Justifying Mechanisms' Outcomes through Disclosure*

[Preliminary, incomplete draft — please do not distribute]

Clayton Thomas†

August 23, 2026

Abstract

How can a mechanism designer justify an allocation to participants by disclosing (as little as possible) information about their reports? For common auction formats, in order to unambiguously justify the outcome by disclosing bids, the designer must disclose essentially all of the bids. In contrast, the designer can justify the claim that no bidder could have gotten a *better* outcome by disclosing only the price-setting and winning bid(s). Turning to student-school matching mechanisms, Serial Dictatorship and Top Trading Cycles have intuitive justifications that disclose only that students prefer their matched outcome over other schools. In contrast, justifying Deferred Acceptance requires revealing information about what students would have preferred more than their matched school.

*We thank seminar participants at Yale and Columbia for helpful discussions. Parts of this work were completed while at Microsoft Research and Yale's CADMY.

†Rensselaer Polytechnic Institute — *E-mail*: thomas.clay95@gmail.com.

1 Introduction

Traditional economic theory provides apt theoretical justifications for why a mechanism designer should choose given outcomes. However, it ignores the question of how a designer should justify these outcomes to the mechanism's participants. For many real-world mechanisms, participants' (in)ability to see *why* the outcome was selected is a first-order concern. For example, matching mechanisms have received complaints that the process "rais[es] concerns about transparency" (Kasman & Valant, 2019) or even "invites accusations of [corruption]" (Robertson et al., 2021). Additionally, many high-profile online auctions have switched from a second-price format (highlighted as desirable by classical economic theory) to a first-price format (Google, 2019; Despotakis et al., 2021), a chose that may have been influenced by the fact that (only) in the first-price format, each bidder knows—solely from his bid and without any additional justification—the price he will pay if he wins.

The classic Second-Price Auction (2PA) illustrates how, according to some notions, justifying a mechanism's outcome can be challenging. Suppose the designer announces the outcome that bidder i won at price p, and she aims to justify the outcome by disclosing a small set of bids. If even a single bid b_j is not revealed, then it remains conceivable (for example) that $p < b_j < b_i$, and the correct price should have been b_j instead of p. Hence, we say that to justify the outcome of 2PA *unambiguously* by disclosing bids, the designer must disclose *all* bids.[1]

On the other hand, the outcome of 2PA can be justified, according to a different notion, much more easily. Suppose the designer announces that bidder i wins at price p, and discloses the highest and second-highest bids, b_i and b_j, with $b_i > p$ and $b_j = p$. Then, each bidder can be sure that no outcome consistent with b_i and b_j is *better* for him than his announced outcome.[2] Hence, we say that the outcome of 2PA is justified *uncontestably*—in that every bidder knows he got the best possible outcome consistent with the bids—when the designer discloses only *two* bids.[3]

In this paper, we study how the outcomes of mechanisms can be justified via ex post *disclosures* the designer makes about the reports of each agent. Our results for 2PA discussed above showcase key general features of our model. First, it accommodates multiple

[1] Grigoryan and Möller (2024) prove that, for a similar reason, 2PA satisfies a property they call *minimal auditability*; see Section 1.1. Our model shows how this non-auditability can be mitigated (for uncontestable justification) by disclosing the price-setting bid.

[2] Specifically, bidder i knows his price could not be lower than b_j, and each $k \neq i$ knows that he could not have won.

[3] Of course, if the designer can make more-flexible disclosures than simply revealing bids, then there is an additional, related approach that achieves unambiguous justification. In particular, if the designer discloses that j's bid was equal to p, that i's bid was *at least* p, and that if $k \notin \{i, j\}$ then k's bid was *at most* p, then the outcome is unambiguous.

justification types (e.g., unambiguous vs. uncontestable) that delineate, to different levels of precision, *why* the announced outcome was selected. Second, we consider a designer who wishes to disclose as little information as possible, and wishes to use specific disclosure types (e.g., fully-disclosing bids of only a few bidders).

Our first main application is to auctions. We formalize the above discussion for 2PA, and investigate our model for other auction formats, focusing on uncontestable justification and disclosures revealing (only) specific bids. For the First-Price Auction, justification can be accomplished by disclosing only a *single* bid: the winner's. For multi-unit auctions with L homogeneous goods, the VCG auction requires disclosing up to L winning bids and L price-setting bids. For multi-unit auctions heterogeneous goods, justification cannot generally be accomplished unless essentially *all* the bids are disclosed.

