The cost of sybils, credible commitments, and false-name proof mechanisms

Bruno Mazorra*[1] and Nicolás Della Penna†[2]

[1]Universitat Pompeu Fabra
[2]GroupLang.ai

July 7, 2024

Abstract

Consider a mechanism that cannot observe how many players there are directly, but instead must rely on their self-reports to know how many are participating. Suppose the players can create new identities to report to the auctioneer at some cost c. The usual mechanism design paradigm is equivalent to implicitly assuming that c is infinity for all players, while the usual Sybil attacks literature is that it is zero or finite for one player (the attacker) and infinity for everyone else (the 'honest' players). The false-name proof literature largely assumes the cost to be 0. We consider a model with variable costs that unifies these disparate streams.

A paradigmatic mechanism can be extended into a Sybil extension mechanism by having the action space by the product of the feasible set of identities to create action where each player chooses how many players to present as in the game and their actions in the original normal form game. A mechanism is (dominant) false-name proof if it is (dominant) incentive-compatible for all the players to self-report as at most one identity. We study mechanisms proposed in the literature motivated by settings where anonymity and self-identification are the norms, and show conditions under which they are not Sybil-proof. We characterize a class of dominant Sybil-proof mechanisms for reward sharing and show that they achieve the efficiency upper bound. We consider the extension when agents can credibly commit to the strategy of their sybils and show how this can break mechanisms that would otherwise be false-name proof.

1 Introduction

The internet naturally motivates the study of mechanisms where players can use fake names and bids to improve their outcomes. Two literatures, largely separate, have studied this. Using the nomenclature of "Sybil attacks" and focusing on single-player deviations from a generally truthful population [6, 20, 21, 28, 39]. The other, using the nomenclature of false-name proofness, has focused on the equilibrium of games where players who can miss-represent as multiple ones [1, 11, 34, 35, 37]. In this work, we jointly formulate both streams of the literature and extend the analysis to situations with the potential for commitment. Since the stream of the literature using the term Sybil appears largely unaware of the equilibrium implications of the phenomenon, we use the Sybil term while focusing on equilibrium to help in correcting this.

*Email: brunomazorra@gmail.com
†Email: nikete@gmail.com

The necessary and sufficient condition for a mechanism to be Sybil-proof (equivalently, false-name proof), meaning that players have no incentives to generate Sybils, is that their payoff is no better as they add extra identities is no more than their payoff when they have a single identity. While several mechanisms have been known not to be Sybil-proof when there is no cost to creating new identities (notably VCG, [1, 32, 37]), we generalise the condition to potential costs in the creation of sybils. Motivated by smart contracts and AI agents, we consider the possibility of credible commitment in the strategies of the Sybil players as a natural extension. Under these conditions, we demonstrate that many previously studied mechanisms in the literature are not Sybil-proof [4, 26, 27], and provide examples of mechanisms that are. These are constructed using the *pie shrinking with crowding*. This consists of shrinking the total welfare as more players present themselves, to make the mechanisms Sybil-proof. In this paper, we use this to construct Sybil and truthful for fair-division cake-cutting mechanisms and Sybil-proof bidding rings in second-price auction.

1.1 Motivation example

In many different fields, the Sybil game emerges as a bigger game. One example is when a reward, R, is distributed fairly among participants, and the chances of winning the reward are based on the number of identities participating. In this situation, a strategic actor may attempt to create multiple identities to increase their chances of receiving the reward, and so, increase its expected payoff. More formally, the action space of agents is to present a number of fake identities x to the mechanism. Then, the expected payoff of any strategic player is $U(x, y) = Rx/(x+y)$ where y is the number of identities reported by other players. Observe that U is strictly increasing on x and therefore, there is no optimal strategy. Similarly, if all players are strategic, the game has no Nash equilibrium. However, in general, creating fake identities induce some cost to the attacker. If creating an identity has some associated costs $c > 0$, then the expected payoff of an attacker creating x identities is $U(x, y) = Rx/(x + y) - cx$. In general, $x = 1$ is not a dominant strategy. For example, assume that $R = 10$, $c = 0.1$, and $y = 3$. Then, $U(1, 3) = 10/4 - 0.1 = 2.4$ and $U(2, 3) = 10 \cdot 2/5 - 0.2 = 3.8$. And so, players have incentives to report more than one identity in equilibrium. In recent years, the security of distributed systems has become increasingly important as more and more of our daily lives are conducted online. One type of attack that has received a lot of attention is the Sybil attack, first identified by John Douceur [10]. A Sybil attack is a type of attack in which a single malicious entity creates multiple fake identities in order to manipulate the system and gain an unfair advantage.

Sybil attacks have been studied in a variety of contexts, including peer-to-peer networks [9, 33], online social networks [40, 41], reputation systems [7] blockchain systems [42], combinatorial auctions [1, 38], and diffusion auctions [5]. In a peer-to-peer network, a Sybil attacker may create multiple fake identities in order to control a large portion of the network and launch a denial-of-service attack. In permissionless anonymous environments, a Sybil attacker may create multiple fake identities in order to spread misinformation or influence public opinion. In a blockchain system, a Sybil attacker may create multiple fake identities in order to control a large portion of the network and carry out a 51% attack. In combinatorial and diffusion auctions, a Sybil attacker may create multiple false identities and bid in different bundles to manipulate the outcome of the auction and increase his gains [38].

2 Related Work

To protect against Sybil attacks, researchers have proposed a variety of mechanisms, including unique identifier systems [14, 22, 31], proof-of-work/proof-of-stake systems [3, 18, 29], and reputation systems [17, 40]. However, these mechanisms are not foolproof, and Sybil attacks can

still occur in practice.

In game theory and auction theory, Sybil attacks are usually noted as false-name strategies or shill bids. The first author studying false-name strategies in internet auctions is made in [37, 38]. In [37] the authors present a combinatorial auction that is robust against false-name bids. In [38], M. Yokoo et. al. prove that Vickrey–Clarke–Groves (VCG) mechanism, which is strategy-proof and Pareto efficient when there exists no false-name bid, is not false name-proof/Sybil-proof and there exists no false-name proof combinatorial auction mechanism that satisfies Pareto efficiency. Moreover, they show that if agents have concave valuations, then the VCG mechanism is false-name proof. In [15], the authors analyzed the worst-case efficiency ratio of false-name-proof combinatorial auction mechanisms. The authors show that the worst-case efficiency ratio of any false-name-proof mechanism that satisfies some minor assumptions is at most $2/(m+1)$ for auctions with m different goods.

In the field of non-monetary mechanisms in [35] the authors study false-name-proof mechanisms in the facility location problem. First, the authors fully characterize the deterministic false-name-proof facility location mechanisms in this basic setting. By utilizing this characterization, they show the tight bounds of the approximation ratios of social cost and maximum cost.

In voting false-name proof mechanisms, different studies have been made [11, 36]. In [36], the authors study voting rules where there is a cost for casting a vote. The authors characterize the optimal (most responsive) false-name-proof with-costs voting rule for 2 alternatives. They prove that as the voting population grows larger, the probability that this rule selects the majority winner converges to 1. Also, the authors characterize the optimal group false-name-proof rule for two alternatives. In [11] the authors characterize all voting rules that verify false-name-proofness, strategy-proofness, unanimity, anonymity, and neutrality as either the class of voting by quota one (all voters can be decisive for all objects) or the class of voting by full quota (all voters can veto all objects).

To our knowledge, no previous work has been made in Sybil-proof collusion mechanisms behaviour in auctions (and bidding rings), and general fair allocation mechanisms such as cake-cutting with homogenous and heterogeneous valuations.

2.1 Our contribution

The main contributions of this paper can be summarized as follows:

- We generalize the games where players can represent more than one identity to the underlying game with some cost that depends on the number of identities, which we call the Sybil extension game. In this game, players can create and utilize Sybils to compete in the underlying normal game. Specifically, we provide a mathematical framework for such games based on the costs associated with creating Sybils and the interactions between a player's Sybils and those of other players.

- We identify a necessary and sufficient condition for a mechanism to be Sybil-proof. That is a necessary condition of a mechanism where players have no incentives to generate Sybils. We use this to demonstrate that many previously proposed mechanisms in the literature do not naturally satisfy this condition, and are therefore not Sybil-proof.

- Motivated by proof-stake mechanism and bidding rings in permisionless environments, we analyse collusive behavior in permissionless environments where players want to share some reward R but can create false identities. We call these Reward Distribution Mechanisms (RDM). We find the necessary conditions for an RDM to be a Sybil-proof mechanism and introduce pie shrinking with crowding. Doing so, we find the symmetric, prior-free, budget-independent, and Pareto optimal RDM. We also prove that if players share some

knowledge of the number of players, other mechanisms have strictly greater social welfare in equilibrium can be found. More generally, we study Sybil-proof of cake-cutting mechanisms, proving the upper bound of efficiency of Sybil-proof cake-cutting mechanisms and giving a worst-case welfare optimal truthful cake-cutting mechanism. Also, we study collusive behavior in the second-price auction, where players can submit false name bids and all players have the same distribution of valuations. In this scenario, we prove there is no efficient Sybil-resistant bidding ring, but there are constructive optimal Sybil-resistant collusion mechanisms.

- Motivated by permissionless credible commitment devices, we define Sybil-commitment games. These arise when players can credibly commit to certain to an identity that behave as a rational agent with specific preferences. This can be seen as an extended game of a normal game with two phases. The Sybil-commitment phase where players commit a set of rational independent identities with different preferences, and the normal game phase where the Sybils in the first phase have full information about the game. We defined the Sybil-commitment equilibrium and the Sybil-commitment-proof. We proved necessary conditions on the price of anarchy of the underlying game to be Sybil-proof. Finally, we give examples of games that are Sybil-proof but not Sybil-commitment-proof.

2.2 Organization of the paper

In this paper, we present a comprehensive study of Sybil-proof mechanisms in mechanism design. The paper is organized as follows:

Section 2 introduces the concept of Bayesian games and games with an unknown number of players. We define the Sybil extension game as a symmetric game with incomplete information and the Sybil Nash equilibrium. We provide examples of non-Sybil proof and Sybil-proof games and discuss the reward distribution mechanism, cake cutting with heterogeneous valuations, and permissionless bidding rings in second-price auctions.

Section 3 introduces Sybil games with identity commitments. We define the Sybil commitment Nash equilibrium and provide the necessary conditions for a mechanism to be Sybil-proof with identity commitments. We give examples of non-Sybil with commitment-proof games such as Cournot and pro-rata mechanisms.

Finally, in Section 4, we provide conclusions and discuss future research directions.

Overall, the paper provides a comprehensive framework for studying Sybil-proof property in mechanism design, highlights the importance of considering Sybil attacks in the design of mechanisms for distributed systems, and computes the cost of having Sybil-proof mechanisms in terms of efficiency.

