Non-Obviously Manipulable Mechanism Design for Single-Parameter Agents and Bilateral Trade

Thomas Archbold
King's College London
London, United Kingdom
thomas.archbold@kcl.ac.uk

Bart de Keijzer
King's College London
London, United Kingdom
bart.de_keijzer@kcl.ac.uk

Carmine Ventre
King's College London
London, United Kingdom
carmine.ventre@kcl.ac.uk

ABSTRACT

A recent line of work in mechanism design has focused on guaranteeing incentive compatibility for agents without contingent reasoning skills: *obviously strategyproof mechanisms* [22] guarantee that it is "obvious" for these imperfectly rational agents to behave honestly, whereas *non-obviously manipulable* (NOM) mechanisms [28] take a more optimistic view and ensure that these agents will only misbehave when it is "obvious" for them to do so. Technically, obviousness requires comparing certain extrema (defined over the actions of the other agents) of an agent's utilities for honest behaviour against dishonest behaviour.

We present a technique for designing NOM mechanisms in settings where monetary transfers are allowed based on *cycle monotonicity*, which allows us to disentangle the specification of the mechanism's allocation from the payments. By leveraging this framework, we completely characterise both allocation and payment functions of NOM mechanisms for single-parameter agents. We then look at the classical setting of bilateral trade and study how much subsidy, if any, is needed to guarantee NOM, efficiency, and individual rationality. We prove a stark dichotomy: no finite subsidy suffices if agents look only at best-case extremes, whereas no subsidy at all is required when agents focus on worst-case extremes. We conclude the paper by characterising the NOM mechanisms that require no subsidies whilst satisfying individual rationality.

KEYWORDS

Mechanism Design; Imperfect Rationality; Bilateral Trade

1 INTRODUCTION

The prevailing solution concept in mechanism design is dominant-strategy incentive-compatibility, otherwise known as strategyproofness, which stipulates that for any joint strategy of the other agents a given agent always weakly prefers to act truthfully than dishonestly. The canonical example of such a mechanism is the second-price sealed-bid auction, in which it is always in an agent's best interests to submit her true valuation for the item to the auctioneer. While this is appealing from a theoretical standpoint it relies on

the assumption that the players of the game act perfectly rationally, meaning they are able to correctly reason about the behaviour of the other players as well as solve an optimisation problem to identify an optimal strategy. This may place a high cognitive or temporal burden on the player and it can be unrealistic to assume this is how they act in the real world.

Empirical evidence in fact shows that people facing mechanisms that are not strategyproof might fail to understand that lying is beneficial. Examples include mechanisms both with and without monetary transfers. The Deferred Acceptance Algorithm, used in many countries to allocate doctors to hospitals, is known to perform well in practice and produce stable outcomes, even though hospitals could lie and hire better candidates [27]. Uniform-price auctions are a natural generalisation of second-price auctions to the multi-unit case: with a supply of k, they charge the winners the $(k + 1)$-th highest bid for each unit received. Variants of uniform-price auctions are used by search engines to charge for sponsored ad slots and governments to sell bonds. These auctions are famously not strategyproof and yet there is little evidence that bidders cheat in practice [19]. On the other hand, there are manipulable mechanisms where people can easily figure out how to lie profitably: the Boston Mechanism for school choice is known to be manipulable [24] and empirical data shows evidence of misreports [11]. Similarly, there is evidence of bid shading in pay-as-bid multi-unit auctions [18].

Motivated by these observations, Troyan and Morrill [28] study *non-obvious manipulability*, an alternative solution concept to strategyproofness for cognitively limited agents. Under this relaxation, simply referred to as NOM, a mechanism is allowed to be vulnerable to certain types of manipulations which are harder to identify by agents with bounded rationality. Instead of requiring the truthful action for each agent i to be a dominant strategy at each possible interaction with the mechanism, under non-obvious manipulability we need only to compare best-case utilities for truthful outcomes (i.e., the maximum utility, taken over the reports of the other agents, for i when she is truthful) to best-case utilities for dishonest outcomes, and worst-case utilities for truthful outcomes to worst-case utilities for dishonest outcomes. This is justified under the assumption that cognitively-limited agents can only identify this type of *obvious* manipulation where the benefits occur at the extremes of her utility function.

In this paper we study NOM mechanisms that use monetary transfers. We wish to: (i) understand the restrictions NOM imposes on the social choice function; (ii) characterise both the allocation functions and payments in NOM mechanisms for the large class of single-parameter agents; and (iii) assess the power of NOM mechanisms for two-sided markets. We develop both conceptual and technical tools to obtain answers to each of these desiderata.

1.1 Our Contribution

Our goal is to understand how permissive NOM is as a solution concept for allocation functions. NOM requires truthtelling to yield a greater utility than misreporting for every player under the best and worst outcomes of the mechanism, taken over all actions of the other players. Therefore the constraints imposed by NOM *depend on* both the allocation and payment function since they define these maximum and minimum utilities, and the utility of every other outcome must lie between these extremes. The first contribution of this work is to abstract this dependency using what we call *labellings* and use these to extend the cycle monotonicity technique to handle NOM direct-revelation mechanisms. Informally a labelling is a choice of bid profiles designating the best and worst outcomes for a given player that the mechanism will define through its allocation and payments. These induce two types of constraints: incentive-compatibility constraints, which enforce the NOM property where the extreme utilities when lying are no better than the extreme utilities from truthtelling, and labelling-induced constraints, which ensure the validity of the labels. Given an allocation function f and labelling λ for a given agent i we can then define a graph $\mathcal{G}_{i,f}^{\lambda}$ representing the constraints imposed by the mechanism and show that the existence of payments satisfying these constraints is related to cycles in this graph.

> **Theorem 1 (informal).** *An allocation function f is NOM implementable if and only if there exists a labelling λ such that $\mathcal{G}_{i,f}^{\lambda}$ has no negative-weight cycles.*

This provides a useful tool for studying the allocation functions for which we can define payments leading to NOM mechanisms. We apply this to the well-studied setting of single-parameter agents where each agent's type can be expressed as a single number, and their utility for an allocation is simply the product of their type with the amount they are allocated. Negative cycles exist in the graph $\mathcal{G}_{i,f}^{\lambda}$ when the labels are antimonotone, roughly meaning that for two types b_i and b_i' with $b_i < b_i'$ the agent is allocated more for the label corresponding to b_i than it does for the label corresponding to b_i'. We then prove that a monotone labelling guarantees no negative cycles and that such a labelling is possible whenever the image of f_i is rich enough. Roughly speaking f is *overlapping* if for each agent i and pair of types b_i, b_i' the sets $\{f_i(b_i, b_{-i})\}_{b_{-i}}$ and $\{f_i(b_i', b_{-i})\}_{b_{-i}}$ have a non-empty intersection.

> **Theorem 2 (informal).** *An allocation function f is NOM-implementable for single-parameter agents if and only if f is overlapping.*

Our second contribution proves that as long as the labels define a monotone restriction of the allocation function then it is possible to define a NOM payment scheme. This uncovers a connection with strategyproofness for single-parameter agents, where the allocation must be monotone for each fixed b_{-i}, and when this is the case it is known what form the payments must take [6]. We observe that we can take the same payments along the monotone restriction defined by the labels, while we can easily characterise the payments associated with the remaining bid profiles. This gives a complete picture on the power of NOM mechanisms for single-parameter agents.