Our second main application is to matching. To illustrate how the designer can justify matchings, consider the classic Serial Dictatorship (SD) mechanism. In SD, each student i is considered in (priority) order $1, 2, \ldots, n$, and is matched to his highest-ranked yet-unmatched school $\mu(i)$. SD selects the outcome μ for a concrete reason: each student i preferred $\mu(i)$ over every other remaining school $h \in \{\mu(i+1), \ldots, \mu(n)\}$. This exactly corresponds to a (partial) disclosure about the report of each student i.[4] In fact, these are a particular type of disclosure we call *prefers-outcome* disclosures, which reveal that some student i prefers their outcome $\mu(i)$ over some set of other schools. Given a fixed priority order, these disclosures force the outcome of SD to be μ, hence they unambiguously justify μ. Thus, SD can be justified using (arguably intuitive) prefers-outcome disclosures that can be specified from only the announced outcome μ.[5]

We consider two classic priority-based matching mechanisms: the efficient Top Trading Cycles (TTC) and the stable (student proposing) Deferred Acceptance (DA). We assume that students know their priority scores, and consider what information about their reported rankings must be revealed in order to justify the announced matching μ. Our findings contribute to a long discourse comparing these two mechanisms, which (both in the literature and in practice) have largely favored DA, with one stated reason being its perceived interpretability. Indeed, when Boston adopted DA over TTC in 2005, their strategic planning

[4] If the designer considers disclose rules that (as in our results for single-item auctions) either fully-reveal of fully-conceal a student's report, then very little is possible without revealing (essentially) all of the reports. For example, disclosure rules must (even for uncontestable justification) justify to student n in SD why each of the seats $\{\mu(1), \mu(2), \ldots, \mu(n-1)\}$ was taken, which requires disclosing at least some information about each student in $\{1, 2, \ldots, n-1\}$.

[5] Möller (2026) proves that SD satisfies a property he calls *transparency*, for reasons related to the fact that justifying SD requires no announcements other than μ; see Section 1.1. Our results for matching show that, with mild additional announcements, more-intricate mechanisms such as TTC can be justified using disclosures very comparable to those of SD.

team wrote that "[In TTC,] the behind the scenes mechanized trading makes the student assignment process less transparent [than in DA]" (BPS, 2005).

Our main result for TTC highlights a potentially-overlooked advantage: TTC can be justified with disclosures quite similar to those justifying SD. In particular, for every profile of student reports, there exists an ordering $(h_1, h_2, \ldots, h_n)$ on the schools such that the TTC matching μ is unambiguously justified by the fact that all students i prefer $h_k = \mu(i)$ over every school in $\{h_{k+1}, \ldots, h_n\}$.[6] Hence, after announcing the TTC matching μ and specifying this ordering of schools, the designer can unambiguously justify μ using these prefers-outcome disclosures, strikingly similar to those used for SD.[7]

For DA, our findings are more nuanced. To begin, we show via a simple example—provided in Figure 1—that DA *cannot* be justified using prefers-outcome disclosures. In other words, unambiguously justifying the outcome μ of DA requires sometimes revealing that $A \succ_i \mu(i)$ for some student i. To get intuition for why this is the case, note that the information $A \succ_i \mu(i)$ can be the reason that some alternative matching μ' with $\mu(i) = \mu'(i)$ becomes unstable, and hence this information must be released to explain why DA did not output μ'.

Students:	Schools:
$1 : A \succ_1 B \succ_1 C$	$A : 2 \succ_A 3 \succ_A 1$
$2 : B \succ_2 A \succ_2 C$	$B : 1 \succ_B 3 \succ_B 2$
$3 : A \succ_3 C \succ_3 B$	$C : 3 \succ_C 2 \succ_C 1$
or : $C \succ'_3 A \succ'_3 B$	

Figure 1: Example showing that DA cannot be justified with prefers-outcome disclosures.

When each student i reports $\succ_i$, DA produces matching $\{(1, B), (2, A), (3, C)\}$. Alternatively, if student 3's report had been $\succ'_3$, DA would have produced $\{(1, A), (2, B), (3, C)\}$. Hence, when each i reports $\succ_i$, justifying the matches of students 1 and 2 requires the designer to disclose the fact that student 3 prefers A *over* her outcome C.

Given the above negative finding, we follow two alternative approaches to justifying DA. The first approach—analogous to our notion of uncontestable justification considered for auctions—is to justify a relaxed, less-specific property of the outcome. For the stable mechanism DA, it may be enough to justify the claim that the announced matching μ is stable. This can be done with prefers-outcome disclosures via a frequently-used representation of stable matchings (Azevedo & Leshno, 2016), namely, by disclosing that for each student i,

[6] In fact, this ordering $h_1, \ldots, h_n$ is exactly the order in which the schools were matched in the TTC algorithm. Given this ordering and the above disclosure, it is easy to see that the outcome of TTC is uniquely determined. Leshno and Lo (2021) also highlight the importance of the order in which schools are matched in the TTC algorithm, but their results focus on communication the outcome, not on justifying it; see Section 1.1.

[7] We remark that our model does not require this ordering to be verified by a trusted third-party in order to provide justification. Instead, our result shows that, if students all independently verify that the announced ordering and outcome matching is consistent with their own information (their priorities and reported ranking), then both the ordering and the matching *must be* as in TTC.

we have $\mu(i) \succ_i h$ for all schools h in the set $\{h \mid i \succ_h \mu(h)\}$.[8] Note, however, that since this will not unambiguously justify DA, the designer could use such disclosures to justify other matchings, such as the outcome of *school-proposing* DA.