2.3 Notation

In this paper, we introduce various mathematical notations that are used throughout the text. We start denoting by $\mathcal{N}$ the set of players, Θ_i the set of types, by U_i as the utility function of player i, which represents the preferences of player i over different outcomes in a game. The action of player i is denoted by x_i, while the joint action of all players except player i is represented by x_{-i}. We also use the notation $\mathbf{1}$ to represent a vector of n elements with all entries being 1. In the context of group theory, we define S_n as the symmetric group of n elements, which consists of all possible permutations of n elements. Furthermore, S_∞ is the symmetric group of an infinite number of elements. In the realm of vector spaces, we use A^∞ to denote the direct sum of infinite copies of a vector space A. Additionally, we introduce the Sybil cost function, denoted by c, which measures the cost incurred by a system when facing Sybil attacks or the cost for an attacker to create Sybil identities. Lastly, we use the notation $\mathrm{NE}(G)$ to

represent the set of Nash equilibria in a game G, where a Nash equilibrium is a stable state in which no player can unilaterally change their strategy and improve their utility.

3 Sybil extension mechanisms and games

In this section, we formalize the notion of Sybil extension mechanism. Informally, the Sybil extension mechanism of a mechanism is modeled as follows. There are a finite but unbounded number of players that are unknown to both the mechanism designer and the players. As a formality, we call the players can first be categorized in two types of players: active players and inactive players. In some sense the set of players is the total potential number of agents that could participate in the mechanism, for example the total number of people in the planet $N \approx 8.1$ Billion. In general, we assume that N is finite but unbounded. The active players are the ones that are willing to participate in the mechanism, and the passive are the ones that do not participate. To analyse mechanisms with unknown number of agents, two different models will be studied, the Bayesian setting where some information about the number of active agents is known, and the prior-free setting where no assumption is made on the number of active agents. In the Bayesian setting players types are drawn from a common knowledge distribution $\mathcal{D}$. Each player can choose the number of players to report to the game and a vector of strategies of each Sybil identity to the game. The mechanism designer, can not distinguish between "real" and Sybil identities. Each Sybil creation induces some costs to the player. The payoff of each player is the sum of the payoffs of the Sybils minus the costs of creation. A game is Sybil-proof if no player has incentives to create false identities. In some motivating applications of internet-enabled permissionless mechanisms, it appears unlikely for there to be a common prior over types. Thus, the natural solution concept for mechanisms is Sybil-proof and dominant strategy incentive compatible (DSIC). In case there is a common prior, the solution concept will Bayes Nash incentive compatible.

3.1 Anonymous mechanisms

In the field of mechanism design, it is typically assumed that the number of agents n is exogenously determined and that the declaration of agents is not part of the strategic behaviors available to them. However, this assumption does not always hold true. For instance, in eBay auctions, sellers can engage in shill bidding—artificially inflating bids on their own items to increase revenue. Similarly, in scenarios where influencers conduct lotteries based on likes or other forms of interaction, agents have incentives to create fake identities to enhance their chances of winning the lottery item. Another example can be found in online multiplayer games, where players might create multiple accounts to gain unfair advantages in in-game events or competitions. Additionally, in public goods provision, individuals might underreport their true valuations or even create phantom identities to manipulate the outcome in their favor, ensuring higher personal benefits at a lower cost.

Therefore, this motivates a formalization of how mechanisms operate when agents can generate Sybils. Before formalizing the definition of Sybil extension mechanism, let's recall the definition of direct revelation mechanisms. A *direct revelation mechanism* or mechanism for short with agents with quasi-linear utilities consists of a tuple $([n], \Theta, \mathcal{X}, \mathbf{x}, \mathbf{p})$, where n is the total number of agents, $\Theta = \prod_{i=1}^{n} \Theta_i$ is the set of all possible types and Θ_i is the types that an agent i can take, $\mathcal{X}$ is the set of all possible outcomes (e.g. in a combinatorial auction which agents are allocated to whom), the social choice function (or the allocation rule) $\mathbf{x} : \Theta \to \mathcal{X}$ that determinates the outcome given the types reported, and finally the payment rule $\mathbf{p}$ that determinates how much each participant has to pay given the reported types. In this setting, if an agent has type $\theta \in \Theta$, then its valuation over the outcome space is given by a publicly known

function $v : \Theta \times \mathcal{X} \to \mathbb{R}$. For example, in a single item auction, the valuation of an agent is $v(\theta, \mathbf{x}(\mathbf{b})) = \theta \cdot x_i(\mathbf{b})$.

Observe that this definition of direct revelation mechanism does not allow agents to report more than one type by construction. Each agent is assigned one input slot for reporting their type, and so they can not use Sybil strategies by construction. However, in reality, mechanisms are designed in a way that they generalize for a more unspecified number of agents. For example, a first-price auction can be naturally extended for any number of agents by allocating the item to the highest bidder and pays to the seller the amount bidded. In so, in the following, we will define the Sybil extension of a set of mechanisms.

First, we have to define the mechanism with unbounded but finite number of players, called *anonymous mechanism*. A deterministic anonymous mechanism $\mathcal{M}$, consists of a tuple Θ, that models all the possible types that an agent can have, a set of outcomes $\mathcal{X}$ endowed with a group action of S_∞, and a sequence of maps $\{(\mathbf{x}^n : \Theta^n \to \mathcal{X}, \mathbf{p}^n : \Theta^n \to \mathbb{R}_+^n)\}_{n\in\mathbb{N}}$ such that the following two properties hold:

- *Inactive type* The set of types Θ has an element 0 that represents not participating in the mechanism.
- *Anonymity*: The maps $\mathbf{x}^n$ and $\mathbf{p}^n$ are equivariant under the action of S_n, that is, for all $\sigma \in S_n$, $b \in \mathbb{R}_+^n$, $\mathbf{x}^n(\sigma b) = \sigma \mathbf{x}^n(b)$ and $\mathbf{p}^n(\sigma b) = \sigma \mathbf{p}^n(b)$.
- *Consistency*: Let $i_{n,m}$ be any inclusion map that comes from taking the identity map on the first n components and zero-filling the remaining $m - n$ components and permutating the m components by a permutation of S_m. Let $p_{n,m} : \mathbb{R}_+^m \to \mathbb{R}_+^n$ be the projection such that $p_{n,m} \circ i_{n,m} = id_{\mathbb{R}_+^n}$. Then, the following diagram commutes[1]:

$$\begin{array}{ccc} \Theta^n & \xrightarrow[\mathbf{x}^n]{} & \mathcal{X} \\ {\scriptstyle i_{n,m}}\downarrow & \circlearrowleft & \downarrow{\scriptstyle id} \\ \Theta^m & \xrightarrow[\mathbf{x}^m]{} & \mathcal{X} \end{array} \qquad \begin{array}{ccc} \mathbb{R}_+^n & \xrightarrow[\mathbf{p}^n]{} & \mathbb{R}_+^n \\ {\scriptstyle i_{n,m}}\downarrow & \circlearrowleft & \downarrow{\scriptstyle i_{n,m}} \\ \mathbb{R}_+^m & \xrightarrow[\mathbf{p}^m]{} & \mathbb{R}_+^m \end{array}$$

Informally, a deterministic anonymous mechanism consists of mechanisms such that a) there is an explicit type that models agents not participating in the mechanism, b) are symmetric, that is defined as one wherein the mechanism's rules, functions, and outcomes are invariant under any permutation of the participants' identities. This implies that the allocation and payment rules are formulated such that they do not inherently privilege or discriminate against any individual based on their identity, and c) the mechanism is consistent, fundamentally, that is the mechanism is well-defined for finite arbitrary number of players.

However, in general, deterministic anonymous mechanisms are not enough to describe a lot of important mechanisms. For example, the first price auction is not a deterministic anonymous mechanism. For example, when two agent report the same bid, to whom those the mechanism allocate the item? If the tie-breaking rule gives the item to the first reported agent, the mechanism is clearly non-symmetric and so is not anonymous. To accept randomized tie-breaking rules, we must extend the definition to randomized anonymous mechanisms. A *randomized anonymous mechanism* consists of a probability space $(\Omega, \mathcal{F}, \Pr)$ and a parametrized random variables $\{(\mathbf{x}^n : \Theta^n \times \Omega \to \mathcal{X}, \mathbf{p}^n : \Theta^n \times \Omega \to \mathbb{R}_+^n)\}_{n\in\mathbb{N}}$ such that, the space Θ have inactive

[1] **Category theory observation**: If we take A_n to be the set of symmetric mechanisms with n agents, and $f_n : A_n \to A_{n-1}$ to be $(\mathbf{x}^n, \mathbf{p}^n) \mapsto (\mathbf{x}^n \circ i_{n-1,n}, \mathbf{p}^n \circ i_{n-1,n})$, then anonymous Sybil mechanisms are elements of the inverse limit

$$\varprojlim A_i = \{(a_i)_{i\in\mathbb{N}} \mid a_i \in A_i \text{ for all } i \in \mathbb{N} \text{ and } f_{ij}(a_j) = a_i \text{ for all } i \leq j\}.$$

type and for every event $w \in \Omega$, the induced maps hold consistency. Moreover, the mechanism is *weak anonymous*, that is that for every $i \in \mathbb{N}$ and every permutation $\sigma \in S_\infty$ holds

$$\mathbb{E}_w[x(w,b)] = \mathbb{E}_w[\sigma^{-1}x(w,\sigma(b)]$$
$$\mathbb{E}_w[p_i(w,b)] = \mathbb{E}_w[p_{\sigma(i)}(w,\sigma b)]$$

As a natural example of *randomized anonymous mechanism* is the first-price auction with randomized tie-breaking rule.

Now, a reader may ask why all these properties are relevant? We claim that the properties are not just technical but also capture some important ideas. First, since the mechanism does not have any prior knowledge on the number of agents, the mechanism must accept an arbitrary number of agents being reported, but also an arbitrary number of non-reported agents. Also, if the mechanism is non-symmetric, it implies the ordering of the identities matters, and so, some players are privileged in some instances. This could imply that before participating to the mechanisms agents will play another game to take the identities that, in expectancy benefit the most to them. For example, suppose that the first identities have privileged benefits. How are these first agents choose? By the alphabetic order? By order of arrival? In either case, this would imply that agents will have incentives to have these privileged identities to increase their payoff, playing another game with unknown externalities to the mechanism designer. This motivates the idea of making the game symmetric since then the agents have no prior incentive to strategies their reported identity. Moreover, in a prior-free model, the following holds.

Proposition 3.1. *Suppose all agents can take types in a set common set Θ. Given a direct revelation mechanism $\mathcal{M}$, there is a symmetric-randomized mechanism $\mathcal{M}'$ such that better worst-case welfare, i.e.* $\min_{\boldsymbol{b}} \mathcal{W}_{\mathcal{M}'}(\boldsymbol{b}) \geq \min_{\boldsymbol{b}} \mathcal{W}_{\mathcal{M}}(\boldsymbol{b})$.

Proof. Let $\mathbf{b}$ be a type vector profile such that $\text{len}(\mathbf{b}) = n$, then consider the mechanism that picks a random permutation $\sigma \in S_n$ with probability $1/n!$ and then outputs the allocation and payments of $\mathcal{M}$ applied to σb. Clearly, this mechanism is a symmetric mechanism. □

In particular, in the prior-free model, if we are considering to maximize the worst-case welfare, we can restrict to symmetric mechanisms. A similar result holds when considering the consumer surplus.