We conclude by applying our findings to NOM mechanisms for two-sided markets where both sides of the market are strategic, a setting which is widespread but notoriously too complex for strategyproofness. We focus on bilateral trade where a single item may be sold from a seller to a buyer. We want to guarantee efficiency, where trade occurs only when the buyer's valuation is at least the seller's production cost; weak budget balance (WBB), for which we never pay the seller more than we charge the buyer; and individual rationality (IR), where the utilities of both buyer and seller are non-negative. NOM cannot guarantee these three properties simultaneously [28] so we investigate the extent to which a subsidy of $\alpha \geq 1$ guaranteeing we pay the seller at most an α factor of what we receive from the buyer (i.e., α-WBB) allows us to obtain NOM, efficiency, and IR simultaneously.

> **Theorem 3 (informal).** *Any NOM, IR and α-WBB mechanism for bilateral trade has unbounded α.*

The approach developed in the first two main theorems can be applied to characterise BNOM and WNOM independently. For our third main result we first characterise the monotone restrictions of a given allocation function that will guarantee efficiency, IR, and either BNOM or WNOM. For the former the unbounded α is implied by the incentive-compatibility constraints: using the explicit payment formulae we show that the buyer must receive the item for free while the seller must receive some positive payment in each of their respective best case profiles. For the latter it is implied by the labelling validity constraints.

We then look beyond using monotone restrictions of the allocation function to construct NOM mechanisms where the labelled profiles need not appear on a "single line". Can we be more flexible in our labelling scheme to derive different payments that would satisfy IR and finite subsidies for an allocation function that is overlapping and efficient? The answer is surprising: even non-single line BNOM mechanisms require infinite subsidies while non-single line WNOM mechanisms can guarantee WBB (i.e., no subsidies). There exists in fact a definition of WNOM labels that implement first-price auctions. We may also strengthen the notion of incentive-compatibility on either side of the market to strategyproofness – for example, the buyer is strategyproof and the seller WNOM – by combining a first-price auction on one side with a second-price auction on the other. This leaves the spread to the market designer, whereas a fully WNOM market would be budget balanced. This setting may be of interest for some classes of double-sided markets where the rationality is asymmetric and depends on the role played in the market.

We conclude by characterising individually rational NOM mechanisms that might be inefficient but require no subsidies. Intuitively, this result says that there exist trading windows defined by payment thresholds that partition the domain of buyer and seller into three sets; bids guaranteeing that the trade occurs, bids guaranteeing that the trade will not occur, and bids where the decision to trade depends on the bid at the other side of the market.

1.2 Related Work

There has been a growing body of work in the algorithmic mechanism design literature that studies agents with bounded rationality. Li [22] introduced the aforementioned notion of obvious

strategproofness (OSP) to study settings in which agents lack contingent reasoning skills, and provides two characterisations of such mechanisms, one from the perspective of contingent reasoning and the other from that of commitment power. While OSP mechanisms provide stronger incentive guarantees than strategyproof ones, they are harder to construct. Where the revelation principle allows us to restrict our attention to direct mechanisms for strategyproofness, the same is not true for OSP: the extensive-form implementation of the mechanism is crucial to a mechanism's obvious strategyproofness. The relationship between the extensive-form implementation and the allocation function of OSP mechanisms is studied in [13], where together with the introduction of the cycle monotonicity variant for OSP mechanisms with transfers, the authors give bounds on the approximation guarantee of these mechanisms for scheduling related machines. Binary allocation problems are then studied in [12] first for single-parameter agents with small domains and more recently in [14] for the general case. Further OSP mechanisms using some form of verification power of the mechanism designer are given in [15, 16, 20]. OSP mechanisms for combinatorial auctions with single-minded bidders are given in [9]. Relaxations of OSP that interpolate between OSP and strategyproofness are studied in [17].

Troyan and Morrill [28] consider agents with bounded rationality from the opposite perspective and introduce non-obvious manipulability (NOM) to study the types of manipulations available to a cognitively limited agent. In contrast to OSP, where the worst-case truthful outcome must be at least as good as the best-case dishonest outcome, NOM stipulates that both the best- and worst-case truthful outcomes are no worse than the best- and worst-case dishonest outcomes, respectively. They provide a characterisation for direct-revelation NOM mechanisms and classify existing mechanisms as either obviously manipulable (OM) or NOM in the settings of school choice, two-sided matching, and auctions. They show that a manipulation is obvious if and only if it can be recognised by a cognitively limited agent. They go on to characterise the class of NOM matching mechanisms for settings subject to both one-sided and two-sided manipulations, then show that for generalisations of first-price and second-price auctions, the former is OM while the latter is NOM. Finally they show that every efficient, IR, WBB mechanism for bilateral trade is OM.

Several papers build directly on [28]. Aziz and Lam [7] study NOM voting rules and provide conditions under which certain classes of rules are NOM, and this may depend on the number of outcomes relative to the number of voters. They also provide algorithms to compute such manipulations using polynomial-time reductions to the unweighted coalitional manipulation problem, which yield polynomial-time algorithms under the k-approval voting rule. Ortega and Segal-Halevi [23] study indirect NOM mechanisms for cake-cutting and show that relaxing strategyproofness to NOM resolves the conflict between truthtelling and fairness: unlike strategyproofness, NOM is compatible with proportionality. Like [22] they note how incentive properties may vary between theoretically equivalent mechanisms for agents with bounded rationality. Psomas and Verma [25] study fairly allocating indivisible goods to agents with additive valuations. They show the existence of deterministic NOM mechanisms which achieve envy-freeness

up to one good (EF1), and highlight a conflict for NOM under different objective functions – they provide a social welfare maximising NOM mechanism for $n \geq 3$ agents and in contrast show that any optimal mechanism for egalitarian or Nash welfare is obviously manipulable. They also provide an efficiency-preserving black box reduction from designing NOM and EF1 mechanisms to designing EF1 algorithms.

2 PRELIMINARIES

We consider mechanism design settings with the possibility of monetary transfers. There is a set of n *agents*, where each agent i has some *type* t_i that comes from a set D_i of possible types (also called agent i's *domain*). We let $D = \times_{i \in [n]} D_i$ be the set of *type profiles*. For a set of possible *outcomes* O an agent's type describes for each $o \in O$ the utility she receives from that outcome; it can therefore be thought of as a function $t_i : O \to \mathbb{R}$ mapping outcomes to numbers. For a type profile $t = (t_1, t_2, \ldots, t_n)$, player i, and type $b_i \in D_i$, we use the standard notation (b_i, t_{-i}) to denote the type profile obtained from t by replacing t_i with b_i. Furthermore we define $D_{-i} = \times_{j \in [n] \setminus \{i\}} D_j$.

The agents interact with a *mechanism M* that solicits their types and returns an outcome and a vector of payments. We focus on *direct-revelation*[1] mechanisms as in [28] in which agents directly report their types to the mechanism, and based on the type profile submitted by the agents the mechanism selects an outcome and vector of payments. We consider a mechanism $M : D \to O \times \mathbb{R}^n$ as a function mapping type profiles to outcome-payment vector pairs, where p_i describes the payment made to agent i. Agents are assumed to maximise their *utility* and are equipped with standard quasi-linear utility functions so that, on output (o, p) of the mechanism, agent i derives utility $t_i(o) + p_i$. We overload notation and denote the utility of agent i under this output as $t_i(o, p)$.