Second, we consider how one can justify DA using more-flexible disclosures than in prefers-outcome disclosures, but while still revealing minimal total information about students' reports. We formalize the amount of information a disclosure protocol reveals as its communication cost. Building on recent results in theoretical computer science on the complexity of DA (Gonczarowski & Thomas, 2024), we present a protocols that require little more communication than just describing the outcome matching μ. Briefly, this protocol works by first justifying stability via the approach above, and additionally disclosing enough preferences $h \succ_i \mu(i)$ in order to guarantee that no Pareto improvement to μ can also be stable. Since the outcome μ of DA is the unique Pareto optimal stable matching, such a disclosure guarantees that μ is the outcome of DA.

We believe that the notions we defined elucidate the precise meaning and transparency guarantees of disclosing validated price-setting bids in auctions. They also highlight intuitive ways to justify the outcomes of the efficient SD and, perhaps-surprisingly, TTC outcomes, and while the situation becomes more-involved for DA, we provide justifications for that popular mechanism as well. Beyond our specific insights discussed above, a key goal in this work is to provide a flexible theoretical framework to reason about disclosure rules, so that these results and others beyond can be formalized in a cohesive language.

1.1 Related Work

Our paper contributes to the literature on simplicity and transparency in mechanism design. Many works in this area focus on incentives, and study the design of *interactive mechanisms*. These papers show that interaction can encourage straightforward preference reporting in strategyproof mechanisms (e.g., Li, 2017, Pycia and Troyan, 2023, Meisner and von Wangenheim, 2023, Bó and Hakimov, 2024, Nagel and Saitto, 2024, 2025, Dreyfuss et al., 2025),[9] and can incentivize the designer to follow the intended protocol (e.g., Akbarpour & Li, 2020; Komo et al., 2024). Another group of papers, also focused on the simplicity of incentives, study the design of ex ante *descriptions* of the mechanism (e.g., Katuščák and Kittsteiner, 2025, Gonczarowski et al., 2024, 2025).

[8] These disclosures, as for SD, can be determined from (only) the priorities and the announced outcome matching μ. (Möller, 2026) proves that, for related reasons, what he calls *safe deviations* from DA must be stable matchings; see Section 1.1.

[9] A complementary line of works study the (non-)simplicity of (deviating from) straightforward preference reporting in non-strategyproof mechanisms, e.g., Fernandez (2018), Börgers and Li (2019), Troyan and Morrill (2020), Hartman et al. (2025), and Shirakawa (2025).

In contrast to the works discussed above, we study the design of ex post *disclosures*. We abstract over the incentives of the designer, and study a fundamental notion of transparency: disclosing enough information to justify—to different levels of precision—the outcome announced by the designer.

We view our paper as conceptually most-related to prior works on auditability and transparency. In particular, Möller (2026) and Grigoryan and Möller (2024) present findings that (while not considered in the context of disclosure rules) capture related insights. Grigoryan and Möller (2024) define the auditability index of a mechanism as the number of agents required to (using their private information) falsify any incorrect outcome, and say that a mechanism is *minimally-auditable* if this requires all n agents. They find that 2PA (as well as DA) is minimally-auditable. Möller (2026) says that a *safe deviation* from a matching mechanism is one that no single agent could falsify, and says that a mechanism is *transparent* if it has no safe deviations. He finds that SD, but not TTC or DA, is transparent. Our work shows how, with moderate additional disclosures, these mechanisms can be made significantly more transparent.

We study static (non-interactive) mechanisms and ex post disclosures, and aim to minimize the amount of communication these disclosures require. Hakimov and Raghavan (2025) study how disclosure of priority cutoffs can make interactive matching mechanisms more transparent, and do not attempt to minimize communication. Our framework is agnostic as to the incentives of the designer, and (for unambiguous justification) requires than no incorrect outcome can be justified with the disclosures. Woodward (2020) considers disclosure rules for auctions garunteeing (only) that the auctioneer cannot profit from deviations.[10]

On a technical level, our work is closely related to the interdisciplinary literature on communication complexity in market design, such as Segal (2007), Gonczarowski et al. (2019), and Gonczarowski and Thomas (2024). Indeed, our notion of unambiguous justification is identical to the verification property required of nondeterministic communication protocols, is a frequently-considered theoretical model used to prove lower bounds on the communication requirements of (potentially interactive) deterministic protocols.[11] We study how such communication protocols can be used ex post to justify the outcomes of (static, non-interactive) mechanisms, and study novel, relaxed forms of justification and important special types of disclosures.

Privacy is an additional desideratum that—while not the focus of our paper—is an important consideration in the design and implementation of disclosure rules. Works studying

[10] Additionally, we require disclosures to separately reveal information about different agents' reports (e.g., disclosing "Agent 1 bid \$4 and agent 2 bid \$5"). In contrast, Woodward (2020) allows disclosures that reveal joint information about agents' reports (e.g., disclosing "the highest bid is \$5").