3.2 Sybil extension of anonymous mechanism

Given an anonymous mechanism $\mathcal{M}$, an arbitrary finite number of types can be reported to the mechanism. The mechanism captures the idea of mechanisms that a) a prior treats all agents equally and b) is well-defined by an arbitrary number of agents. However, in our setting, the mechanism designer does not neceserly have the ability to distinguish between different parties. And so, users can exploit this asymetry of information from the mechanism designer to try to increase its payoff by reporting more than one type via creating fake identities with multiple types to the mechanism. And so, we can think of this as the extension mechanism of the underlying mechanism, where each agent can report a finite arbitrary number of agents to the mechanism. However, creating a fake identity incurs to some costs to the agent employing the strategy. This can be thought as the cost of buying bots, or the cost of generating different identities. Different examples have been seen in different contexts, such creating new EOAs in Ethereum to wash-trade [25], the cost of generating new mails and bank accounts to participate in eBay auctions [2], or to rent the use of third-party IDs to create new League of legend account and resell it in a secondary market once its ELO-boosted [3].

[2] `https://www.ebay.com/help/policies/selling-policies/selling-practices-policy/shill-bidding-policy`

[3] `https://www.playerauctions.com/lol-account/challenger/`

We call the mechanism that takes into account the costs of agents to create Sybils and the ability of agents to report multiple identities to the mechanism the Sybil extension mechanism and is formally defined as follows. A *Sybil extension mechanism* of an anonymous mechanism $\mathcal{M}$ is a tuple $(\mathcal{M}, C)$ where:

- $\mathcal{M}$ is an anonymous mechanism.
- $C : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is the cost function that models the cost that an agent must bear when it generate x identities once other players have generated y identities.

To the Sybil extension mechanism, an agent can report any type vector profile $b_k \in \bigoplus_{i\in\mathbb{N}} \Theta$ of finite length and if other agents report type vector profile b_{-K} its utility is:

$$u_i(b_K, b_{-K}, \theta) = \mathbb{E}\left[v_i(x(b_K, b_{-K}), \theta) - \sum_{k\in K} p_k(b_k, b_{-k})\right] - C(|K|, \text{len}(b)). \tag{1}$$

In other words, the utility of the agent is the expected valuation of the outcome minus the payments made by the Sybils to the auctioneer minus the cost of generating the Sybils. Observe that the underlying mechanism $\mathcal{M}$ can be thought as a Sybil extension mechanism $\mathcal{M}$ with cost function $C(x, y) = \infty \cdot \mathbb{1}_{x\geq 2}$, since in this case, no agent will have incentives to misreport identities since will have to pay an infinite amount to generate extra sybils. To motivate the study, lets start with an example.

Let's consider the mechanism where some divisible good X with common value R is shared among reported participants proportionally. In case that agents can not use sybils, each agent is allocated an n fraction of the good X and so the utility of each agent is R/n. Now, let's consider the Sybil extension where agents can report arbitrary number of identities, and each creating of identity has a cost $c \geq 0$. Each identity has a trivial set of actions (participate). And so, the Sybil extension game can be modelled as a game with unknown number of players and payoff:

$$u_i(x, y) = \frac{x}{x + y}R - cx, \tag{2}$$

where x is the number of players reported by the ith player and y is the total number of identities reported by the other players. In this scenario, as we show in the following proposition, the welfare loss is of the order of magnitude of n.

Proposition 3.2. *The game defined by 2 has a symmetric mixed Nash equilibrium. The strategy consists of randomizing between two actions. The mixed strategy is defined as*

$$\pi = \begin{cases} \lfloor \frac{R/c(n-1)}{n^2} \rfloor, & \textit{with probability } p, \\ \lceil \frac{R/c(n-1)}{n^2} \rceil, & \textit{with probability } 1-p, \end{cases}$$

for some $p \in [0, 1]$. The welfare in equilibrium is $\Theta(R/n)$ and so the ratio of the optimal social welfare and the welfare in equilibrium is $\Theta(n)$.

Proof. First, assume that players can take actions in $x \in \mathbb{R}_+$. Then, the vector

$$(h_1, ..., h_n) = \frac{n-1}{n^2}R \cdot \mathbf{1}^T.$$

is a Nash equilibrium. With vector of payoffs $(U_1, ..., U_n) = \frac{R}{n^2} \cdot \mathbf{1}^T$.

Assume that other players are consuming y computation resources, since the player i is individually rational, solves

$$\begin{aligned} &\underset{x}{\text{maximize}}\ U_i(x, y) \\ &\text{subject to } x \geq 0 \end{aligned}$$

Clearly U is twice differentiable, and we have that

$$\frac{\partial U(x,y)}{\partial x} = \frac{y}{(x+y)^2}R - 1 \tag{3}$$

$$\frac{\partial^2 U(x,y)}{\partial x^2} = -\frac{y}{(x+y)^3}R < 0. \tag{4}$$

By 4 we have that U is concave, and therefore the local maximum are global. By 3, we have that the maximum is realized in $x^\star(y) = \sqrt{Ry} - y$. Now, assume that $(x_1, ..., x_n)$ is a Nash equilibrium. So, we have that $x_i = \sqrt{Rx_{-i}} - x_{-i}$ for all i. Then $x_T^2 = Rx_{-i}$ for all i and therefore we have that $x_i = x_j$ for all i, j. Implying we have a unique Nash equilibrium. Now, let's compute it. Let $x := x_1$, we have that $n^2x_1^2 = R(n-1)x_1$, so $x_1 = \frac{R(n-1)}{n^2}$. The payoff of this equilibrium is $\frac{R}{n^2}$ and so the social welfare in equilibrium is R/n. Now, there is a p such that if all players play π, then $U_i(\lfloor\frac{R/c(n-1)}{n^2}\rfloor, \pi_{-i}) = U_i(\lfloor\frac{R/c(n-1)}{n^2}\rfloor + 1, \pi_{-i})$. And so, we deduce that with this p, $(\pi, ..., \pi)$ is a Nash equilibrium. □

In other words, there is a loss of welfare of order of $\Theta(n)$ due to the ability of players to submit false identities. And so this leads to the following questions:

- When agents do not have incentives to create Sybil identities?
- For any given allocation problem is there a mechanism that does not incentive creating Sybil identities and does not sacrifice the overall welfare of the participants?

In this paper, we will formalize these questions and solve them in some circumstances.

3.3 Sybil proofness

Sybil-proofness, or false-name proofness, is a crucial property in mechanism design, ensuring resilience against attacks where a single entity generates multiple identities to manipulate the system. A mechanism is Sybil-proof if it maintains its intended outcomes and utility functions even when participants can freely create and control multiple identities. This property is vital in contexts like distributed systems, blockchain networks, and online voting mechanisms, where identity verification is limited or non-existent. Technically, Sybil-proof mechanisms are designed so that no participant can increase their utility by splitting their true identity into several false ones. This involves structuring incentives and rules such that the strategy of behaving as a single entity yields the highest utility, regardless of the number of identities a participant might control.

Definition 3.3. *Formally, an anonymous mechanism $\mathcal{M}$ is Sybil-proof, if the subspace $\Theta \times \prod_{i\geq 2}\{0\} \subseteq \bigoplus_{i\in\mathbb{N}}\Theta$ is a weak-dominant strategy set for every Sybil extension $(\mathcal{M}, C)$. In other words, for every agent i with type θ for every type vector profile b_{-i} reported to the mechanism and b_K Sybil strategy of agent i, there exists a type $b_i' \in \Theta \times \prod_{i\geq 2}\{0\}$ such that $u_i(b_i', b_{-K}, \theta) \geq u_i(b_K, b_{-K}, \theta)$.*

Examples of Sybil-proof mechanisms are first-price auction and second-price auction when just taking into account the strategic behaviour of the buyers. When sellers have valuation over their item and they can bid in their own auction, neither the first-price nor the second-price auction are Sybil-Proof. In case the sellers can report a reserve price to the mechanism, the first-price auction is Sybil-proof also for sellers, while it is not the case in second-price auction. Also, cake-cutting mechanisms, cost-sharing mechanisms, and VCG combinatorial auctions are in general not Sybil-proof (more details in following sections and chapters). In this chapter however, we will be more interested in coalition Sybil-Proof mechanisms such as bidding rings.

4 Reward sharing Sybil-Proof mechanisms

The distribution of resources among multiple players is a vital concern in both economics and game theory. One key and straightforward category of challenges in this area involves creating mechanisms for distributing rewards when agents are able to complete an specific task. These mechanisms seek to allocate a divisible asset with a common value of R to players who have a linear payoff function and can complete a specific task. For example, suppose that agents are asked to provide a solution to a Diophantine equation or a formal proof that does not have solution. The creator of the task does not want to give all the rewards to the first provided but rather wants to reward all participants that were able to provide the solution fairly. However, the task is knowledge/information-based. Once you know the solution, you can report it an arbitrarily finite number of times with a small extra marginal cost, and for the person that does know the solution of the task does not have extra costs to produce it. In this section we will analyse the limits of this reward mechanisms.

Lets assume that the main goal is to maximize the total payoff constraint to the distribution being Sybil-Proof.

In this section, we propose a dominant strategy incentive compatible (DSIC) and Sybil-proof mechanism for allocating a part of the fungible item. We also prove that this mechanism is the optimal solution for this class of mechanisms. The mechanism will consist of sufficiently shrinking the "pie" per self-reported identity. We call this technique the *pie shrinking with crowding*. Furthermore, we investigate the case where players have some information about the number of players, and explore how this impacts the design and performance of the mechanism.

More formally, a reward distribution mechanism consists of a map $r : \{0,1\}^\star \to \mathbb{R}$ such that:

- $r_i(\mathbf{b})$ is the total amount of reward obtained by the identity i.
- For any arbitrary number of reported agents, the total reward distributed is less or equal than R, i.e. $\sum_{i=1}^{\infty} r_i(\mathbf{b}) \leq R$ for all $\mathbf{b} \in \{0,1\}^\star$.

As a constraint, we will ask that the reward distribution mechanism is symmetric, and so, it will treated all reported agents equally. This implies that the reward distribution mechanism can be written has $r_i(b) = r(\#\{j : b_j = 1\} \cdot \mathbb{1}_{b_i=1})/\#\{j : b_j = 1\}$. Now let $r : \mathbb{N} \to \mathbb{R}$ be that holds previous property. Assume that n players are reported to the mechanism, then $r(n)$ is split among the reported players. So, the total payoff distributed is $r(n)$ and each player obtains $r(n)/n$. Clearly, if $r(n) = 0$ for all $n \geq 3$ this mechanism is Sybil-Proof, however, the total users utility is zero. Another example of mechanism is $r(n) = R$, however, as mentioned before, this mechanism is not Sybil-proof in general. In the following, we will compute the optimal symmetric Sybil-proof mechanism. That is, given the family of mechanisms

$$\mathcal{M} = \{r : \mathbb{N} \to [0, R] \mid r \text{ induces a Sybil-Proof mechanism } \forall c > 0, n\},$$

the optimal social welfare mechanism is $r_{max}(n) = \text{argmax}_{r\in\mathcal{M}}\{r(n)\}$.