A *social choice function* $f : D \to O$ is a function that maps type profiles to outcomes, and we say that a mechanism M *implements* a social choice function f if and only if for all $t \in D$ the outcome returned by $M(t)$ is equal to $f(t)$. Note that agent types are private information, hence each agent i may misreport her type as $b_i \neq t_i$ to the mechanism. We will therefore refer to the reports as *bids* and the profile submitted to the mechanism as a *bid profile*. We use payments to realign the agents' incentives to satisfy the NOM property (defined precisely later in this section). In this paper we focus on designing NOM mechanisms that implement a given social choice function and therefore, given a social choice function f, we may view our mechanism as the tuple $M = (f, p)$ where the outcome returned by M on input b is equal to $f(b)$ for all bid profiles b, and $p : D \to \mathbb{R}^n$ is a rule defining the payments made to each agent for each bid profile b. Therefore our objective, given f, is to find payment functions p such that the resulting mechanism $M = (f, p)$ is NOM.

From Section 4 we focus on a specific class of social choice functions known as *allocation functions*. Here we assume outcomes to be real-numbered vectors of length n and refer to them as *allocations*, and assume an agent's utility for an allocation a depends only on

---

[1]In an upcoming paper [4] we show that NOM direct mechanisms are just as powerful as indirect ones for single-parameter agents; whether this is the case in general is an open problem.

the ith coordinate of a. This is done purely for interpretational purposes, since we focus on auction settings in which the mechanism assigns an amount of some object to each agent and an agent's type describes the utility she gets from that allocation.

A familiar solution concept in mechanism design is strategyproofness: a mechanism M is *strategyproof* if and only if $t_i(M(t_i, b_{-i})) \geq t_i(M(b_i, b_{-i}))$ for all $t_i, b_i \in D_i$, all $b_{-i} \in D_{-i}$, and all players i. Nonobvious manipulability, on the other hand, requires agents only to compare the extremes of their utility function. Informally stated it says that a manipulation is *obvious* if either the best-case dishonest outcome is strictly greater than the best-case truthful outcome or the worst-case dishonest outcome is strictly greater than the worst-case truthful outcome. The best and worst cases are defined over all bid profiles of the other players excluding i. As per [28] a direct-revelation mechanism is *not obviously manipulable (NOM)* if both of the following two properties hold:

$$\sup_{b_{-i}}\{t_i(M(t_i, b_{-i}))\} \geq \sup_{b_{-i}}\{t_i(M(b_i, b_{-i}))\}, \tag{1}$$

$$\inf_{b_{-i}}\{t_i(M(t_i, b_{-i}))\} \geq \inf_{b_{-i}}\{t_i(M(b_i, b_{-i}))\} \tag{2}$$

for every $t_i, b_i \in D_i$ and for every player i. Note that the b_{-i} are not necessarily the same on each side of the inequalities. If (1) holds then M is *best-case not obviously manipulable (BNOM)* and if (2) holds then M is *worst-case not obviously manipulable (WNOM)*. If either inequality is violated for some $t_i, b_i \in D_i$ then b_i is an *obvious manipulation* of M. If $M = (f, p)$ implements f and we can define payments p such that M is BNOM (respectively, WNOM) then we say that f is *BNOM-implementable* (respectively, *WNOM-implementable*), and if f is both BNOM- and WNOM-implementable then f is *NOM-implementable*.

In Sections 4 and 5 we consider several additional properties. A mechanism M is *individually rational (IR)* if $t_i(M(t_i, b_{-i})) \geq 0$ for all $t_i \in D_i$, all $b_{-i} \in D_{-i}$, and for each player i, and it makes *no positive transfers (NPT)* if $p_i(b) \leq 0$ for each player i and each $b \in D$. In Section 5 we also study efficiency and budget-balance in the context of bilateral trade, which we will define when needed.

In order to refer to specific types of some player i we enumerate her domain as $D_i = \{t_1, t_2, \ldots, t_d\}$. In Section 4 we focus on single-parameter agents where player types are treated as scalars and we assume that $t_1 < t_2 < \ldots < t_d$. We conclude this section with the following remark.

Remark 1. If mechanism M is strategyproof then it is also NOM, however the reverse is not necessarily true.

This says that there is some additional flexibility that we may be able to exploit when designing incentive-compatible mechanisms for NOM versus strategyproofness. In the following section we formalise this flexibility in the form of profile labellings and use it to derive a useful technique for designing NOM mechanisms.

3 PROFILE LABELLINGS TO MONOTONICITY

While strategyproofness requires satisfying a constraint on each pair of bid profiles (t_i, b_{-i}) and (b_i, b_{-i}) for each pair of types $t_i, b_i \in D_i$ and each $b_{-i} \in D_{-i}$, for NOM we need only compare the extremes of i's utility function. Since the social choice function f is given then we may design the payments p for some mechanism $M = (f, p)$ to define these extremes and thus effectively select

which bid profiles to compare for incentive-compatibility. We then only need to satisfy the respective BNOM and WNOM constraints for these chosen profiles, in addition to the implicit constraints imposed by this ordering on the profiles. We model this with *profile labellings*, which allow us to designate the type profiles leading to an agent's highest and lowest utilities when interacting with the mechanism, based on her true type and the type she reports.

Definition 1 (Profile labelling). Fix player i with domain D_i and let $d = |D_i|$. A *best-case labelling* β and *worst-case labelling* ω of mechanism M for player i are matrices $\beta, \omega \in |D_{-i}|^{d \times d}$ such that

$$\beta_{jk} \in \operatorname*{arg\,sup}_{b_{-i} \in D_{-i}}\{t_j(M(t_k, b_{-i}))\},$$
$$\omega_{jk} \in \operatorname*{arg\,inf}_{b_{-i} \in D_{-i}}\{t_j(M(t_k, b_{-i}))\}.$$

The entry in row j and column k of β (respectively, ω) therefore represents the partial bid profile that, when submitted to mechanism M along with player i's bid t_k, results in i's greatest (respectively, least) utility when she has type t_j. Often we will use λ to denote either a best- or worst-case labelling of some mechanism. For a labelled profile λ_{jk} for player i, given that we can infer i's bid by looking at the second subscript we will use λ_{jk} to refer to the (full) bid profile (t_k, λ_{jk}) for brevity. When dealing with labellings for multiple players as in Section 5 we will use a superscript to differentiate between the two, i.e. λ^i and λ^j for players i and j, respectively.

Profile labellings induce two types of constraint. The *labelling constraints* induced by $\lambda \in \{\beta, \omega\}$ ensure that a labelled bid profile leads to the extreme values of i's utility function when she has type t_j and bids type t_k:

$$t_j(M(\beta_{jk})) \geq t_j(M(t_k, b_{-i})) \quad \text{for all } b_{-i} \in D_{-i}, \tag{3}$$

$$t_j(M(\omega_{jk})) \leq t_j(M(t_k, b_{-i})) \quad \text{for all } b_{-i} \in D_{-i}, \tag{4}$$

for all $t_j, t_k \in D_i$. Meanwhile the *incentive-compatibility constraints* ensure that the extreme values of i's utility function when she reports truthfully to the mechanism are no worse than the extreme values of her utility function when she misreports her type. Since the label λ_{jk} denotes a bid profile where i has type t_j and bids type t_k then for both best- and worst-case labellings we have:

$$t_j(M(\lambda_{jj})) \geq t_j(M(\lambda_{jk})) \quad \text{for all } t_j, t_k \in D_i. \tag{5}$$

It is worth emphasising that the social choice function f is fixed and that we will use the profile labellings to decide whether payments exist such that the resulting mechanism is NOM. Not every labelling will admit such payments since the above constraints may be unsatisfiable.