[11] For background on communication complexity, see Kushilevitz and Nisan (1997).

the privacy of interactive mechanisms include Brandt and Sandholm (2005, 2008), Milgrom and Segal (2020), and Haupt and Hitzig (2025). Further from our work is cryptographic (e.g., Izmalkov et al., 2011, Canetti et al., 2025), differential (e.g., Dwork et al., 2006, Pai and Roth, 2013), or stochastic (e.g., Eilat et al., 2025; Liu, 2025; He et al., 2026) notions of privacy.[12] Our results identify minimal-disclosures (in the sense of communication cost), serving as a natural starting point for the investigation of maximal-privacy disclosure rules achieving justification.

1.2 Paper Organization

Our paper proceeds as follows. In Section 2, we give preliminaries and define our general model of justification through disclosure. In Section 3, we study auctions, focusing on unambiguous justification through disclosing bids. In Section 4, we study matching, focusing on justification through prefers-outcome disclosures.

2 Model

In this section, we provide preliminaries, define our general framework for justification, discuss how we reason about disclosures, and present two simple examples.

2.1 Preliminaries on Environments and Mechanisms

An *environment* specifies a set of outcomes and a set of agents, where each agent ("he") has independent (i.e., private value) preferences over the outcomes. Formally, an environment $\left(N, \mathcal{A}, (\mathcal{T}_i)_{i\in N}\right)$ is defined as follows. There is a set $N = \{1, 2, \ldots, n\}$ of *agents* and a set of *outcomes* $\mathcal{A}$. Each agent $i \in N$ has a set $\mathcal{T}_i$ of *types*, and each $t_i \in \mathcal{T}_i$ corresponds to a *preference order* $\succeq_i^{t_i}$, which is a weak order over $\mathcal{A}$.[13] In *ordinal environments* we take these preference orders as a primitive. In *cardinal environments* each $t_i \in \mathcal{T}_i$ corresponds to a utility function $u_i^{t_i} = u^{t_i} : \mathcal{A} \to \mathbb{R}$, and we write $a \succeq_i^{t_i} b$ to mean $u_i^{t_i}(a) \geq u_i^{t_i}(b)$.

We write $t_N = (t_1, \ldots, t_n) \in \mathcal{T}_N = \mathcal{T}_1 \times \ldots \times \mathcal{T}_n$ to denote a type profile. For $I \subseteq N$, we let (t_I, t_{-I}) denote the type profile in $\mathcal{T}_N$ combining $t_I \in \mathcal{T}_I = \bigtimes_{i\in I} \mathcal{T}_i$ and $t_{-I} \in \mathcal{T}_{-I} = \bigtimes_{i\in N\setminus I} \mathcal{T}_i$.

[12] Liu (2025) shows that DA has the highest privacy loss (in terms of a novel worst-case privacy loss score) among all stable matching mechanisms. This complements our finding that justifying DA requires "privacy violating" (i.e., non-prefers-outcome) disclosures. In contrast to our result, Liu's holds because in DA, students' matches are more highly-correlated with their preferences, whereas (as illustrated by Figure 1) our result holds because students *might not* be matched to some school that they like.

[13] A weak order $\succeq$ is a total order with ties. Formally, it is a binary relation that is reflexive ($\forall a : a \preceq a$), transitive ($\forall a, b, c : a \preceq b \wedge b \preceq c \implies a \preceq c$), and connected ($\forall a, b : a \preceq b \vee b \preceq a$). As is standard, we write $a \succ b$ when $a \succeq b$ but we do not have $b \succeq a$.

When $I = \{i\}$, we write (t_i, t_{-i}) with $t_i \in \mathcal{T}_i$ and $t_{-i} \in \mathcal{T}_{-i}$.

The designer ("she") implements a (static, non-interactive) *mechanism* $f : \mathcal{R}_N \to \mathcal{A}$, which determines the selected outcome based on the *reports* r_i that each agent i chooses from a *report space* $\mathcal{R}_i$.[14] Most of the mechanisms we consider have $\mathcal{R}_i = \mathcal{T}_i$ for all $i \in N$. Many are furthermore *strategyproof* (i.e., $\forall i \in N, t_i, t_i' \in \mathcal{T}_i, t_{-i} \in \mathcal{T}_{-i} : f(t_i, t_{-i}) \succeq_i^{t_i} f(t_i', t_{-i})$). For such mechanisms, we assume i reports her true type t_i.

Throughout the paper, we assume that the designer's intended mechanism f is publicly known ex ante, and that the outcome a selected by the designer is publicly known ex post. For each agent i, the designer collects his report r_i via a private communication channel. In the status quo setting without disclosure, the designer then simply announces the selected outcome $a \in \mathcal{A}$.

2.2 Model of Disclosures and Justification

We consider a designer who wishes to select outcomes according to a mechanism $f : \mathcal{R}_N \to \mathcal{A}$ and justify her selection using disclosures. She collects reports $r_i \in \mathcal{R}_i$ from each agent $i \in N$. Then, she uses a protocol to announce (a, D), where $a \in \mathcal{A}$ is the selected outcome and D is a disclosure conveying information about agents' reports.