Observe that if l Sybils are reported to the mechanism, the payoff obtained by a Sybil is $r(l)/l$ and so, the payoff of the player is $U_i(x, y) = x/(x + y)r(x + y) - C(x, y)$. Now take a player i and assume that there are y of reported players in the game. Then, the best response is $x \in \mathbb{Z}_{\geq 0}$ such that

$$\begin{aligned} &\underset{x}{\text{maximize}}\ x\frac{r(y+x)}{y+x} - C(x,y) \\ &\text{subject to } x \in \mathbb{Z}_{\geq 0} \end{aligned}$$

Therefore, a mechanism $r \in \mathcal{M}$ is Sybil-proof for all cost-functions C, if and only if, $\frac{r(1+y)}{1+y} \geq x\frac{r(y+x)}{y+x}$ for all $x \geq 1, y \geq 0$. In particular, r holds $r(1 + y)/(1 + y) \geq 2r(2 + y)/(2 + y)$ for all $y \geq 0$. With this inequality, it can be deduced the following proposition.

Proposition 4.1. *The distribution mechanism given by* $r(n) = \frac{n}{2^{n-1}}R$ *is a Sybil-proof mechanism for all* $c \geq 0$*. Moreover,* $r = r_{max}$ *i.e. the mechanism proposed, is the social welfare optimal in* $\mathcal{M}$*.*

Proof. First, if a mechanism is Sybil-proof for all increasing cost functions C then in particular Sybil-proof for $c = 0$. As mentioned previously, we have that for all $y \geq 1$, $r(1+y)/(1+y) \geq 2r(2+y)/(2+y)$. And so, we have that

$$r(1+n) \leq \frac{r(n)}{2}\frac{n+1}{n}, \text{ for } n \geq 1.$$

And so, recursively we deduce that

$$r(1+n) \leq \frac{r(1)}{2^n}\prod_{k=1}^{n}\frac{k+1}{k} = \frac{r(1)}{2^n}(n+1)$$

Since $r(1) \leq R$, we deduce that $r(n) \leq \frac{n}{2^{n-1}}R = r_{max}$. □

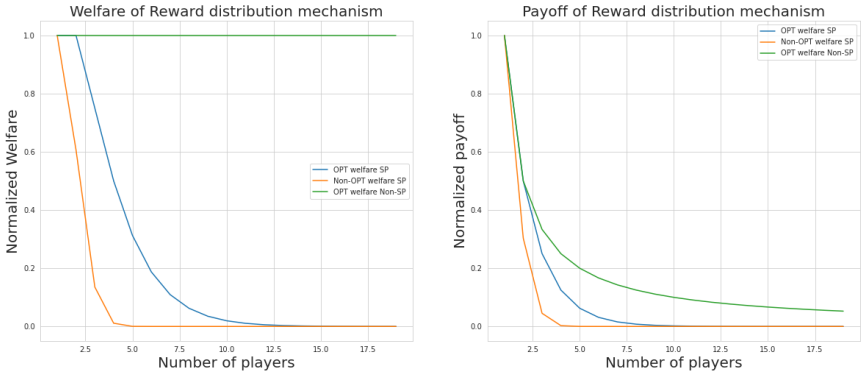

Figure 1: Welfare and payoff of reward distribution mechanisms

Observe that with the reward mechanism r_{max}, is by construction, strictly dominating strategy to self-report just once. Also, an implication of the proposition is that there does not exist an efficient prior-free reward mechanism. Informally, these mechanisms consists of shrinking the pie optimally to have an optimal Sybil-proof reward mechanism. This technique can also be used if the item is non-divisible.

If the item is non-divisible and players are risk-neutral and share the same valuation on the item, then a similar Sybil-proof mechanism can be achieved by random allocating the item to any player with probability $\Pr[\text{item allocated to the } i\text{th player}] = \frac{1}{2^{n-1}}$. Clearly, there is (large) inefficiency lost since with probability $1 - \frac{n}{2^{n-1}}$ the item is not allocated to any player.

This simple study has the following interpretation: when a set of agents observe a common value R and want to create a coalition to share the reward symmetrically but they do not know the number of counterparties, then the amount of the value shared among the participants decreases exponentially when we impose that the mechanism is Sybil-proof.

However, when there is some common prior information over the number of agents, the mechanism can exploit that to obtain a better welfare.

Let's assume that the distribution of number of players $\mathcal{D}$ is common knowledge. And that agents completely complement each other, i.e. they generate a symmetric game. The mechanism designer objective is to find a symmetric mechanism that maximizes the expected total payoff.

Let $\mathcal{D}'$ be the distribution of $\mathcal{D}$ conditioned to $\mathcal{D} \geq 1$. We write $p_n = \Pr[\mathcal{D}' = n]$. Our objective is to find a Sybil proof symmetric mechanism that maximizes the expected sum of

payments, i.e. the total payoff. We want to split part of the total value $R \in \mathbb{R}_{\geq 0}$ generated by the complementarity among the agent. We can model the distribution with a vector $X \in \mathbb{R}^{\infty}$ such that if there are a total number of n players reported, each player receives X_n. Therefore, if there are n identities (already) reported and a player reports k identities more, he obtains $k \cdot X_{k+n}$. Our objective is to find X such that:

1. (Ex-ante Sybil-proof) All players reporting one identity is an equilibrium (Weaker than Syibil-proof), formally

$$\mathbb{E}_{n\sim\mathcal{D}'}[X_n] \geq \mathbb{E}_{n\sim\mathcal{D}'}[y \cdot X_{y-1+n}] \text{ for all } y \geq 1.$$

2. (Budget Balance) For every reported number of identities n, it holds $n \cdot X_n \leq R$.

3. (Efficiency) X maximizes the expected total payoff of all players in equilibrium. That is

$$X^* = \text{argmax}_X \quad \mathbb{E}_{n\sim\mathcal{D}'}[n \cdot X_n]$$

And so, we can rewrite this as the Linear optimization problem

$$\begin{aligned} &\underset{X}{\text{maximize}} \sum_{n=1}^{+\infty} n \cdot X_n \cdot p_n \\ &\text{subject to } 0 \leq X_n \leq R/n \text{ for all } n \geq 1 \\ &\qquad \sum_{n=1}^{\infty} (X_n - X_{y-1+n} \cdot y) p_n \geq 0 \text{ for all } y \geq 2 \end{aligned}$$

As we will show in the figure 2, the optimal prior is in general strictly higher than the prior-free optimal rebate for full complementary transactions. An easy example to show this is when $p_n = 1$ for some n, since $X_n = R$, $X_k = 0$ for $k \neq 0$ is a solution of the optimization problem and has total payoff R. [4]

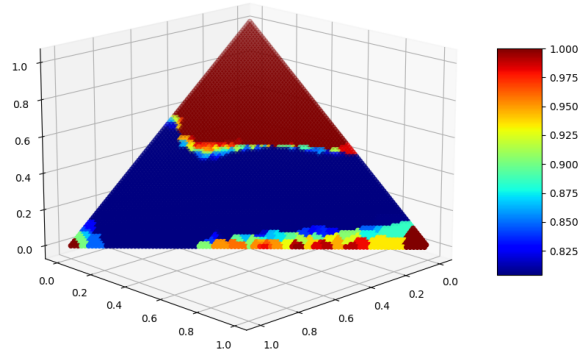

Figure 2: Plot of the solution of the optimization problem for the elements in $\{(0, 0, x_3, x_4, x_5) \in \mathbb{R}^5_+ | x_3 + x_4 + x_5 = 1\}$.

Sybil-proof prior-free truthful cake cutting mechanism.

A natural generalization of the reward distribution problem is the cake-cutting problem, where a divisible good is to be divided among multiple players with heterogeneous valuations. This problem has been well-studied in the literature [4, 27]. Previously, we presented a specific

[4] Code of the following figure is available in `https://github.com/BrunoMazorra/COFA.git`.

cake-cutting mechanism for the case of divisible goods with homogeneous valuations. However, many real-world scenarios involve a heterogeneous good, such as a cake with different flavors or a public good with different features. In this section, we will provide a more general framework for the cake-cutting problem that can handle such heterogeneous goods. We will build upon the work of [27] to present a modified mechanism that incorporates Sybil resistance, providing a solution for this important problem. We will provide a detailed description of the mechanism and its underlying principles, as well as an analysis of its performance. To do so, we will define cake-cutting mechanisms for finite but unbounded number of players. We will reformulate different notions of fairness, such as envy-freeness and proportionality. We will prove that no Sybil-proof mechanism is efficient, and we will upper bound the worst-case social welfare. To do so, we will again use pie shrinking with crowding.

Cake-cutting model: Let $\mathfrak{C}$ be a $\sigma-$algebra of the set $\mathcal{C}$ ($\mathfrak{C}$ model the slices of the cake, and $\mathcal{C}$ the cake). Let there be n players and let $\mu_1, ..., \mu_k$ be non-atomic probability measures on $(\mathcal{C}, \mathfrak{C})$, so the value of the slice $C \in \mathfrak{C}$ is $\mu_i(C)$. Let $\mathcal{L} = \{\mu : \text{non-atomic probability measure of } (\mathcal{C}, \mathfrak{C})\} \cup \{0\}$, where 0 is the map that sends everything to zero. A cake-cutting mechanism with finite but unknown bounded number of players consists of a (random) allocation rule[5] $\mathfrak{A} : \mathcal{L}^\infty \times \Omega \to \mathfrak{C}^\infty$ an where if players $i = 1, ..., k$ report to the mechanism the valuations $\mu_1, ..., \mu_k$ and an event $w \in \Omega$ is drawn, then

$$\mathfrak{A}((\mu_1, ..., \mu_k, 0, ...), w)_j = \begin{cases} \text{slice allocated to player } i, \text{ if } 1 \leq j \leq k \\ \emptyset, \text{ otherwise} \end{cases}$$

and for all $(v, w) \in \mathcal{L}^\infty \times \Omega$ it holds $\mathfrak{A}(v, w)_i \cap \mathfrak{A}(v, w)_j = \emptyset$ for all i, j. Moreover, we will assume that the allocation is symmetric. That is, for all $(v, w) \in \mathcal{L}^\infty \times \Omega$ and all permutation $\sigma \in S^\infty$ it holds $\mathfrak{A}(v, w)_i = \mathfrak{A}(\sigma(v), w)_{\sigma(i)}$. In cake-cutting literature [30], different fairness criteria are taken into account for designing cake-cutting mechanisms. In the following, we reformulate the most important once in cake-cutting mechanisms with a finite (but not bounded) number of players.

A mechanism $\mathfrak{A}$ is:

- **Envy-free in expectation** if, for every allocation, no player prefers another player's slice of cake. More formally, if for all vector of valuations $\mu \in \mathcal{L}^\infty$ and all i, j it holds $\mathbb{E}[\mu_i(\mathfrak{A}(\mu)_i)] \geq \mathbb{E}[\mu_i(\mathfrak{A}(\mu)_j)]$.
- **α-proportional** if for all vector of valuations $\mu \in \mathcal{L}^\infty$ and all i such that $\mu_i \neq 0$ it holds $\mu_i(\mathfrak{A}(\mu, w)_i) \geq \alpha$. We say that is in expectancy α-proportional if the same holds in expectancy.
- **Non-wastefulness** if for all vector of valuations $\mu \in \mathcal{L}^\infty$ and event $w \in \Omega$ we have that $\bigcup_{i=1}^\infty \mathfrak{A}(\mu, w)_i = C$.
- **Truthful in expectation** if truth-telling is a weak-dominant strategy. More formally, it holds $E[\mu(\mathfrak{A}(\mu, v_{-i})_i)] \geq \mathbb{E}[\mu(\mathfrak{A}(\nu, v_{-i})_i)]$ for all $\mu, \nu \in \mathcal{L}, v \in \mathcal{L}^\infty$ and $i \in \mathbb{Z}_{\geq 0}$.