We use the labellings to define a graph in order to characterise the class of NOM-implementable social choice functions. This is achieved using the cyclic monotonicity technique in a similar manner to Rochet [26], Lavi and Swamy [21], and Ventre [29], among others. For some social choice function f with labelling $\lambda \in \{\beta, \omega\}$ for player i we construct the weighted directed multigraph $\mathcal{G}^\lambda_{i,f} = (V, E^\lambda, w)$ whose node set $V = D$ is the domain of f and whose edge set encodes the constraints imposed by a NOM mechanism for f. The precise edges in E^λ will vary depending on whether λ is a best- or worst-case labelling. We annotate each edge of the graph with a type from i's domain and for two nodes x and y connected

by an edge annotated with type t we write $(x, y; t)$, or equivalently $x \rightarrow^t y$. The weight of this edge is $w(x, y; t) = t(f(x)) - t(f(y))$. The weight $w(C)$ of a cycle C is simply the sum of edge weights on each edge in the cycle.

The edges of the graph represent the constraints in (3), (4), and (5). For any labelling λ the edge set E^λ contains the edges $\{\lambda_{jj} \rightarrow^{t_j} \lambda_{jk} : t_j, t_k \in D_i\}$ representing the incentive-compatibility constraints. The labelling constraints for β and ω are the reverse of each other and we have $\{\beta_{jk} \rightarrow^{t_j} (t_k, b_{-i}) : t_j, t_k \in D_i, b_{-i} \in D_{-i}\} \subset E^\beta$ and $\{(t_k, b_{-i}) \rightarrow^{t_j} \omega_{jk} : t_j, t_k \in D_i, b_{-i} \in D_{-i}\} \subset E^\omega$.

For any labelling λ of f for player i, given some bid $t_k \in D_i$ every edge encoding a labelling constraint is contained within the subgraph of $\mathcal{G}_{i,f}^\lambda$ that consists only of nodes (t_k, b_{-i}), i.e. profiles where i bids t_k. We refer to this subgraph as the "k-island" of $\mathcal{G}_{i,f}^\lambda$. Thus all labelling constraint edges for profile λ_{jk} are contained in the k-island of $\mathcal{G}_{i,f}^\lambda$, while incentive-compatibility constraint edges are those which for any two types $t_j, t_k \in D_i$ cross from the j-island to the k-island of $\mathcal{G}_{i,f}^\lambda$. This graph allows us to use the cycle monotonicity technique to characterise the social choice functions that are NOM-implementable. The proof appears in the full version of the paper [3] and largely follows standard arguments from the literature, with some extra care required in the proof of sufficiency. We note that the result holds only for finite domains.

Theorem 1. *Social choice function f is BNOM-implementable (respectively, WNOM-implementable) if and only if for each agent i there exists a best-case labelling β (respectively, worst-case labelling ω) such that $\mathcal{G}_{i,f}^\beta$ (respectively, $\mathcal{G}_{i,f}^\omega$) contains no negative weight cycles.*

Certain cycles are guaranteed to exist in $\mathcal{G}_{i,f}^\beta$ and $\mathcal{G}_{i,f}^\omega$ regardless of the specific labelling. We identify two classes of such cycles: those contained within a j-island of $\mathcal{G}_{i,f}^\lambda$, which purely describe labelling constraints, and those that traverse different islands, which correspond to incentive-compatibility constraints. Intuitively the former class of cycles appears because given a bid t_j of player i, each labelled profile for this bid and some true type from i's domain must yield a higher (respectively, lower) utility from the mechanism than every other profile where i bids t_j, including other profiles in the j-island labelled with a different true type. We can therefore select any number of true types from D_i and form a cycle between their corresponding labelled profiles in the j-island. Take any p types $t_1, t_2, \ldots, t_p \in D_i$. There always exist cycles C_p^β and C_p^ω of length p in the j-island with weight:

$$w(C_p^\beta) = \sum_{k \in [p]} t_k(f(\beta_{k,j})) - t_k(f(\beta_{k+1,j})), \tag{6}$$

$$w(C_p^\omega) = \sum_{k \in [p]} t_k(f(\omega_{k-1,j})) - t_k(f(\omega_{k,j})). \tag{7}$$

For the latter class the labelled truthful profiles must yield a greater utility than a labelled profile where i bids, say, t_j dishonestly (to uphold the NOM property), and this dishonest labelled profile must then yield a higher (respectively, lower) utility than every other labelled profile in the j-island. We can again choose any p types and alternate between traversing incentive-compatibility edges and

labelling edges, yielding cycles C_p^β and C_p^ω of length $2p$ with weight:

$$w(C_p^\beta) = \sum_{k \in [p]} t_k(f(\beta_{k,k})) - t_k(f(\beta_{k+1,k+1})), \tag{8}$$

$$w(C_p^\omega) = \sum_{k \in [p]} t_k(f(\omega_{k-1,k})) - t_k(f(\omega_{k,k+1})). \tag{9}$$

In both cases the indices -1 and $p + 1$ wrap around to p and 1 respectively. The weights above must be non-negative for any labellings β and ω. In the next section we will use these in our characterisation of implementable functions for single-parameter agents.

4 SINGLE-PARAMETER AGENTS

In this section and the remainder of the paper we apply our analysis to single-parameter agents. In this setting types can be described by a single number: when agent i has type t_i she values the outcome of f on input b as $t_i(f(b)) = t_i \cdot f_i(b)$, i.e. the product of her "value per unit" and her allocation, which allows us to decouple an agent's type from the outcome of the mechanism to prove properties of the latter in isolation. This approach has seen much success in the literature: Archer and Tardos [6] use it to show that designing truthful mechanisms reduces to designing monotonic algorithms (whether they are increasing or decreasing depends on whether agent types describe costs or valuations), and in such instances we can derive explicit formulae for truthful payments using the area under the allocation curve. Other settings for single-parameter agents include combinatorial auctions [5], machine scheduling [1], and payment computation [10].

Our goal is to characterise the class of NOM-implementable allocation functions for single-parameter agents. In the case of strategyproofness it is well known that the allocation rule must be monotone in each player's bid for each fixed b_{-i}. For our setting the shape of the allocation function will depend on the labelling, so we use the guaranteed cycles of the previous section to argue some necessary properties of any labelling. Taking $p = 2$ and any $t_1, t_2 \in D_i$, the weights of all guaranteed labelling cycles in (6) and (7) must be $(t_1 - t_2)(f_i(\beta_{11}) - f_i(\beta_{22})) \geq 0$ and $(t_1 - t_2)(f_i(\omega_{21}) - f_i(\omega_{12})) \geq 0$. Similarly, for all guaranteed incentive-compatibility cycles in (8) and (9) we must have $(t_1 - t_2)(f_i(\beta_{1j}) - f_i(\beta_{2j})) \geq 0$ and $(t_1 - t_2)(f_i(\omega_{2j}) - f_i(\omega_{1j})) \geq 0$ for any bid $t_j \in D_i$. We need these inequalities to be satisfied for any two types from player i's domain. The following definition describes the "shape" of the allocation functions guaranteeing this and hence the allocation functions for which NOM payment schemes exist, allowing us to recover an analogue to the monotonicity property for strategyproof mechanisms with the theorem that follows.