Formally, a *disclosure* D is a set $D \subseteq \mathcal{R}_N$ of the form $D = \bigtimes_{i \in N} D_i$ for some $D_i \subseteq \mathcal{R}_i$. An *announcement* is a pair (a, D), where $a \in \mathcal{A}$ and D is a disclosure. When $r_N \in D$, we say that the disclosure D is *consistent with* r_N, and that r_N is consistent with D. A (disclosure) *protocol* $\mathcal{P}$ is any set of announcements the designer might make. The *disclosure set* $\mathcal{D} = \{D \mid \exists (a, D) \in \mathcal{P}\}$ of protocol $\mathcal{P}$ is the set of disclosures $\mathcal{P}$ might make. When the designer selects an announcement (a, D) from the protocol, outcome a is implemented, and agents learn that the true report profile r_N is consistent with D. Hence, disclosures provide verifiably true partial information about the reports, which is publicly revealed ex post.

The designer's goal is to use a disclosure to justify the selected outcome to the agents. We say that D *unambiguously justifies* a (as the outcome of f) at r_N if D is consistent at r_N, and for all $\widetilde{r}_N \in D$, we have $f(\widetilde{r}_N) = a$. Protocol $\mathcal{P}$ *unambiguously justifies* the mechanism f if, for all $r_N \in \mathcal{R}_N$, there exists an $(a, D) \in \mathcal{P}$ such that D unambiguously justifies $a = f(r_N)$ at r_N, i.e.,

$$\forall\, r_N \in \mathcal{R}_N : \exists\, (a, D) \in \mathcal{P} : r_N \in D \wedge \big[\forall\, \widetilde{r}_N \in D : f(\widetilde{r}_N) = a\big]. \tag{*}$$

Additionally, when $\mathcal{P}$ is an unambiguous justification protocol, we require that for all $(a, D) \in \mathcal{P}$, there is some $r_N \in \mathcal{R}_N$ such that D unambiguously justifies a at r_N.[15]

[14] We write $r_N = (r_I, r_{-I})$ for $I \subseteq N$, and other related notation, using the same conventions as for agents' types.

[15] For unambiguous justification, this implies that for all $(a, D) \in \mathcal{P}$, we have $a = f(\widetilde{r}_N)$ for all $\widetilde{r}_N \in \mathcal{R}_N$.

When $\mathcal{P}$ unambiguously justifies f, the designer can always select announcements (a, D) that disclose enough information about the report profile to demonstrate—even to a third-party observer—that outcome of f is (unambiguously) a. Motivated by Equation (*), we will often describe a protocol $\mathcal{D}$ by specifying, for any report profile r_N, a generic disclosure $D \in \mathcal{D}$ that justifies $f(r_N)$ at r_N.

Justification types. The first key set of desiderata in our paper are justification types. Unambiguous justification is one type of justification, and is the strongest notion of justification we consider. We introduce additional, more-relaxed justification types throughout the paper.

Disclosure types. The second key set of desiderata in our paper are disclosure types. These are restrictions on the disclosures D used by a protocol. For instance, a *report revealing* disclosure D for agent i is one in which i's report is fully revealed, and no other information is given (i.e., $D = D_{\mathsf{rev}}(i, r_i) = \{r_i\} \times \mathcal{R}_{-i}$). We introduce additional, more-flexible disclosure types throughout the paper.

We study protocols $\mathcal{P}$ that use specific combinations of disclosure types. Formally, this means that each $D \in \mathcal{D}$ is the intersection of disclosures of the specified types. For example, we say that $\mathcal{D}$ uses at most k report-revealing disclosures when all $D \in \mathcal{D}$ are of the form $\{(r_u)_{u \in U}\} \times \mathcal{R}_{N \setminus U} = \bigcap_{u \in U} D_{\mathsf{rev}}(u, r_u)$ for some $U \subseteq N$ with $|U| \leq k$.

2.3 Quantifying and Minimizing Disclosure

Justification can always be achieved in a straightforward, generic way. In particular, every mechanism can be unambiguously justified by a *Full Disclosure* protocol $\mathcal{P}_{\mathsf{Full}}$ that completely specifies all agents' reports (i.e., the protocol $\mathcal{P} = \big\{\big(f(r_N), \{r_N\}\big) \mid r_N \in \mathcal{R}_N\big\}$). While this protocol indeed provides a reasonable justification of outcome a—after all, given r_N any party can simply calculate that $a = f(r_N)$—our paper goes beyond this trivial baseline, and characterizes what disclosures are required to achieve justification. In particular, we aim to minimize disclosure.

We primarily formalize the amount of information disclosed by a protocol $\mathcal{P}$ as its *communication cost* $\log_2 |\mathcal{P}|$. This measures the required length of messages that describe announcements $(a, D) \in \mathcal{P}$ (up to constant factors, regardless of the encoding language).[16] For

Hence, such protocols are uniquely determined from their disclosure sets $\mathcal{D}$. This will not be true for other notions, such as uncontestable justification, that we examine later in the paper.