When extending the cake-cutting mechanism to include the possibility of Sybil/false-name valuations, where each player can report multiple (but finite) identities to the mechanism, traditional notions of fairness such as proportionality and envy-freeness no longer apply. For example, in a cake-cutting game where one player uses a Sybil attack to submit multiple valuations, a proportional cake-cutting mechanism would result in an asymmetric distribution as the attacker would receive a larger portion of the cake. Similarly, envy-freeness, which ensures that no player prefers another player's portion of the cake, cannot be guaranteed as players do not know which

[5]here we denote by $\mathcal{L}^\infty = \{v \in \prod_{i=1}^\infty \mathcal{L} : \text{ with finite number of non-trivial components}\}$, and similar for $\mathcal{C}^\infty$.

identities reported to the mechanism belong to each player. A natural way of keeping these properties in the Sybil extension game is by making the mechanisms Sybil-proof. If players have no incentives to report more than one valuation, then all mathematical definitions stated before maintain their intrinsic characteristics. Using the definition of Sybil-proofness, we have that a cake-cutting mechanism $\mathfrak{A}$ is Sybil-proof if holds:

$$\mathbb{E}[\mu_i(\mathfrak{A}(\mu_i, \mu_{-i}, \cdot)_i)] \geq \mathbb{E}[\sum_{j=1}^{k} \mu_i(\mathfrak{A}(\nu_{i_1}, ..., \nu_{i_k}, \mu_{-i}, \cdot)_{i_j})] \text{ for all } \mu_i, \nu_{i_1}, ..., \nu_{i_k} \in \mathcal{L}, \mu_{-i} \in \mathcal{L}^{\infty}. \quad (5)$$

In the following proposition, we will see that Sybil-proofness is a strong condition. This condition implies that the mechanism waste (a large) part of the cake if other properties such as proportionality want to be guaranteed. Moreover, we upper bound the worst-case social welfare of Sybi-proof cake-cutting mechanisms.

Proposition 4.2. *No truthful Sybil-proof cake cutting is α-proportional for $\alpha > 1/2^{n-1}$. In particular, truthful Sybil-proof cake cutting are not non-wasteful. So the worst-case social welfare is upper bounded by $n/2^{n-1}$, i.e. $\min_{\mu \in \mathcal{L}^n} \mathbb{E}[\sum_{i=1}^{n} \mu_i(\mathfrak{A}(\mu, \cdot)_i)] \leq n/2^{n-1}$.*

Proof. Assume all players have the same valuation μ. Also, we assume that there is a Sybil-resistant cake-cutting algorithm that is α-proportional. Then, if players report μ each player obtains some slice with exactly the same valuation $\beta(n) \geq \alpha$ (by symmetry). Since is Sybil-proof, it holds $\beta(n - 1 + k)k \leq \beta(n)$ for all $k \geq 1$. In particular, $2\beta(n + 1) \leq \beta(n)$. And so, we deduce that $\beta(n) \leq 1/2^{n-1}$, then $\alpha \leq 1/2^{n-1}$. The worst case welfare is in particular smaller than $n\beta(n) \leq n/2^{n-1}$. □

In the previous proposition, we have seen that all mechanisms hold

$$\min_{\mu \in \mathcal{L}^{\infty}} \mathbb{E}[\sum_{i=1}^{\infty} \mu_i(\mathfrak{A}(\mu, \cdot)_i)] \leq n/2^{n-1}. \quad (6)$$

In the following, we will define a (non-constructive) mechanism that is truthful, Sybil-proof, and is $1/2^{n-1}-$proportional. In order to do so, we will use Neyman's theorem (similar to [27]) that establishes that there exists a partition $C_1, ..., C_k$ of the cake $\mathfrak{C}$ such that for all players i and slices j it holds that $\mu_i(C_j) = 1/k$. The mechanisms works as follows:

Assume that players' true valuations are $\mu_1, ..., \mu_n$ and that each one declares some measure ν_i. First, find a partition $C_1, ..., C_n$ such that for all i, j holds $\nu_i(C_j) = 1/n$. Then choose a random permutation $\sigma \in S_n$ from the uniform distribution. Afterward, toss a biased coin X with probability $\Pr[X = 1] = n/2^{n-1}$. If $X = 1$, then allocate $C_{\sigma(i)}$ to the ith player, otherwise allocate the empty set.

Proposition 4.3. *The mechanism Sybil-proof (non-constructive) cake-cutting mechanism that is in expectancy $1/2^{n-1}-$proportional.*

Proof. Observe that the expected size of the slice of player i is

$$\Pr[X = 1] \left(\sum_j \mu_i(C_j) \Pr[\sigma(i) = j] \right) = \frac{n}{2^{n-1}} \sum_j \mu_i(C_j)/k = \frac{n}{2^{n-1}} \mu_i(\cup_j C_j) = \frac{1}{2^{n-1}}.$$

Since this quantity is independent of ν_i then player i has no incentive to declare $\nu_i \neq \mu_i$. And so, the mechanism is truthful and expectancy $\frac{1}{2^{n-1}}$-proportional. Now, if the player declares k

identities and other players report y. We write $n = y + k$, then the expected payoff is:

$$\begin{aligned}\sum_{\mathcal{I}\subseteq[n]:|\mathcal{I}|=k}\left(\sum_{j\in\mathcal{I}}\mu(C_j)\Pr[X=1]\right)\Pr[\sigma([k])=\mathcal{I}] &= \Pr[X=1]\Pr[\sigma([k])=[k])\sum_{j\in[n]}\sum_{\mathcal{I}\subseteq[n]:|\mathcal{I}|=k,j\in\mathcal{I}}\mu_i(C_j)\\ &= \frac{n}{2^{n-1}}\frac{y!k!}{n!}\sum_{j\in[n]}\mu_i(C_j)\binom{n-1}{k-1}\\ &= \frac{k}{2^{y+k-1}}\end{aligned}$$

Since the function $f(x) = \frac{x}{2^{y+x-1}}$ restricted to $\mathbb{Z}_{\geq 0}$ is maximized in $x = 1$, we have that the mechanism is Sybil-proof. □

Putting it all together, we have proved that

$$\max_{\mathfrak{A}\in\mathcal{M}}\min_{\mu\in\mathcal{L}^n}\mathbb{E}[\sum_{i=1}^{n}\mu_i(\mathfrak{A}(\mu,\cdot)_i)] = n/2^{n-1}$$

where $\mathcal{M}$ is the set of all truthful Sybil-proof cake-cutting mechanisms.

4.1 Bidding ring in second price auction with an unknown number of players

Another generalization of reward sharing arises in mechanisms such that players want to form a coalition in order to reduce competition and increase their expected profits. For example, in auctions, the bidders can act collusively and engage in a collusion with a view to obtaining lower prices, see [19]. The resulting arrangement, usually nominated by *bidding ring*, is studied in [23, 24, 26] in the different auctions and different conditions in the properties of the cartel.

In a second-price auction, bidders submit sealed bids and the highest bidder wins the auction, but pays the price of the second-highest bid rather than their own. If there are ties, the item is allocated randomly among the identities that made the highest bid. Clearly, second-price auctions (and standard auctions with one item) are Sybil-proof (in case that the valuation of the players have no mass points). For example, in a first-price auction, if two or highest bids coincide, then players have incentives to bid more than once in order to increase their chances to obtain the item. However, if the distribution has no mass points, no player can benefit from bidding more than once. We will discuss mechanisms where players try to form a coalition, known in context of auctions as bidding rings.

We will consider the model proposed in [26]. There are n risk-neutral bidders, denoted by $i = 1, ..., n$, each player i knows his own valuation v_i of the item. In the prior-free setting no extra assumptions are made over the valuation v_i, while in the Bayesian setting the private valuations v_i are independently drawn from a cumulative distribution F. The seller sells the item through a second-price auction. In this context, the second-price auction is social-welfare optimal, that is the item is allocated to the buyer with highest valuation. However, is not consumer surplus optimal, that is, the sum of utilities of the members of the cartel (in our case the buyers) is not necessarily optimal. More formally, given a truthful mechanism $(\mathbf{x}, \mathbf{p})$ for allocating the item, the consumer surplus is defined as

$$CS(\mathbf{v}) = \sum_{i=1}^{n} x_i(\mathbf{v})v_i - p_i(\mathbf{v}),$$

when agents report truthfully and have valuation profile $\mathbf{v}$. The optimal consumer surplus is CS-OPT $= \sum_{i=1}^{n} x_i^\star(\mathbf{v})v_i$, where $x^\star$ is the VCG mechanism. Now, while the VCG mechanism is

the most efficient one in terms of welfare, is not in terms of consumer surplus. For example, if two agents have valuation $v_1 = v_2 = 10$ over the item, the consumer surplus is 0 while the social welfare is 10. On the other hand, if a mechanism is DSIC and optimal consumer surplus, would also be social welfare optimal and so would be VCG, leading to a contradiction. Therefore, no prior-free DSIC mechanism is consumer surplus optimal.

For this reason, let's first focus on the Symmetric Bayesian case. Lets suppose that the buyers valuations are drawn i.i.d from a cdf F has differentiable density f with support $[0, v_h]$ and F is common knowledge. In this case, by [26] there is efficient incentive-compatible (truthful bidding) and efficient (bidding ring) mechanism. The mechanism works as follows. Before the auction, the cartel members report their valuations to the mechanism. If no report exceeds r, the cartel does not bid in the auction. If at least one player i exceeds the bid r, the bidder making the highest report v obtains the item and pays a total of

$$T(v) = (n-1)F(v)^{-n} \int_r^v (x-r)F(x)^{n-1} f(x)dx + r. \tag{7}$$

Each loser bidder receives from the winner $(T(v)-r)/(n-1)$ and the seller receives r. Moreover, they prove that every incentive-compatible and efficient mechanism has the property that the winner transfers to each loser an amount equal to $V(n) = \mathbb{E}[v_{(2)} - r \mid v_{(1)}]/n$ (where v_j is the jth order statistic). Observe that this mechanism is ex-ante budget-balance and consumer surplus optimal in expectancy. Two questions arise from this mechanism

- What if we impose that the mechanism to be Sybil-Proof?
- Is there a mechanism that is DSIC, Sybil-Proof and close to consumer surplus optimal?