Definition 2 (Overlapping). *Let $O_{ij} = \{f_i(t_j, b_{-i}) : b_{-i} \in D_{-i}\}$ denote the set of allocations of f for player i when she submits bid t_j, and let $O_i = \times_{j \in D} O_{ij}$. Allocation function f is overlapping for i if there exists $\mathbf{o} = (o_1, \ldots, o_d) \in O_i$ such that $o_1 \leq o_2 \leq \ldots \leq o_d$, and f is overlapping if it is overlapping for each player $i \in [n]$.*

Theorem 2. *Allocation function f is either BNOM- or WNOM-implementable for single-parameter agents if and only if it is overlapping.*

Due to space restrictions we omit the proof and note that it can be found in the full version of the paper. As a corollary of Theorem 2 we can show that as long as f is overlapping then it is straightforward to achieve both BNOM and WNOM simultaneously. The only extra condition imposed is that unless the function f is constant for player i on bid t_j – that is, it allocates the same amount $f_i(t_j, \cdot)$ to i whenever she bids t_j, no matter the bid profile of the other players – then the best- and worst-case labelled profiles must be different.

We use the fact that f is overlapping to prove sufficiency in Theorem 2 by defining a particular "simple" labelling in which the best- and worst-case outcomes for a given bid of player i all occur at the same bid profile, no matter her type. Such labellings, which we call *single-line labellings*, allow us to precisely characterise the payments that lead to incentive-compatibility.

Definition 3 (Single-line mechanism). *A labelling λ is single-line if for every player i, for every bid $t_j \in D_i$ we have $\lambda_{kj} = \lambda_{\ell j}$ for every pair of types $t_k, t_\ell \in D_i$. A single-line mechanism is one whose payments are defined on a social choice function f whose labelling is single-line.*

When a labelling λ is single-line then we will index it simply by the agent's bid: if player i bids t_j in a single-line mechanism then labelled profile associated with this bid is denoted λ_j. After using these labellings to derive explicit payment formulae for single-line mechanisms we will then show in Section 5 that more general labellings can lead to different payments.

Archer and Tardos [6] show that a work function that allocates chores (where types describe *costs* as opposed to valuations) admits truthful payments if and only if it is non-increasing and characterise these payments using the area under the allocation function. Apt and Heering [2] recently provided an elementary proof of the uniqueness of these payment functions using a little-known result from the theory of real functions. It is straightforward to derive a similar result in our setting for single-line mechanisms: we show that NOM payments exist if and only if one can find a non-decreasing collection of outcomes for the best- and worst-cases of the function. Fixing such a labelling λ, we may treat f as a function only in player i's bid b_i and derive the following payments by solving the resulting differential equations from [6] in the same way.

Theorem 3. *Social choice function f is NOM-implementable if and only if f is overlapping. When this is the case and the resulting mechanism (f, p) uses a single-line labelling, the payments p take the form*

$$p_i(t_j, \lambda_j) = h_i(\lambda) - t_j f_i(t_j, \lambda_j) + \int_0^{t_j} f_i(u, \lambda_u) \, \mathrm{d}u, \qquad (10)$$

for each agent i, where $h_i(\lambda) = h_i(\lambda_1, \ldots, \lambda_d)$ is an arbitrary function that depends only on the (best- or worst-case) labelling $\lambda \in \{\beta, \omega\}$.

Requiring IR and NPT further restricts the above payments. When agents value the outcome of f we have $h_i(\beta) = h_i(\omega) = 0$, while when they incur some cost then $h_i(\beta) \geq \int_0^{t_j} f_i(u, \beta_u) \, \mathrm{d}u$ and $h_i(\omega) \geq \int_0^{t_j} f_i(u, \omega_u) \, \mathrm{d}u$ for all $b \in D_i$. The remaining payments (i.e., for bid profiles which are not best- or worst-case inputs for player i) must simply satisfy the labelling constraints induced

by β and ω. For each $t_j, t_k \in D_i$ we simply require $t_j(f_i(\beta_j)) + p_i(\beta_k) \geq t_j(f_i(t_k, b_{-i})) + p_i(t_k, b_{-i})) \geq t_j(f_i(\omega_k)) + p_i(\omega_k)$ for each $b_{-i} \in D_{-i}$. It is important to note the transition from finite to infinite domains, done both to mirror the setting of [6] and to use in the upcoming section. Conceptually this is not an issue and in the case of finite domains we can simply replace the above integral with a summation.

5 BILATERAL TRADE

We now apply our previous analysis to bilateral trade. In this setting there are two agents, B and S, where B is a potential buyer of an item that S may produce and sell. The buyer has a valuation $v \in D_B$ for the item and the seller incurs a cost $c \in D_S$ for producing the item. We assume $D_B = D_S = [0, 1]$, and let $D = D_B \times D_S$. A mechanism for bilateral trade is a tuple $M = (f, p)$ where $f : D \to \{0, 1\}$ indicates whether a trade takes place and $p : D \to \mathbb{R}_{\geq 0}$ describes the payments. The buyer's utility is therefore given by $u_B(M(x, y)) = v \cdot f(x, y) - p_B(x, y)$ and the seller's by $u_S(M(x, y)) = p_S(x, y) - c \cdot f(x, y)$.

In addition to individual rationality (defined in Section 2) we focus on the following properties: a mechanism $M = (f, p)$ is *efficient* when $f(x, y) = 1$ if and only if $x \geq y$ for all $(x, y) \in D$; and *weakly budget balanced* if $p_S(x, y) \leq p_B(x, y)$ for all $(x, y) \in D$. Troyan and Morrill [28] show that every efficient, individually rational, weakly budget balanced mechanism for bilateral trade is obviously manipulable, hence we relax budget balance and say that M is α-*weakly budget balanced* if $p_S(x, y) \leq \alpha \cdot p_B(x, y)$ for $\alpha \geq 1$. We begin by showing that single-line mechanisms cannot resolve the impossibility result with any bounded subsidy.

5.1 A single line induces unbounded subsidy

We first note the form taken by the explicit payment formulae for single-line mechanisms from Theorem 3. In this setting the buyer must pay $p_B(x, \lambda_x^B) = h_B(\lambda^B) + x f(x, \lambda_x^B) - \int_0^x f(u, \lambda_u^B) \, \mathrm{d}u$ while the seller must receive $p_S(\lambda_y^S, y) = h_S(\lambda^S) + y f(\lambda_y^S, y) - \int_0^y f(\lambda_u^S, u) \, \mathrm{d}u$, where λ^B is any labelling for the buyer (and thus) represents a bid of the seller and λ^S is any labelling for the seller (and thus represents a bid of the buyer). Again notice that if we have IR and NPT then $h_B(\lambda^B)$ goes to zero, while for the seller $h_S(\lambda^S)$ must be at least the area under the curve.

The following facts will be useful. Let x and y be bids of the buyer and seller, respectively. For any efficient IR mechanism we require $p_B(x, y) \leq x$ and $p_S(x, y) \geq y$ for all $x \geq y$, while for $x < y$ we have $p_B(x, y) = 0$. Therefore whenever trade does not take place the utility of the buyer is 0 no matter her valuation. In Lemma 1 we show that for a single-line mechanism the curves representing the best- and worst-case outcomes of f are constant. Since the buyer values the item while the seller incurs a cost for it then by Theorem 2 the best- and worst-case outcomes of f must be non-decreasing for the buyer and non-increasing for the seller. If $M = (f, p)$ is NOM and f takes values in $\{0, 1\}$ then there must be a threshold at which, for all bids that are at least this threshold for some player, their best- or worst-case allocation from the mechanism flips from one value to the other. We formalise this in the following lemma and show for both the buyer and the seller that these thresholds must be placed at the extremes, i.e. 0 or 1.