[16] In particular, communicating an element of a finite set X requires $\lceil \log_2 |X| \rceil$ bits in the worst case, since a string of bits of length k can take exactly 2^k values. Throughout this paper, we consider asymptotic communication costs as the number of agents n goes to infinity, and use the standard notation $\mathcal{O}, \Omega, \Theta$ for

most mechanisms we consider, we achieve protocols with cost roughly that of the status quo: announcing the outcome (i.e., $\Theta(\log|\mathcal{A}|)$). This is typically much less than the cost Full Disclosure (i.e., $\log|\mathcal{R}_N|$).

We also introduce a partial order on disclosures and protocols based on what information they might reveal. For disclosures, we say D_1 *reveals less than* D_2 when $D_1 \supseteq D_2$. For protocols, we say that $\mathcal{P}_1$ *reveals less than* $\mathcal{P}_2$, written $\mathcal{P}_1 \lesssim \mathcal{P}_2$, when $\forall(a, D_1) \in \mathcal{P}_1 : \exists(a, D_2) \in \mathcal{P}_2 : D_1 \supseteq D_2$, and additionally $\forall(a, D_2) \in \mathcal{P}_2 : \exists(a, D_1) \in \mathcal{P}_1 : D_1 \supseteq D_2$. Hence, $\mathcal{P}_1 \lesssim \mathcal{P}_2$ means that in order to justify a given outcome, $\mathcal{P}_1$ can always reveal less than $\mathcal{P}_2$, and $\mathcal{P}_2$ can always reveal more than $\mathcal{P}_2$. For all settings we study, Full Disclosure is a maximum element under $\lesssim$. However, there are often multiple $\lesssim$-incomparable, $\lesssim$-minimal protocols.[17]

Finally, we say that a protocol $\mathcal{P}$ is *outcome-communicated* when there is exactly one announcement $(a, D) \in \mathcal{P}$ for each possible outcome $a \in \mathcal{A}$. This means that the protocol is particularly minimal, and (under mild conditions) implies that $\mathcal{P}$ has minimum communication cost and is $\lesssim$-minimal. However, outcome-communicated protocols can only justify a few simple mechanisms.

Before proceeding, we also remark on one of our foundational assumptions: that justifications must be made by specifying information about each individual agent's report separately (i.e., that all disclosures are of the form $\bigtimes_{i\in N} D_i$). If we instead allowed joint information about all agents reports (i.e., if we allowed disclosures to be arbitrary subsets of $\mathcal{R}_N$), then any mechanism could be unambiguously justified by the protocol containing an announcement $\big(a, \big\{\{r_N \in \mathcal{R}_N \mid f(r_N) = a\}\big\}\big)$ for each $a \in \mathcal{A}$. While such a protocol matches the communication of announcing the outcome, it provides no nontrivial justification for outcome a.[18]

We are agnostic as to whether the designer will ever actually *try* to deviate and select an incorrect outcome; we aim to provide important notions of justification even when the designer is always honest. However, we highlight one particularly important facet in the interpretation of our framework: that agents perceive the designer as being unable to "fabricate" reports (e.g., to submit shill bids in auctions); for instance, because they trust that other agents will validate that the disclosures are consistent with their reports.

---

an asymptotic relation that ignores constant factors, as well as $\widetilde{\mathcal{O}}, \widetilde{\Omega}, \widetilde{\Theta}$ for asymptotic relations that ignore logarithmic factors.

[17] In Section A.2, we provide an explicit example of two incomparable $\lesssim$-minimal protocols. We also provide additional exposition into the properties of $\lesssim$; this ordering is related to the refinement order on partitions of $\mathcal{R}_N$, with the main difference is that, unlike for partitions, disclosures need not be disjoint.

[18] Our assumption that disclosures are of the form $\bigtimes_{i\in N} D_i$ also implies that agents can always validate that the true reports r_N are consistent with D by separately checking that $r_i \in D_i$, without resorting sophisticated coordination, lending additional transparency to our protocols.

2.4 Introductory Examples and Discussion

We now provide two concrete examples that illustrate the primitives of our model and our notion of (unambiguous) justification.

Two-Trader Mechanism. For our first example, consider an environment with two agents $N = \{1, 2\}$ and two outcomes $\mathcal{A} = \{a_0, a_T\}$, where a_0 represents a default outcome, and a_T represents the two agents engaging in trade (at an exogenously fixed price), to which they both must concent. Each agent i has two types $\mathcal{T}_i = \{\succ_i^0, \succ_i^T\}$, where $a_0 \succ_i^0 a_T$, and $a_T \succ_i^T a_0$. The *Two-Trader* mechanism $\mathsf{2\text{-}Trade} : \mathcal{T}_N \to \mathcal{A}$ selects outcome a_T when both agents i report $\succ_i^T$, and selects a_0 otherwise.[19]

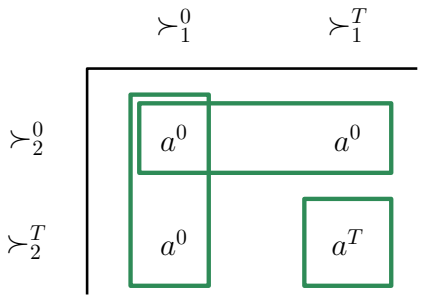
Protocol 1, illustrated in Figure 2, provides a disclosure protocol for 2-Trade. Protocol 1 either announces the outcome was a^0 ("no trade") and justifies this by disclosing a specific agent i who did not want to trade, or it announces a^T ("trade") and discloses that both agents wanted to trade. This unambiguously justifies 2-Trade since, for each possible disclosure, there is a unique outcome consistent with the disclosure. Moreover, this protocol reveals the minimum information of any unambiguous justification protocol for 2-Trade.[20]