First, observe that previous mechanism is not Sybil-Proof. The proof of this is straightforward. For $nV(n) \uparrow v_h$ and so, for n sufficiently large, holds $2V(n+1) \geq V(n)$. And so, with a sufficient number of players, any bidder i with valuation $v_i \geq r$ has incentives to bid at least twice (one truthfully and another one r). In particular, any consumer surplus optimal bidding ring (the bidder with the highest valuation wins the auction and all the surplus generated in the collusion is divided among the bidding ring members) is not Sybil-proof. To find permissionless bidding rings that are Sybil-proof and bayesian incentive compatible, first, we define the following family of bidding rings defined by (T, g):

- Registration phase: All players register an identity responsible for bidding $i = 1, ...k$ (potentially some identities will be Sybils).
- All players submit their bids $w_1, ..., w_k$ to the ring center (potentially some players will submit more than one Sybil bid).
- The bidder with the highest bid pays $T(v; k)$ to the ring center and submits the highest value to the seller.
- W.l.o.g assumes that the winning bid is $i = 1$. Then, the ring center pays to each loss identity $i = 2, ..., k$ the amount $g(k)T(v; k)$ and credible burns the remaning part.

Since we want the mechanism to be budget balance, from now on we will assume that $g \leq 1/(k-1)$. Observe that if $g(k) = 1/(k-1)$, and T defines as 7, then we have the previous mechanism. We denote by $\mathcal{B}$ the set of bidding rings of this family such is incentive compatible to report the bid truthfully. In the following proposition, we give in terms of g, the function T in order to be incentive compatible to report the true valuation. For simplicity, from now on, we will assume that the reserve price r equals zero.

Proposition 4.4. *For a given g, the bidding ring (T, g) with:*

$$T(v) = F(v)^{-n-l(n)+1} \int_r^v (n-1)vF(u)^{n-2+l(n)} f(u)du \tag{8}$$

with $l(n) = (n-1)g(n)$ is incentive compatible.

Proof. First, we define $T(v)$ as the amount being payed by the player with highest valuation. And we define $g : \mathbb{N} \to [0, 1]$ to be the fraction of $T(v)$ being paid to a reported player, that is losing player will obtain will get paid $g(n)(T(v) - r)$. Since the protocol is budget balance, we have that $g(n) \leq 1/(n-1)$. Now, let $\pi(w, v, m)$ be the expected payoff of a player with valuation v, report w and reporting m identities to the protocol. From now on we will denote by n the number of actual players (not the reported ones) that are members of the cartel. In this scenario, it holds:

$$\begin{aligned}
\pi(w, v, 1) &= (v - T(w))F(w)^{n-1} + \\
&\quad (1 - F(w)^{n-1}) \int_w^{v_h} g(m+n-1)(T(u) - r) \frac{(n-1)F(u)^{n-2}f(u)}{1 - F(w)^{n-1}} du \\
&= (v - T(w))F(w)^{n-1} + \int_w^{v_h} g(n+m-1)(n-1)(T(u) - r)F(u)^{n-2}f(u)du
\end{aligned}$$

And so, if we denote by $l(n) = (n-1)g(n)$, we have

$$\begin{aligned}
\frac{\partial \pi}{\partial w} &= (n-1)(v - T(w))F(w)^{n-2}f(w) - T'(w)F(w)^{n-1} - l(n)F(w)^{n-2}f(w)(T(w) - r) \\
&= \left((n-1)v - (n-1+l(n))T(w) + l(n)r\right) F(w)^{n-2}f(w) - T'(w)F(w)^{n-1}
\end{aligned}$$

Since $\partial^2\pi/\partial w \partial v \geq 0$, incentive compatibility is characterized by

$$\frac{\partial \pi}{\partial w} \mid_{w=v} = 0$$

This induces the following differential equation:

$$((n-1)v)F(v)^{n-2}f(v) = ((n-1) + l(n))T(v)f(v)F(v)^{n-2} + T'(v)F(v)^{n-1} \tag{9}$$

Multiplying by $F(v)^{l(n)}$ we obtain

$$((n-1)v)F(v)^{n-2+l(n)}f(v) = (n-1+l(n))T(v)f(v)F(v)^{n-2+l(n)} + T'(v)F(v)^{n+l(n)-1} \tag{10}$$

Integrating in both sides, we obtain

$$\int_r^v ((n-1)v)F(u)^{n-2+l(n)}f(u)du = T(v)F(v)^{n+l(n)-1} \tag{11}$$

$$F(v)^{-n-l(n)+1} \int_r^v ((n-1)v)F(u)^{n-2+l(n)}f(u)du = T(v) \tag{12}$$

□

Observe that if $g = 0$, then $T(x) = \mathbb{E}[v_{(2)} \mid v_{(1)} = x]$ and so the expected profit of the Bidding ring (T, g) is exactly the expected profit without the bidding ring. Also, as we have seen, for not all g, the bidding ring (T, g) is Sybil-proof. Let $\mathcal{BS} \subseteq \mathcal{B}$ be the set of incentive-compatible bidding rings that are Sybil-proof in equilibrium (all reporting exactly one identity is an equilibrium). Observe that $\mathcal{BS}$ is non-empty since for $g = 0$, the corresponding bidding

ring is Sybil-proof. Therefore, we can consider the optimal Sybil-proof bidding ring g_{max}, and the optimal consumer surplus OPT:

$$g_{max} = \underset{g;(T,g)\in\mathcal{BS}}{\operatorname{argmax}} \mathbb{E}\left[\sum_{i=1}^{n} u_i(w_i, n_i, w_{-i}, n_{-i}) \mid w_i = v_i, n_i = 1, \forall i\right]$$

$$\text{SP-OPT} = \max_{g;(T,g)\in\mathcal{BS}} \mathbb{E}\left[\sum_{i=1}^{n} u_i(w_i, n_i, w_{-i}, n_{-i}) \mid w_i = v_i, n_i = 1, \forall i\right]$$

with u_i being the utility function of a player. With an easy manipulation, we obtain the following result.

Proposition 4.5. *If the valuations are non-trivial i.e.* $\mathbb{E}[v_{(1)}] \neq 0$*, then SP-OPT* $> \mathbb{E}[v_{(1)} - v_{(2)}]$*. In other words, there are profitable Sybil-proof bidding rings in second-price auctions.*

Now, let's move to the prior-free case. As mentioned previously there is no efficient consumer surplus IR-DSIC and budget-balance mechanism. This leads to the following questions:

- What is the second-best IR-DSIC and budget-balance mechanism, that is, the one that maximizes the worst-case consumer surplus? What if we impose that the mechanism is Sybil-proof?

The first question is partially addressed in [12]. Here, we propose a simpler mechanism to construct an approximate consumer surplus optimal IR-DSIC mechanism. The mechanism involves the following steps:

1. Elicit bids from all agents.
2. Select an agent $i \in [n]$ uniformly at random. This agent will act as the seller.
3. Conduct a second-price auction without a reserve price among the remaining agents $[n]\backslash\{i\}$, with the highest bidder's payment transferred to i.

This mechanism is IR-DSIC as it is a randomization over IR-DSIC mechanisms. Additionally, it is $(1 - 1/n)$-approximate in terms of consumer surplus, meaning $\min_{\mathbf{v}} \text{CS}(\mathbf{v}) \geq (1-1/n)\text{CS-OPT}$. The proof is straightforward. First, the mechanism is budget-balance among the buyers since all payments are transferred to the virtual seller. Second, the item is allocated to the agent who values it most with a probability of $1-1/n$, thereby ensuring the result. While this mechanism is not necessarily consumer surplus optimal, it is asymptotically close.

However, this mechanism is not Sybil-proof. If the cost of generating Sybils is zero, an agent is incentivized to create as many Sybils as possible to increase their chances of becoming the virtual seller and thus enhance their expected payoff. Using similar reasoning as the reward distribution mechanism, the best worst-case Sybil-proof, no-deficit, and IR-DSIC mechanism can achieve at most $\frac{n}{2^{n-1}}$ consumer surplus approximation, considering the valuation vector profile $\mathbf{v} = \underbrace{(R, ..., R)}_{n}$.

In essence, consumer surplus decreases rapidly when cartel members cannot distinguish between real and Sybil agents. This has a beneficial implication for sellers who sell items valued by a large number of agents through the internet or a blockchain, as it makes coordination among agents to maximize consumer surplus significantly more difficult.

5 Sybil Proofness with Commitments

In the previous section, we considered Sybil extension mechanisms as an extension of mechanism where agents can report multiple identities. Now, we will explore Sybil commitments, which offer valuable insights into strategic decision-making when agents credibly delegate decisions to third parties, such as AI or reinforcement learning algorithms, for optimizing preferences in complex environments. This delegation is related to the revelation principle, which states that agents can truthfully reveal their private information to a mechanism while maximizing their utility. Sybil commitment games are an extension of normal games with two phases. In the first phase, each player commits a finite number of Sybils that will act as individual rational agents in the next phase. In the second phase, each Sybil acts under the rationality committed in the previous round and with public knowledge of the number of players in the games. These Sybils are indistinguishable from actual players and are therefore treated as such by other Sybils. These games have the following interpretation: Imagine the case have not completely formed their priors over how many agent they think they are competing against and agent have the option to shift their priors towards thinking there are k more players playing the game, but with the cost that each one of these identities will play as independent (competing) identities in the next phase without the ability to communicate.

The Sybil commitment game has the structure of a two-phase game, we can analyse its equilibrium points, which we will call Sybil commitment equilibrium. These games have natural applications in environments where the players can credibly commit to act in a certain way up to some predicates. For example, in [13], the authors propose a model for games in which the players have shared access to a blockchain that allows them to deploy smart contracts to act on their behalf. Also, in the second phase, Sybils do not need to constrain their actions based on their belief since they have a perfect signal on the number of players. However, we will prove that in general, it is not incentive-compatible to commit a unique agent or, in other words, reveal the agents' identity. In general, when players play a game with incomplete information on the number of agents, they do not have incentives to truthfully reveal their identity. Finally, we will give a necessary condition for players to truthfully report their identity in the commitment phase. The study of these games is crucial for designing mechanisms where agents credibly delegate their preferences and decision-making to third parties or to an AI, ultimately promoting trust, accountability, and efficiency in multi-agent settings. Understanding Sybil commitment games helps develop robust systems across various contexts, from financial markets and online platforms to distributed systems and influence campaigns.

In this context, suppose that all agents have utility functions $u : \mathcal{X} \to \mathbb{R}$ over an outcome space $\mathcal{X}$ and can each identity (or Sybil) can take an action on an action space $\mathcal{A}$, and there is a function that maps all possible (finite but unbounded) vector of actions $a = (a_1, ..., a_n, ...) \in \bigoplus_{i=1}^{\infty} \mathcal{A}$ to an outcome $f(a)$ in $\mathcal{X}$. For every game of this form, we consider the following two-stage game:

1. For every player i the set $\mathcal{N}$ and u_{-i} is unknown.

2. In the first phase, the *Sybil-commitment-stage*, each player i chooses and commits a number of Sybils n_i with Sybil set $\mathcal{I}_i$ with utility functions $\{u_i^1, ..., u_i^j\}$. The set of utility functions lay in $\mathcal{L} \subseteq C(\mathcal{X}, \mathbb{R})$[6]. In other words, the set of preferences that the players can declare lay in $\mathcal{L}$. The cost of generating Sybils is modelled by a function $C : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ where $C(x, y)$ is the cost of a player that generates x identities and the other players generate y identities.

3. In the second phase, all players know the total number $n = \sum_{i=1}^{N} n_i$ of players and the

[6]The set of continuous functions from $\mathcal{X}$ to $\mathbb{R}$.

set of utility functions $\{u_i^j : \mathcal{X} \to \mathbb{R} : i \in \mathcal{N}, j \in \mathcal{I}_i\}$. In this phase, the Sybils play a normal-form game (complete information) with the number of players and utility specified.