Lemma 1. *In any efficient, IR, single-line BNOM mechanism the best-case outcomes always occur for both the buyer and the seller when the trade is executed, while in any efficient, IR, single-line WNOM mechanism the worst-case outcomes always occur for both the buyer and the seller when the trade is not executed.*

Proof. Let $M = (f, p)$ be an efficient, IR, single-line NOM mechanism and first consider the threshold for best-case outcomes from the buyer's side. Since f is non-decreasing let θ_B be a threshold bid of the buyer such that $f(x, \beta_x^B) = 1$ for all $x \geq \theta_B$. Now take $x < \theta_B$ and let y be a bid of the seller such that $f(x, y) = 1$. By the BNOM labelling constrains we have $u_B(M(x, \beta_x^B)) = 0 \geq v - p_B(x, y)$ for all $v \in D_B$, so $1 \leq p_B(x, y) \leq x$, where the second inequality follows from IR and efficiency. Now we have $1 \leq x < \theta_B$ where $\theta_B \in [0, 1]$, hence there is no x such that $x < \theta_B$. Thus $\theta_B = 0$ and trade always occurs in the buyer's best-case outcome.

Now consider the seller. Since f must be non-increasing in her bid there again must be a threshold $\theta_S \in [0, 1]$ such that $f(\beta_y^S, y) = 1$ for all $y \leq \theta_S$ and 0 otherwise. For some bid $y > \theta_S$ of the seller her best case utility is $u_S(M(\beta_y^S, y)) = p_S(\beta_y^S, y) = h_S(\beta^S) + yf(\beta_y^S, y) - \int_0^y f(\beta_u^S, u)\, du$. Since c describes the seller's cost then $h_S(\beta^S)$ must be at least the area under the curve. Setting $h_S(\beta^S)$ to this area gives $p_S(\beta_y^S, y) = yf(\beta_y^S, y) + \int_y^1 f(\beta_u^S, u)\, du$ so her best-case utility is 0.[2] Now take any x such that $f(x, y) = 1$. By the BNOM constraints we have $u_S(M(x, y)) = p_S(x, y) - c \leq 0$ for all c. Combining this with individual rationality we have $c \geq p_S(x, y) \geq y$, which gives us a contradiction if we take $c < y$. Thus $y \leq \theta_S$ for all y and so $\theta_S = 1$, meaning trade always occurs in the seller's best-case outcome.

We can apply similar reasoning to show that trade does not occur for the buyer or seller in their worst-case outcome, no matter their bid. Full details are provided in the full version of the paper. □

As a corollary of Lemma 1 we have $p_B(x, \omega_x^B) = p_S(\omega_y^S, y) = 0$ for all x, y since trade does not occur at the worst-case profiles. Conversely the buyer's best outcome must be when she gets the item and pays nothing, while the seller's must occur when she produces the item and is paid at least her bid. For any IR, NPT, single-line BNOM mechanism M best-case payments are as in (10), so $p_B(x, \beta_x^B) = 0$ and $p_S(\beta_y^S, y) \geq y + \int_y^1 f(\beta_u^S, u)\, du = 1$. The following result shows that single-line mechanisms are not flexible enough to avoid the impossibility of [28] without unbounded subsidy.

Theorem 4. *Any efficient, IR, single-line BNOM or WNOM mechanism with α-WBB has unbounded α.*

Proof. We first prove the result for BNOM mechanisms. From Lemma 1 we know that $f(x, \beta_x^B) = f(\beta_y^S, y) = 1$ for any $x \in D_B$ and $y \in D_S$, while $p_B(x, \beta_x^B) = 0$ and $p_S(\beta_y^S, y) \geq y$. Since in the best case the buyer pays zero while the seller receives at least her cost, we may prove the claim by providing a bid profile that appears on the best-case curves of both the buyer and seller. Take any bid x of the buyer and let $y = \beta_x^B > 0$ be a bid of the seller. Since $f(x, y) = 1$ then $p_S(x, y) \geq y$ by efficiency and IR. The labelling constraints

require $u_S(M(x, y)) \geq u_S(M(\beta_y^S, y))$ for all y and so $p_S(\beta_y^S, y) \geq p_S(x, y) \geq y$. Since $y = \beta_x^B$ then $p_S(\beta_y^S, y) \geq p_S(x, \beta_x^B) \geq y > 0$, while $p_B(x, \beta_x^B) = 0$. Therefore at (x, β_x^B) the buyer pays 0 while the seller must receive some positive amount.

Now let $M = (f, p)$ be an efficient, IR, single-line WNOM mechanism and consider a bid $x \in D_B$ of the buyer for which $f(x, \omega_x^B) = 0$ and thus $p_B(x, \omega_x^B) = u_B(M(x, \omega_x^B)) = 0$. Let $y \in D_S$ be a bid of the seller such that $f(x, y) = 1$ and note that by efficiency $x \geq y$. Since the worst utility of the buyer is 0 then $p_B(x, y) \leq v$ and thus we need $p_B(x, y) \leq \min\{x : x \geq y\} = y$. Now consider the seller with bid $y \in D_S$ such that $f(\omega_y^S, y) = 0$. Let $x \in D_B$ be a bid of the buyer such that $f(x, y) = 1$. By the labelling constraints we have $p_S(x, y) - c \geq p_S(\omega_y^S, y) \geq 0$, hence $p_S(x, y) \geq c$ for all c (since M is single-line) and thus $p_S(x, y) \geq \max\{y : x \geq y\} = y$.

Now consider the bid profile $(1, \varepsilon)$. The buyer must pay at most ε while the seller is to receive at least 1, and since ε is arbitrary we get the desired result. □

5.2 Avoiding subsidies by changing the labelling

Single-line mechanisms are overly restrictive if we want to design efficient, IR, and α-WBB NOM mechanisms for bilateral trade, so we now explore if the issue of unbounded subsidy is avoided by considering more general labellings and hence payments. We show that while this works in a strong sense for WNOM mechanisms, where we can achieve exact (i.e. $\alpha = 1$) budget balance, surprisingly unbounded subsidies are inherent in any BNOM mechanism. For the following two results it is enough to consider domains $D_B = [0, 1]$ and $D_S = \{0, 1\}$.

Theorem 5. *There is an efficient, IR, WBB, WNOM mechanism.*

Proof. Consider the labelling ω^B for the buyer such that $\omega_{jk}^B = 1$ if and only if $t_j \geq t_k$, and 0 otherwise. Observe that for all $x \in D_B$ every cycle within the x-island of $\mathcal{G}_{B,f}^\omega$ have non-negative weight: the x-island contains exactly two nodes $(x, 0)$ and $(x, 1)$, and ω^B ensures that each two-cycle has weight at least $x(f(x, 0) - f(x, 1)) + (x - \varepsilon)(f(x, 1) - f(x, 0)) = \varepsilon > 0$ (i.e. for $t_j = x$ and $t_k = x - \varepsilon$). Recall that incentive-compatibility edges are the only ones to traverse different islands, and note with the following that incorporating these into a cycle of labelling edges results in no negative cycles. The shortest such cycle has the form $(x, 1) \to^x (x', 0) \to^{x'} (x', 1) \to^{x'} (x, 1)$ for $x' > x$, and its weight is minimised when $x' = x + \varepsilon$ for some $\varepsilon > 0$; the weight of the cycle therefore always non-negative. It is straightforward to verify that these are the only types of cycle induced by the labelling, and that increasing the length of the cycle will only increase its weight. Thus f is WNOM-implementable. It now remains to derive the payments this labelling produces.