Protocol 1: Unambiguously Justifying 2-Trade

The protocol contains three announcements:

- Selecting a^0 and disclosing that agent 1 reported $\succ_1^0$, i.e., $\big(a^0, \{(\succ_1^0, \succ_2^0), (\succ_1^0, \succ_2^T)\}\big)$,
- Selecting a^0 and disclosing that agent 2 reported $\succ_2^0$, i.e., $\big(a^0, \{(\succ_1^0, \succ_2^0), (\succ_1^T, \succ_2^0)\}\big)$, and
- Selecting a^T and disclosing that both agents $i \in N$ reported $\succ_i^T$, i.e., $\big(a^T, \{(\succ_1^T, \succ_2^T)\}\big)$.

Figure 2: Illustration of Protocol 1



Protocol 1 illustrates how concealing information can provide more-general justifications of outcomes, compared to Full Disclosure. For example, at reports $(\succ_1^0, \succ_2^0)$, the protocol can justify the outcome by disclosing that agent 1 did not want to trade, where Full Disclosure would specify that *both* agents do not want to trade. This example also shows that there can be multiple distinct announcements (a, D) that are consistent with a given report profile

[19] Up to relabeling, this mechanism is also coincides with a two-agent voting rules that requires unanimous consent to adopt policy a_T. It also coincides with a special case of both the Top Trading Cycles and Deferred Acceptance matching mechanisms, as follows. Agents are matched to schools X, Y, whose priorities over agents are $1 \succ_X 2$ and $2 \succ_X 1$. Outcome a_0 corresponds to the matching $\mu_0 = \{(1, X), (2, Y)\}$, and a_T corresponds to $\mu_T = \{(1, Y), (2, X)\}$. Agent i's preference $\succ_i^0$ in the Trade mechanism corresponds to a preference for i's match in μ_0, and vice-versa for $\succ_i^T$.

[20] Formally, for all protocols $\mathcal{P}_2$ unambiguously justifying 2-Trade, we have (Protocol 1) $\lesssim \mathcal{P}_2$.

r_N, and the protocol does not specify which the designer will choose. This fact can allow the designer to reveal less information than in protocols in which all disclosures must be disjoint.[21]

Majority Voting. For our second example, consider the Majority Rule mechanism Maj in an environment with voters $N = \{1, \ldots, n\}$, candidates $\mathcal{A} = \{a, b\}$, and binary types (preferring either a or b). For simplicity, assume n is odd.

Protocol 2 presents a disclosure protocol for Maj, which announces the candidate $x \in \{a, b\}$ that won the election, and discloses some majority of exactly $(n+1)/2$ voters who voted for x. As in our first example above, this protocol unambiguously justifies Maj, and reveals the minimum information required to do so.[22]

Protocol 2: Unambiguously Justifying Maj

Outcome: Candidate $x \in \{a, b\}$ elected.
Disclosure: For some $S \subseteq N$ with $|S| = (n+1)/2$, disclose that all $i \in S$ voted for x.

Protocol 2 illustrates how (unambiguous) justification requires specifying *why* an outcome is selected. Candidate $x \in \{a, b\}$ wins by receiving support from a majority of voters, so the protocol must reveal such a majority to justify the claim that x wins. Moreover, in both of Protocols 1 and 2, each announcement (a, D) has an intuitive verbal description that explains why the outcome is a for all reports consistent with D. While our formal model abstracts over precisely how justifications are verbally conveyed, this feature is true—to various extents—of many protocols we construct.

While Protocol 2 announces the minimal disclosures that justify Maj, it requires substantially more communication than the status quo approach of announcing (only) the winning candidate.[23] In particular, Protocol 2 has communication cost $\widetilde{\Theta}(n)$ bits.[24] In contrast, for the allocation problems we consider in our main results, we will achieve justifications that require only a constant-factor increase in communication over announcing the outcome.

[21] For example, if all the disclosures in Figure 2 were required to be disjoint, then the protocol would necessarily fully reveal $\{r_N\}$ for at least one of $r_N = (\succ_1^0, \succ_2^T)$ or $r_N = (\succ_1^T, \succ_2^0)$, disclosing more information than necessary to justify a^0 as the correct outcome.

[22] To see that Protocol 2 is $\lesssim$-minimal, note that for each $x \in \{a, b\}$ and $S \subseteq N$ of size $(n+1)/2$, an unambiguous justification protocol must sometimes announce (x, D) for some disclosure D that (at least) reveals that S voted for x, and additionally must always announce some disclosure of this form.

[23] Here we show that justifying Majority Rule requires disclosing (just over) half the votes. In Section A.3, we show that justifying Median Rule in a single-peaked voting environment can require even more, namely, completely revealing every single vote.