4. Finally, if the set of actions taken by all Sybils is $a = (a^1, ..., a^N)$ the total payoff of the player i is

$$u_i(a) - C(|\mathcal{I}_i|, |\bigcup_{k \neq i} \mathcal{I}_k|)$$

We call this game the **Sybil commitment extension game** of the tuple $G = (u_1, ..., u_n, C)$ and denoted by $\mathbf{SyC}(\mathbf{G}, \mathcal{L})$.

Sybil-commitment equilibrium: For a Sybil commitment extension game $\mathbf{SyC}(\mathbf{G}, \mathcal{L})$ we say that a point $(\mathbf{U} = \{U_i^j : \mathcal{X} \to \mathbb{R} : i \in \mathcal{N}, j \in \mathcal{I}_i\}, \mathbf{x}) \in (\mathcal{L}^\infty)^N \times \mathbf{A}^\infty$ is a Sybil Nash equilibrium if:

1. $\mathbf{x}$ is a Nash equilibrium of the normal game $G_n = ([n], \mathbf{A}^n, U)$ with $\mathbf{n}$.
2. Let $\mathbf{U}_i = \{U_i^j : j \in \mathcal{I}_i\} \subseteq \mathbf{U}$. For all $\mathbf{U'}_i \neq \mathbf{U}_i$ we have that $u_i(\mathbf{x}) \geq \max_{\mathbf{y} \in \mathrm{NE}(\mathbf{U'})} u_i(\mathbf{y})$.

We say that a mechanism is *Sybil-commitment-proof (SCP)* if for every number of players, reporting only one identity is a dominant strategy. In other words, providing that information to other players is incentive compatible.

An example of *Sybil-commitment-proof (SCP)* are first-price and second-price auctions.

5.1 Non-SYWC examples

Consider a homogeneous oligopoly with n firms, where the Cournot equilibrium is regular and unique. The inverse demand function is given by $P(x^T \cdot \mathbf{1})$. Each firm (player) chooses an output x_i and its costs are given by $C_i(x_i)$. For an output vector $x = (x_1, ..., x_n)$, the profit to each firm is:

$$u_i(x_i, x_{-i}) = x_i P(x^T \cdot \mathbf{1}) - C_i(x_i)$$

Under sufficient good conditions [8], this game has a unique locally stable Nash equilibrium. Moreover, if the firms have the same cost function C, this equilibrium is symmetric. Now, let's consider the Cournot oligopoly game with $p(x) = \alpha - x$ and $C(x) = cx$ for some $\alpha, c \in \mathbb{R}_{\geq 0}$ such that $c < \alpha$. So in this game, we have that

$$u_i(x_i, x_{-i}) = x_i(\beta - x^T \cdot \mathbf{1})$$

where $\beta = \alpha - c$. In this scenario, the unique equilibrium point is given by $q^\star(n) = \frac{\beta}{n+1}$. The payoff of a player $u_i(q^\star) = \frac{\beta^2}{(n+1)^2}$. And so, the welfare is $\Theta(1/n)$. Now, let's take the Sybil Commitment extension of this game. We know by theorem 5.1 that if $c = \mathcal{O}(1/n^2)$, game is not Sybil-commitment-proof with $\mathcal{L} = \{U_1\}$ (observe all u_i are the same up to reorder of x_i, x_{-i}). More specifically, since the utility of a firm in equilibrium is $u_i(q^\star) = \frac{\beta^2}{(n+1)^2}$, if the total number of players is x_{-i}, the ith player the best response strategy is to generate x sybils such that

$$\begin{aligned} &\underset{x}{\text{maximize}}\ x\frac{\beta^2}{(x + x_{-i} + 1)^2} - cx \\ &\text{subject to } x \in \mathbb{Z}_{\geq 0}. \end{aligned}$$

For example, if there are $n = 10$ players that commit in the Sybil-commitment phase, $\beta = 10$ and $c = 0.001$, then the best-response of a player is $x = 11$.

More general games such as concave pro-rata games [16] with differential function f. We know that these games are Sybil-proof, and by [16] that concave pro-rata games have welfare in equilibrium $\Theta(1/n)$. Therefore, if $c = \mathcal{O}(1/n^2)$, we have that this type of game are not SCP.

Concave pro-rata games are Sybil-proof but are not SCP. In the following, we plot the strategy of committing more than one player to the aggregated arbitrage game in decentralized Batch exchanges. We explore through simulation the payoff of generating one Sybil in the aggregated arbitrage game provided in [16]. The aggregated arbitrage game consists of game played by arbitrageurs that can exploit the price difference between a Constant function market maker C with two assets A and B (for more details, see [2]) and an off-chain market maker essentially risk-free. Moreover, the trades in the constant function market maker are batched before they are executed. Specifically, the trades are aggregated in some way (depending on the type of batching performed) and then traded 'all together' through the CFMM, before being disaggregated and passed back to the users. In this scenario, the authors assume that the forward function g is differentiable (if a player offers Δ assets to the CFMM obtains $g(\Delta)$ of B assets). Also, they assume that the price of the external market maker is c. Therefore, the profit of an arbitrageur trading t assets is $g(t)/c - t$. And so, the optimal arbitrage problem consists of maximizing $g(t) - ct$ constraint to $t \geq 0$.

When the trades of the players are batches, arbitrageurs cannot directly trade with the CFMM, but must instead go through the batching process. Assuming that there are n arbitrageurs, the pro-rata game induced is

$$u_i(x_i, x_{-i}) = \frac{x_i}{x_i + x_{-i}} g(x_i + x_{-i}) - cx_i$$

As shown in [16] this game is a pro-rata concave game and has a unique Nash equilibria. Fixing the CFMM to be a constant product market maker [2] and fixing the reserves, we can compute the payoff of each arbitrageur in equilibrium. In the case that a player can commit more than one identity, we show that this player has incentives to, at least, commit one more identity if the payoff of each player is the payoff in equilibrium with $n+1$ players, and the payoff of the Sybil commitment generator is twice that amount, see figure 3[7].

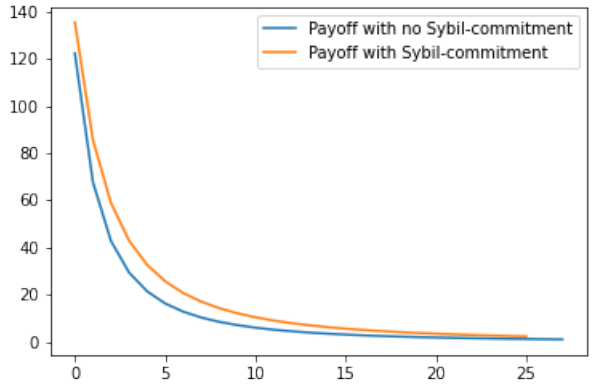

Figure 3: Aggregated arbitrage game with Sybil commitment strategy

An interpretation of this is that if a player can credibly convince the other players that there are two independent players (one himself, and the other a Sybil identity), then he has incentives to do so even if he has to commit to acting as individual rational agents. Clearly, this will decrease the social welfare, since the social welfare is $\mathcal{O}(1/n)$. In other words, the cost of

[7]Code available in `https://github.com/BrunoMazorra/CostsOfSybils`

credibly convincing the other players about the number of competitors is lower than the payoff obtained by employing the strategy. Another interpretation of Sybil commitments is that agents have the ability to change other agents' priors about the number of competitors before the game starts.

Consider a realistic scenario where agents engage in a pro-rata game as a repeated game. The agents aim to automate their strategy in a way that optimally reacts to the environment and responds to other players. Additionally, suppose that agents want to learn during the game, updating their actions each time the game is played. The automated strategy must be robust against any number of agents and capable of learning on the fly how to play optimally. For instance, in concave games, no-regret algorithms exist within the bandit feedback model. If these algorithms guarantee no-regret performance, an agent would not deviate from using this strategy to maximize their average payoff. For example, consider a scenario in a Cournot oligopoly game where $n-1$ players are using a no-regret algorithm. The last agent could create an additional identity, running two independent instances of a no-regret algorithm. This strategy would increase their expected profit, as we have previously observed, because the algorithm would converge to the Nash equilibrium considering $n+1$ agents. Therefore, this game consists of agents picking automated strategies, and then playing the repeated game by updating their actions by the rules defined by the strategy. In this game, if an automated strategy forms a symmetric Nash equilibrium for any number of agents, then the automated strategy form a Sybil commitment equilibrium.

5.2 Analysis on Welfare of Sybil Commitment proofness

In the following theorem, we will see that is very expensive for a symmetric game or mechanism with common value to be SCP (all agents have the same type). Assume that we have a social choice function f and utility functions $u_i : \mathcal{A} \times \mathcal{A}^\infty \to \mathbb{R}$ such that:

- $u_i(y_i, y_{-i}) = u_i(y_i, \sigma y_{-i})$ for all $(y, y_{-i}) \in \mathcal{A} \times \mathcal{A}^\infty$ $\sigma \in S_\infty$.
- $u_i(y_i, y_{-i}) = u_{\sigma(i)}(y_{\sigma(i)}, y_{-\sigma(i)})$ for all $i \in \mathbb{N}$ and σS_∞.

An example of this games are concave pro-rata games.

Theorem 5.1. *Consider a game with the previous properties. Assume that $c = \mathcal{O}(1/n^2)$. Then if the Sybil commitment extension game is SPC, then the welfare in equilibrium of the underlying game is $\mathcal{O}(n/2^n)$.*

Proof. Let $R = \max_{\mathbf{x}\in\mathcal{A}^\infty} W(\mathbf{x})$. Since is Sybil-proof, we have that $R = \max_{\mathbf{x}\in\mathcal{A}^1} W(\mathbf{x})$. By definition of the Sybil extension, we have that

$$u_i(\mathbf{x}_i, \mathbf{x}_{-i}) - c \geq u_i(\mathbf{y}_i, \mathbf{y}_{-i}) - c + u_j(\mathbf{y}_j, \mathbf{y}_{-j}) - c$$

for all $i \in [n]$, $j \in [n+1]$, $\mathbf{x} \in \mathrm{NE}(n)$ and $\mathbf{y} \in \mathrm{NE}(n+1)$. By symmetry of U, we can assume that $U_{n+1}(\mathbf{y}_n, \mathbf{y}_{-(n+1)}) \leq W(\mathbf{y})/(n+1)$. Fixing i, we have that

$$\begin{aligned} nu_i(\mathbf{x}_i, \mathbf{x}_{-i}) &\geq nu_i(\mathbf{y}_i, \mathbf{y}_{-i}) + \sum_{j\neq i}^{n+1} u_j(\mathbf{y}_j, \mathbf{y}_{-j}) - cn \\ &\geq (n-1)u_i(\mathbf{y}_i, \mathbf{y}_{-i}) + W(\mathbf{y}) - cn \end{aligned}$$

and so, adding all i, we have that

$$\begin{aligned} nW(\mathbf{x}) &\geq \sum_{i=1}^{n}[(n-1)u_i(\mathbf{y}_i, \mathbf{y}_{-i}) + W(\mathbf{y})] - cn^2 \\ &= (n-1)\left(W(\mathbf{y}) - U_{n+1}(\mathbf{y}_n, \mathbf{y}_{-(n+1)})\right) + nW(\mathbf{y}) - cn^2 \\ &\geq (n-1)W(\mathbf{y})(1 - \frac{1}{n+1}) + nW(\mathbf{y}) - cn^2 \end{aligned}$$

and so

$$\begin{aligned} W(\mathbf{x}) &\geq W(\mathbf{y})\left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n+1}\right) + W(\mathbf{y}) - cn \\ &= 2\left(1 - \frac{1}{n+1}\right)W(\mathbf{y}) - cn \end{aligned}$$