Fix a bid t_j of the buyer. Since $\omega_{jj} = 1$ then trade does not occur and we have $u_B(M(t_j, 1)) = 0$ for all v. Since $\omega_{ij} = 0$ for $t_i < t_j$ then trade occurs and we have $v - p_B(t_j, 0) \leq u_B(M(t_j, 1)) = 0$. Thus for all $v < t_j$ we have $p_B(t_j, 0) \geq v$ and so $p_B(t_j, 0)$ tends to t_j. Combining this with IR we have $t_j \leq p_B(t_j, 0) \leq t_j$ and therefore when the trade occurs the buyer pays her bid, and 0 otherwise. Therefore this labelling induces first-price payments for the buyer. Now consider payments for the seller where she gets paid her own bid when trade occurs. It is straightforward to verify that these

---

[2]We can of course pay the seller more, as long as it doesn't interfere with incentives. We will be using this lemma to prove the impossibility of (bounded) α-WBB, so if we can show this when we are being frugal then the claim will hold even when using more generous payments. Thus we set $h_S(\beta)$ to be as small as possible, i.e. exactly the area under the curve, and this is enough to prove the unboundedness.

payments for the seller are strategyproof and therefore NOM by Remark 1. When trade occurs the payments are $p_B(x, y) = x$ and $p_S(x, y) = y$, and 0 otherwise. Since trade occurs only when $x \geq y$ then the payment given to the seller never exceeds the amount extracted from the buyer, hence we achieve (exact) weak budget balance. □

With the following we show that moving away from single-line mechanisms does not avoid the issue of unbounded subsidies for BNOM as it does for WNOM. The proof, which we omit in the interest of space, is by case analysis on all possible labellings between two types of the buyer's domain and reveals that no labelling results in a mechanism that bypasses the negative result, even if we restrict to these two types. Each labelling falls into one of four cases: it is single-line, for which we have already proved subsidies are unbounded; it is invalid and leads to a negative cycle, meaning NOM payments are impossible by Theorem 1; it induces no negative cycles but otherwise leads to contradiction when taken with IR and efficiency; or it leaves us with the same payments as in the single-line case, meaning unbounded subsidies are unavoidable.

Theorem 6. *There is no efficient, IR, α-WBB, BNOM mechanism with bounded α.*

5.3 Characterising IR, WBB, NOM bilateral trade mechanisms

We close our analysis of NOM bilateral trade mechanisms by providing a full characterisation of the class of IR, WBB, NOM mechanisms for this setting. We introduce some notation to conveniently formulate our characterisation. For a number x we denote by $\text{succ}_B(x)$ and $\text{prec}_B(x)$ respectively the least element in D_B strictly greater than x and the greatest element strictly less than x. We use $\text{succ}_S(y)$ and $\text{prec}_S(y)$ analogously for the seller's domain D_S. Let H_B denote the least element in D_B that is greater than every element in D_S, and L_B the greatest element in D_B that is less than every element in D_S. For a mechanism we may partition the type space into three sets: M_0^B, M_1^B, and M_{01}^B. M_0^B consists of all bids of the buyer for which trade is guaranteed not to happen, M_1^B consists of all bids for which trade surely happens, and M_{01}^B consists of the remaining elements of D_B, where trade depends on the seller's bid. The sets M_0^S, M_1^S, and M_{01}^S can be defined analogously for the seller.

For a mechanism M define the *utility interval* $I_B(M, v, x)$ for the buyer as the two numbers (ℓ, h) such that $\ell = \min\{u_B(M(x, y)) : y \in D_S\}$ and $h = \max\{u_B(M(x, y)) : y \in D_S\}$, and define $I_S(M, c, y)$ analogously for the seller. The NOM property imposes that $I_B(M, v, v) \geq I_B(M, v, x)$ for all $v, x \in D_B$ and that $I_S(M, c, c) \geq I_S(M, c, y)$ for all $c, y \in D_S$.

Theorem 7. *A deterministic bilateral trade mechanism M for domains D_B, D_S is IR, WBB, and NOM if and only if there are thresholds $p_B^{\min}, p_B^{\max}, p_S^{\min}, p_S^{\max}$, where $p_B^{\min} \leq p_B^{\max}$ and $p_S^{\min} \leq p_S^{\max}$, such that:*

(1) *When trade doesn't occur the buyer's and the seller's price are both zero. When trade occurs the buyer's price always exceeds the seller's price, both prices are non-negative, the buyer's price is less than their valuation, and the seller's price is at least the buyer's valuation.*

(2) *The set M_0^B contains all types in D_B less than $p_B^{\min}$, the set M_1^B contains all types in D_B greater than $p_B^{\max}$, and M_{01}^B contains all types in D_B in between $p_B^{\min}$ and $p_B^{\max}$. If $p_B^{\min}$ is itself in D_B then $p_B^{\min}$ is in either M_0^B or M_{01}^B. Similarly if $p_B^{\max}$ is in D_B then $p_B^{\max}$ is in either M_1^B or M_{01}^B.*

(3) *For every type $x \in M_{01}^B \cup M_1^B$ the buyer's payment in outcome $M(x, y)$ is at least $p_B^{\min}$ for all $y \in D_S$, and there exists some $y \in D_S$ such that in outcome $M(x, y)$ trade occurs and the buyer pays $p_B^{\min}$. For every type $x \in M_1^B$ the buyer's payment in outcome $M(x, y)$ is at most $p_B^{\max}$ for all $y \in D_S$, and there exists a $y \in D_S$ such that in outcome $M(x, y)$ trade occurs at price $p_B^{\max}$.*

(4) *(Analogous to point 2 but for the seller) The set M_0^S includes all types in D_S greater than $p_S^{\max}$, the set M_1^S includes all types in D_S less than $p_S^{\min}$, and the set M_{01}^S consists of all types in D_S in between $p_S^{\min}$ and $p_S^{\max}$. If $p_S^{\min}$ itself is in D_S, then $p_S^{\min}$ is either in M_1^S or M_{01}^S. Similarly, if $p_B^{\max}$ is in D_S then $p_B^{\max}$ is in M_{01}^S or M_1^S.*

(5) *(Analogous to point 3 but for the buyer) For every type $y \in M_{01}^S \cup M_1^S$ there is some $x \in D_B$ such that in outcome $M(x, y)$ trade occurs with seller's price $p_S^{\max}$. For every type $y \in M_1^S$ there is furthermore a $x \in D_B$ such that in outcome $M(x, y)$ trade occurs at price $p_S^{\min}$.*

6 CONCLUSIONS

This paper introduces a framework for designing NOM mechanisms with payments. We formalise the added flexibility of such mechanisms in the form of labellings which allows us to characterise the set of implementable allocation functions. We apply our analysis to bilateral trade and show a surprising dichotomy between BNOM and WNOM with regards to budget balance – we provide an efficient, IR, α-WBB WNOM mechanism and show that no BNOM mechanism exists with bounded α. In light of our results we believe that NOM mechanisms deserve to be better understood, and reasoning about the labellings inherent to such mechanisms appears to be a useful way to do so. There are a number of settings in the mechanism design literature to which our framework could be applied next, particularly the various auction settings not covered by [28]. Following [7] it would of course be useful to explore computational aspects to these mechanisms and further how the two sides of NOM relate to one another. Our general characterisation can also be helpful in better understanding multidimensional domains for which strategyproofness is known be restrictive [8].

