcome of `true` (resp., `false`) in a particular election e via bribery for any expenditure greater than $\mathsf{bribe}_P(e,\texttt{true})$ (resp., $\mathsf{bribe}_P(e,\texttt{false})$). Note that we make the simplifying assumption that bribery costs are independent across elections. We assume that bribing a given P to flip its vote from `true` to `false` (respectively, `false` to `true`) costs $\max(2 \cdot \mathsf{util}_P(e,\texttt{true}) + \varepsilon, 0)$ (respectively, $\max(2 \cdot \mathsf{util}_P(e,\texttt{false}) + \varepsilon, 0)$), for some constant ε. Successful bribery to achieve an outcome of `true` in e means flipping enough votes to cross a certain threshold q of votes for `true` in e (and vice versa for `false`). For example, a typical value for q may be $q = 0.5$. We note that this represents the threshold to *ensure* the desired outcome in an election, and not just to win it.

Theorem A.5 (Bribery and VBE)**.** Let $(E,U'_{E,\mathcal{P}},\mathsf{tokens}) = T_{\texttt{bribe}}(\mathcal{P},E,U_{E,\mathcal{P}},\mathsf{tokens})$ be the transformation where an entity successfully bribes players $\hat{\mathcal{P}} \subseteq \mathcal{P}$ in elections E to achieve an outcome of, without loss of generality, `true`. The rest of the system remains unchanged. Then, it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U_{E,\mathcal{P}},\mathsf{tokens}) > \mathsf{VBE}_{C_\varepsilon,\min}(E,\mathcal{P},U'_{E,\mathcal{P}},\mathsf{tokens}). \quad (1)$$

Proof. Let B be the largest voting bloc by token holdings before $T_{\texttt{bribe}}$. First, note that, after $T_{\texttt{bribe}}$, all voters in $\hat{\mathcal{P}}$ belong to the same voting bloc B'. Recall that, in our DAO abstraction, bribing a player P to flip its vote in election e from `false` to `true` costs $\max(2 \cdot \mathsf{util}_P(e, \texttt{false}) + \varepsilon, 0)$. So, for every $P \in \hat{\mathcal{P}}$ and $e \in E$, either $\mathsf{util}_P(e, \texttt{true})$ was already positive to begin with, or it is now $|\mathsf{util}_P(e, \texttt{false})| + \varepsilon$. Then, since $\mathrm{sgn}(\mathsf{util}_P(e, \texttt{true})) = 1$ for all $e \in E$, all of $\hat{P}$ consists of a single voting bloc B' according to ε-TOC.

It follows then that $\mathsf{tokens}(B') \geq \mathsf{tokens}(B)$, as either the "new" voting bloc B' is now the largest bloc, or the same bloc is the largest before and after $T_{\texttt{bribe}}$. Then, from Theorem 4.1, it follows that

$$\mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}) \geq \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U'_{E,\mathcal{P}}, \mathsf{tokens})$$

as desired. □

Theorem A.6 (Internal Bribery and VBE)**.** Let $(E, U'_{E,\mathcal{P}}, \mathsf{tokens}) = T_{\texttt{bribe}}(\mathcal{P}, E, U_{E,\mathcal{P}}, \mathsf{tokens})$ be the transformation where $U'_{E,\mathcal{P}}$ is some arbitrary change in the utilities of the players. The rest of the system remains unchanged. Assume that an entity in $\mathcal{P}$ needs to bribe other players holding a total of at least n_1 and n_2 tokens to guarantee an outcome of `true` in elections E before and after $T_{\texttt{bribe}}$, respectively. Then, it follows that

$$\begin{aligned} n_1 > n_2 \iff & \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U'_{E,\mathcal{P}}, \mathsf{tokens}) \\ & < \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}). \end{aligned}$$

This result sheds light on the scale of bribery in the case where the briber is a malicious tokenholder a priori. Conversely, the briber may instead be some external entity. In this case, decentralization also raises the risk of systemic bribery: if there are large players in the system, the briber can directly coordinate with whales to achieve their desired election outcome. If, however, the DAO is highly decentralized, the outcome of the election depends on many stakeholders, which thus requires large-scale coordination among these. More formally:

Theorem A.7 (External Bribery and VBE)**.** Let $(E, U'_{E,\mathcal{P}}, \mathsf{tokens}) = T_{\texttt{bribe}}(\mathcal{P}, E, U_{E,\mathcal{P}}, \mathsf{tokens})$ be the transformation where $U'_{E,\mathcal{P}}$ is some arbitrary change in the utilities of the players. The rest of the system remains unchanged. Let n_1 and n_2 be the minimum number of players that an external entity needs to corrupt to guarantee an outcome of `true` in elections E before and after $T_{\texttt{bribe}}$, respectively. Then, it follows that

$$\begin{aligned} n_1 > n_2 \iff & \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U'_{E,\mathcal{P}}, \mathsf{tokens}) \\ & < \mathsf{VBE}_{C_\varepsilon,\min}(E, \mathcal{P}, U_{E,\mathcal{P}}, \mathsf{tokens}). \end{aligned}$$

Result #7: quadratic voting. Our formalism captures the relationship between quadratic voting and bribery (which has been informally identified by prior work [27]). We define "small" accounts to be, concretely, those whose fraction of the total tokens increases with quadratic voting in place, and thus have their impact amplified. More formally, we denote that a player $P \in \mathcal{P}$ benefits from quadratic voting by $\mathsf{quad}(P, \mathsf{tokens}) = 1$, where

$$\mathsf{quad}(P, \mathsf{tokens}) = 1 \iff \frac{\mathsf{tokens}(P)}{\sum\limits_{p \in \mathcal{P}} \mathsf{tokens}(p)} < \frac{\sqrt{\mathsf{tokens}(P)}}{\sum\limits_{p \in \mathcal{P}} \sqrt{\mathsf{tokens}(p)}}.$$

The relationship between quadratic voting and bribery hinges on whether the cost to bribe a player is the same with or without quadratic voting. Whether quadratic voting changes a player's utility or not will vary across systems. Broadly speaking, if DAO members take governance seriously and are invested in election outcomes, quadratic voting indeed changes utilities: since smaller accounts become more "pivotal" as a result of quadratic voting, their utilities increase correspondingly. Conversely, if members have little interest in governance, the fact that their vote can now have a greater impact in the election will not change their utilities. As such, the nature of a community must be taken into account when deciding to use quadratic voting.

Theorem A.8 (Quadratic Voting and Bribery)**.** Let $(E', U_{E',\mathcal{P}}, \mathsf{tokens}) = T_{\texttt{quad}}(\mathcal{P}, E, U_{E,\mathcal{P}}, \mathsf{tokens})$ be the transformation where all elections E employ quadratic voting. We denote the election corresponding to $e \in E$ by $e' \in E'$. Let f and f' be the fraction of total votes that a bribing entity is able to control for some fixed expenditure t in elections E and E', respectively. Then, it follows that

$$f < f' \iff \exists \hat{\mathcal{P}} \subseteq \mathcal{P} \,|\, \forall P \in \hat{\mathcal{P}},\ \Big(\mathsf{quad}(P, \mathsf{tokens}) = 1 \wedge U_{E,P} = U_{E',P}\Big)$$

This result thus shows that quadratic voting may be favorable for a bribing entity. In particular, if there are enough small voters whose utilities are unchanged, the cost to guarantee successful bribery decreases:

Corollary 2.1. *Assume that, for $\hat{P}$ as defined in Theorem A.8, $\mathsf{tokens}(\hat{P}) > q \cdot \sum_{P \in \mathcal{P}} \mathsf{tokens}(P)$. Let t and t'' be the expenditure required to guarantee an outcome of `true` in elections E and E', respectively. Then, it follows that $t' < t$.*

This corollary simply follows from the fact that, as proved in Theorem A.8, the expenditure t' required to control a fraction of q votes in E', and thus guarantee successful bribery in E', would only be enough to acquire a fraction of $q - \varepsilon$ votes in E. As such, some additional expenditure is required to cross the threshold of q votes.