By induction, we have that for all equilibrium $\mathbf{x} \in NE(n)$,

$$\begin{aligned} W(\mathbf{x}) &= \frac{1}{2^{n-2}\prod_{k=2}^{n-1}\left(1-\frac{1}{k}\right)}R + \sum_{k=1}^{n-2}\frac{c(n-k)}{2^k\prod_{l=1}^{k}\left(1-1/(n-l)\right)} \\ &\leq \frac{1}{2^{n-2}\prod_{k=2}^{n-1}\left(1-\frac{1}{k}\right)}R + \frac{c}{2}n^2 = \frac{n-1}{2^{n-2}}R + \frac{c}{2}n^2 \sim \frac{n}{2^n}(R/4) \text{ as } n \to +\infty \end{aligned}$$

using that $\prod_{k=2}^{n-1}\left(1-\frac{1}{k}\right) = \frac{1}{n-1}$ and $c = \mathcal{O}(1/n^2)$. □

As a corollary, if the cost of creating Sybils is sufficiently small and the Welfare $\mathbf{W}$ in equilibrium is $o(n/2^n)$, then the game is not SCP. Also, taking the reward distribution mechanism, we have that this bound is tight. In the following, we will provide a set of popular games that are not truthful self-reporting. But first, we will give two examples of truthful self-reporting games. In the following we assume that $\mathcal{L}$ is the set of one function, the actual payoff function of the players.

- The trivial game defined by $U : \mathbb{R}_+^\infty \to \mathbb{R}_+^\infty$ with cost function $C(x) = cx$ defined as $u_i(x) = \frac{3c}{2}$ for all $i \in \mathbb{N}$ and $x \in \mathbb{R}^\infty$. This game is obviously truthful self-reporting since no action can modify the outcome and reporting two identities has payoff $-c/2$ and reporting one identity has payoff $c/2$.

- The game defined by $u_i(x_i, x_{-i}) = \frac{2x_i}{e^{x_i+x_{-i}}}$ and cost function $C(l,k) = \frac{1}{e^{l+k}}$. Observe that $x = 1$ is strictly dominant strategy. Since $x = 1$ is the unique critical point of U and U is increasing in $0 \leq x \leq 1$. If there are k reported identities, then $\tilde{U}(l,k) = \frac{l}{e^{l+k}}$ and this is maximized with $l = 1$. Therefore, the game is truthful self-reporting.

6 Conclusions

We explored the theoretical challenges of Sybil strategies in mechanism design, particularly focusing on Sybil-proof mechanisms for scenarios like reward sharing. The authors introduce the "Sybil extension mechanism" to account for the costs associated with creating multiple identities and their impact on outcomes, enhancing the understanding of how to mitigate such threats in practical settings.

Future research could explore different cost functions for generating Sybils to refine the anylisis further. Another area of interest could be the probability of a mechanism designer identifying Sybils, which could improve strategies for preventing attacks and increase the welfare

in equilibrium. Additionally, investigating the implications of agents credibly committing to their Sybils' actions could provide deeper insights into strategic interactions under manipulation and information asymmetry. These areas offer promising directions for advancing the theory and practical applications of mechanism design in the face of identity manipulations.

References

[1] Colleen Alkalay-Houlihan and Adrian Vetta. "False-name bidding and economic efficiency in combinatorial auctions". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 28. 1. 2014.

[2] Guillermo Angeris et al. "Constant function market makers: Multi-asset trades via convex optimization". In: *Handbook on Blockchain*. Springer, 2022, pp. 415–444.

[3] Mohamed Baza et al. "Detecting sybil attacks using proofs of work and location in vanets". In: *IEEE Transactions on Dependable and Secure Computing* 19.1 (2020), pp. 39–53.

[4] Steven J Brams and Alan D Taylor. "An envy-free cake division protocol". In: *The American Mathematical Monthly* 102.1 (1995), pp. 9–18.

[5] Hongyin Chen et al. "Sybil-Proof Diffusion Auction in Social Networks". In: *arXiv preprint arXiv:2211.01984* (2022).

[6] Xi Chen, Christos Papadimitriou, and Tim Roughgarden. "An axiomatic approach to block rewards". In: *Proceedings of the 1st ACM Conference on Advances in Financial Technologies*. 2019, pp. 124–131.

[7] Alice Cheng and Eric Friedman. *Manipulability of PageRank under sybil strategies*. 2006.

[8] Krishnendu Ghosh Dastidar. "Is a unique Cournot equilibrium locally stable?" In: *Games and Economic Behavior* 32.2 (2000), pp. 206–218.

[9] Jochen Dinger and Hannes Hartenstein. "Defending the sybil attack in p2p networks: Taxonomy, challenges, and a proposal for self-registration". In: *First International Conference on Availability, Reliability and Security (ARES'06)*. IEEE. 2006, 8–pp.

[10] John R Douceur. "The sybil attack". In: *International workshop on peer-to-peer systems*. Springer. 2002, pp. 251–260.

[11] Federico Fioravanti and Jordi Massó. "False-name-proof and strategy-proof voting rules under separable preferences". In: *Available at SSRN 4175113* (2022).

[12] Mingyu Guo and Vincent Conitzer. "Optimal-in-expectation redistribution mechanisms". In: *Artificial Intelligence* 174.5-6 (2010), pp. 363–381.

[13] Mathias Hall-Andersen and Nikolaj I Schwartzbach. "Game theory on the blockchain: a model for games with smart contracts". In: *International Symposium on Algorithmic Game Theory*. Springer. 2021, pp. 156–170.

[14] Sarosh Hashmi and John Brooke. "Authentication mechanisms for mobile ad-hoc networks and resistance to sybil attack". In: *2008 Second International Conference on Emerging Security Information, Systems and Technologies*. IEEE. 2008, pp. 120–126.

[15] Atsushi Iwasaki et al. "Worst-case efficiency ratio in false-name-proof combinatorial auction mechanisms". In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. 2010, pp. 633–640.

[16] Nicholas AG Johnson et al. "Concave Pro-rata Games". In: (2022).

[17] Sepandar D Kamvar, Mario T Schlosser, and Hector Garcia-Molina. "The eigentrust algorithm for reputation management in p2p networks". In: *Proceedings of the 12th international conference on World Wide Web*. 2003, pp. 640–651.

[18] Sunny King and Scott Nadal. "Ppcoin: Peer-to-peer crypto-currency with proof-of-stake". In: *self-published paper, August* 19.1 (2012).

[19] Vijay Krishna. *Auction theory*. Academic press, 2009.

[20] Brian Neil Levine, Clay Shields, and N Boris Margolin. "A survey of solutions to the sybil attack". In: *University of Massachusetts Amherst, Amherst, MA* 7 (2006), p. 224.

[21] Jian Lin et al. "Sybil-proof incentive mechanisms for crowdsensing". In: *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE. 2017, pp. 1–9.

[22] Deepak Maram et al. "Candid: Can-do decentralized identity with legacy compatibility, sybil-resistance, and accountability". In: *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2021, pp. 1348–1366.

[23] Robert C Marshall and Leslie M Marx. "Bidder collusion". In: *Journal of Economic Theory* 133.1 (2007), pp. 374–402.

[24] Robert C Marshall and Leslie M Marx. *The economics of collusion: Cartels and bidding rings*. Mit Press, 2014.

[25] Bruno Mazorra, Victor Adan, and Vanesa Daza. "Do not rug on me: Leveraging machine learning techniques for automated scam detection". In: *Mathematics* 10.6 (2022), p. 949.

[26] R Preston McAfee and John McMillan. "Bidding rings". In: *The American Economic Review* (1992), pp. 579–599.

[27] Elchanan Mossel and Omer Tamuz. "Truthful fair division". In: *International Symposium on Algorithmic Game Theory*. Springer. 2010, pp. 288–299.

[28] Wolf Müller et al. "Sybil proof anonymous reputation management". In: *Proceedings of the 4th international conference on Security and privacy in communication netowrks*. 2008, pp. 1–10.

[29] Satoshi Nakamoto. "Bitcoin: A peer-to-peer electronic cash system". In: *Decentralized Business Review* (2008), p. 21260.

[30] Jack Robertson and William Webb. *Cake-cutting algorithms: Be fair if you can*. AK Peters/CRC Press, 1998.

[31] David Cerezo Sánchez. "Zero-knowledge proof-of-identity: Sybil-resistant, anonymous authentication on permissionless blockchains and incentive compatible, strictly dominant cryptocurrencies". In: *arXiv preprint arXiv:1905.09093* (2019).

[32] Itai Sher. "Optimal shill bidding in the VCG mechanism". In: *Economic Theory* 50 (2012), pp. 341–387.

[33] Jung Ki So and Douglas S Reeves. "Defending against sybil nodes in bittorrent". In: *International Conference on Research in Networking*. Springer. 2011, pp. 25–39.

[34] Takayuki Suyama and Makoto Yokoo. "Strategy/false-name proof protocols for combinatorial multi-attribute procurement auction". In: *Autonomous Agents and Multi-Agent Systems* 11 (2005), pp. 7–21.

[35] Taiki Todo, Atsushi Iwasaki, and Makoto Yokoo. "False-name-proof mechanism design without money". In: *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 2011, pp. 651–658.

[36] Liad Wagman and Vincent Conitzer. "Optimal False-Name-Proof Voting Rules with Costly Voting." In: *AAAI*. Vol. 8. 2008, pp. 190–195.

[37] Makoto Yokoo, Yuko Sakurai, and Shigeo Matsubara. "Robust combinatorial auction protocol against false-name bids". In: *Artificial Intelligence* 130.2 (2001), pp. 167–181.

[38] Makoto Yokoo, Yuko Sakurai, and Shigeo Matsubara. "The effect of false-name bids in combinatorial auctions: New fraud in Internet auctions". In: *Games and Economic Behavior* 46.1 (2004), pp. 174–188.

[39] Haifeng Yu et al. "Dsybil: Optimal sybil-resistance for recommendation systems". In: *2009 30th IEEE Symposium on Security and Privacy*. IEEE. 2009, pp. 283–298.

[40] Haifeng Yu et al. "Sybilguard: defending against sybil attacks via social networks". In: *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*. 2006, pp. 267–278.

[41] Haifeng Yu et al. "Sybillimit: A near-optimal social network defense against sybil attacks". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 3–17.

[42] Shijie Zhang and Jong-Hyouk Lee. "Double-spending with a sybil attack in the bitcoin decentralized network". In: *IEEE transactions on Industrial Informatics* 15.10 (2019), pp. 5715–5722.