REFERENCES

[1] Nir Andelman and Yishay Mansour. 2006. A Sufficient Condition for Truthfulness with Single Parameter Agents. In *Proceedings of the 7th ACM Conference on Electronic Commerce* (Ann Arbor, Michigan, USA) *(EC '06)*. Association for Computing Machinery, New York, NY, USA, 8–17. https://doi.org/10.1145/1134707.1134709

[2] Krzysztof R. Apt and Jan Heering. 2022. Characterization of incentive compatible single-parameter mechanisms revisited. *CoRR* abs/2212.02344 (2022). https://doi.org/10.48550/arXiv.2212.02344 arXiv:2212.02344

[3] Thomas Archbold, Bart de Keijzer, and Carmine Ventre. 2022. Non-Obvious Manipulability for Single-Parameter Agents and Bilateral Trade. https://doi.org/10.48550/ARXIV.2202.06660

[4] Thomas Archbold, Bart de Keijzer, and Carmine Ventre. 2022. Non-Obvious Manipulability in Extensive-Form Mechanisms: the Revelation Principle for Single-Parameter Agents. Manuscript under review.

[5] Aaron Archer, Christos Papadimitriou, Kunal Talwar, and Éva Tardos. 2003. An Approximate Truthful Mechanism for Combinatorial Auctions with Single Parameter Agents. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (Baltimore, Maryland) *(SODA '03)*. Society for Industrial and Applied Mathematics, USA, 205–214.

[6] Aaron Archer and Éva Tardos. 2001. Truthful mechanisms for one-parameter agents. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. 482–491. https://doi.org/10.1109/SFCS.2001.959924

[7] Haris Aziz and Alexander Lam. 2021. Obvious Manipulability of Voting Rules. In *Algorithmic Decision Theory*, Dimitris Fotakis and David Ríos Insua (Eds.). Springer International Publishing, Cham, 179–193.

[8] George Christodoulou, Elias Koutsoupias, and Annamária Kovács. 2021. On the Nisan-Ronen conjecture. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*. IEEE, 839–850. https://doi.org/10.1109/FOCS52979.2021.00086

[9] Bart de Keijzer, Maria Kyropoulou, and Carmine Ventre. 2020. Obviously Strategyproof Single-Minded Combinatorial Auctions. In *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 168)*, Artur Czumaj, Anuj Dawar, and Emanuela Merelli (Eds.). Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 71:1–71:17. https://doi.org/10.4230/LIPIcs.ICALP.2020.71 ISSN: 1868-8969.

[10] Shahar Dobzinski and Shiri Ron. 2021. *The Communication Complexity of Payment Computation*. Association for Computing Machinery, New York, NY, USA, 933–946. https://doi.org/10.1145/3406325.3451083

[11] Umut Dur, Robert G. Hammond, and Thayer Morrill. 2018. Identifying the harm of manipulable school-choice mechanisms. *American Economic Journal: Economic Policy* 10 (2018), 187–213.

[12] Diodato Ferraioli, Adrian Meier, Paolo Penna, and Carmine Ventre. 2019. Automated Optimal OSP Mechanisms for Set Systems. In *Web and Internet Economics*, Ioannis Caragiannis, Vahab Mirrokni, and Evdokia Nikolova (Eds.). Springer International Publishing, Cham, 171–185.

[13] Diodato Ferraioli, Adrian Meier, Paolo Penna, and Carmine Ventre. 2019. Obviously Strategyproof Mechanisms for Machine Scheduling. In *27th Annual European Symposium on Algorithms (ESA 2019) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 144)*, Michael A. Bender, Ola Svensson, and Grzegorz Herman (Eds.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 46:1–46:15. https://doi.org/10.4230/LIPIcs.ESA.2019.46

[14] Diodato Ferraioli, Paolo Penna, and Carmine Ventre. 2022. Two-Way Greedy: Algorithms for Imperfect Rationality. In *Web and Internet Economics (Lecture Notes in Computer Science)*, Michal Feldman, Hu Fu, and Inbal Talgam-Cohen (Eds.). Springer International Publishing, Cham, 3–21. https://doi.org/10.1007/978-3-030-94676-0_1

[15] Diodato Ferraioli and Carmine Ventre. 2018. Probabilistic Verification for Obviously Strategyproof Mechanisms. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 240–246. https://doi.org/10.24963/ijcai.2018/33

[16] Diodato Ferraioli and Carmine Ventre. 2021. Approximation Guarantee of OSP Mechanisms: The Case of Machine Scheduling and Facility Location. *Algorithmica* 83, 2 (Feb. 2021), 695–725. https://doi.org/10.1007/s00453-020-00771-x

[17] Diodato Ferraioli and Carmine Ventre. 2022. Obvious Strategyproofness, Bounded Rationality and Approximation. *Theory of Computing Systems* 66, 3 (June 2022), 696–720. https://doi.org/10.1007/s00224-022-10071-2

[18] Ali Hortaçsu and David McAdams. 2010. Mechanism Choice and Strategic Bidding in Divisible Good Auctions: An Empirical Analysis of the Turkish Treasury Auction Market. *Journal of Political Economy* 118, 5 (2010), 833–865. http://www.jstor.org/stable/10.1086/657948

[19] Jakub Kastl. 2011. Discrete bids and empirical inference in divisible good auctions. *The Review of Economic Studies* 78 (2011), 974–1014.

[20] Maria Kyropoulou and Carmine Ventre. 2019. Obviously Strategyproof Mechanisms without Money for Scheduling. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1574–1581.

[21] Ron Lavi and Chaitanya Swamy. 2009. Truthful mechanism design for multidimensional scheduling via cycle monotonicity. *Games and Economic Behavior* 67, 1 (Sept. 2009), 99–124. https://doi.org/10.1016/j.geb.2008.08.001

[22] Shengwu Li. 2017. Obviously Strategy-Proof Mechanisms. *American Economic Review* 107, 11 (Nov. 2017), 3257–3287. https://doi.org/10.1257/aer.20160425

[23] Josué Ortega and Erel Segal-Halevi. 2022. Obvious manipulations in cake-cutting. *Social Choice and Welfare* (June 2022). https://doi.org/10.1007/s00355-022-01416-4

[24] Parag A. Pathak and Tayfun Sönmez. 2008. Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism. *The American Economic Review* 98, 4 (2008), 1636–1652. http://www.jstor.org/stable/29730139

[25] Alexandros Psomas and Paritosh Verma. 2022. Fair and Efficient Allocations Without Obvious Manipulations. https://doi.org/10.48550/ARXIV.2206.11143

[26] Jean-Charles Rochet. 1987. A necessary and sufficient condition for rationalizability in a quasi-linear context. *Journal of Mathematical Economics* 16 (1987), 191–200. https://doi.org/10.1016/0304-4068(87)90007-3

[27] Alvin E. Roth. 1991. A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom. *The American Economic Review* 81, 3 (1991), 415–440. http://www.jstor.org/stable/2006511

[28] Peter Troyan and Thayer Morrill. 2020. Obvious manipulations. *Journal of Economic Theory* 185 (2020), 1–26. https://doi.org/10.1016/j.jet.2019.104970

[29] Carmine Ventre. 2014. Truthful optimization using mechanisms with verification. *Theoretical Computer Science* 518 (Jan. 2014), 64–79. https://doi.org/10.1016/j.tcs.2013.07.034