[24] Intuitively, this means that specifying a subset of half the voters requires a string of length proportional to n. To prove this formally, note that there is a distinct announcement in Protocol 2 for each of possible $2\binom{n}{(n+1)/2}$ pairs (x, S), where $x \in \{a, b\}$ and S is a set of $(n+1)/2$ voters, and apply Stirling's approximation.

3 Auctions

In this section, we study single- and multi-item auctions. We focus on *bid revealing* disclosures. We show that a relaxed notion called *uncontestable* justification can often be achieved using only a modest number of bid-revealing disclosures.

Preliminaries on auctions. In an auction environment, the agents $i \in N = \{1, \ldots, n\}$ are known as *bidders*. Each $i \in N$ will receive a *bundle* $B_i \in \mathcal{B}_i$, and his type space $\mathcal{T}_i$ is a set of *valuations* $v_i : \mathcal{B}_i \to \mathbb{R}_{\geq 0}$. There is a set of *feasible allocations* $\mathcal{F} \subseteq \bigtimes_{i \in N} \mathcal{B}_i$, and each outcome $(B_N, p_N) \in \mathcal{A}$ consists of a feasible allocation $B_N \in \mathcal{F}$ and prices $p_N = (p_i)_{i \in N}$ charged to the agents. When the select outcome is (B_N, p_N), agent i's utility is $v_i(B_i) - p_i$. For concreteness, we ignore the possibility of ties, and assume that communicating prices requires $\Theta(\log n)$ bits.[25] Throughout, we focus on (strategyproof or non-strategyproof) static, non-interactive auctions in which the report space $\mathcal{R}_N$ equals the type space $\mathcal{T}_i$, and refer to the reports $v_N \in \mathcal{R}_N$ as *bids*.

In auction environments, agents always prefer outcomes where they are charged less money but receive the same bundle of goods. We introduce terminology to capture this fact, and any more-general situation where all possible types of agent i agree on how some outcomes compare. For $a, b \in \mathcal{A}$, we write $a \succeq_i^* b$ when all types of i prefer a over b (i.e., $\forall t_i \in \mathcal{T}_i : a \succeq_i^{t_i} b$); we then say that i *always prefers* a to b.

Bid Disclosures and Uncontestable Justification. For auctions with report space $\mathcal{R}_N = \mathcal{T}_N$, a particularly salient form of disclosure is revealing bids.[26] A *bid revealing* disclosure is one where—for some agent i, value b, and bundle $B_i \in \mathcal{B}_i$—the disclosure reveals that i reported v_i such that $v_i(B_i) = b$. Formally, it is thus a disclosure D of the form

$$D = \{v_i \in \mathcal{T}_i \mid v_i(B_i) = b\} \times \mathcal{T}_{-i}.$$

[25] We formalize these technicalities as follows. For each bidder $i \in N$, we assume all valuations $v_i \in \mathcal{T}_i$ have support $\mathcal{V}_i \subseteq \mathbb{R}_{\geq 0}$, where $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$ for all $i \neq j$. We also assume $\log_2(|\mathcal{V}_i|) = C \log n$ for some constant C. For example, we can take $\mathcal{V}_i = \{v + 2^{-i} \mid v \in \{0, 1, \ldots, C^n\}\}$. Our results remain unchanged if type spaces allow for ties, but auctions always break ties in some consistent way (e.g., in favor of the lowest-index agent).

[26] Depending on the context, there may be many reasons to prefer bid-revealing disclosures over arbitrary disclosures. For one, their meaning may be more concrete and interpretable, which may be helpful in explaining auction outcomes. For another, if cryptographic signatures are needed in order to prevent shill bidding, then validating the authenticity of a bid using digital signature can be easily done using off-the-shelf and widely-available cryptography, while in contrast, it may be computationally far more challenging to implementing a more-complicated cryptographic protocol that validates a claim such as $b_i \leq 100$ (without revealing more information about b_i).

As we will formalize momentarily, unambiguous justification is too strong a requirement to achieve (nontrivial) bid-revealing disclosure protocols. Hence, we introduce a relaxed notion based on the idea that no agent should have a reason to "complain" about the announced outcome. In other words, the disclosure should make it clear, for each agent i, that no outcome consistent with the disclosure could've been better for i than the announced one.

Formally, given a mechanism f, we say that disclosure D *uncontestably justifies* a at r_N if D is consistent at r_N, and for all agents i and $\widetilde{r}_{-i} \in D_{-i}$, we have $a \succeq_i^* f(r_i, \widetilde{r}_{-i})$. Protocol $\mathcal{P}$ *uncontestably justifies* f if, for all $r_N \in \mathcal{R}_N$, there exists an $(a, D) \in \mathcal{P}$ such that D uncontestably justifies $a = f(r_N)$ at r_N, i.e.,

$$\forall\, r_N \in \mathcal{R}_N : \exists \left(f(r_N), D\right) \in \mathcal{P} : r_N \in D \wedge \left[\forall\, i \in N, \widetilde{r}_{-i} \in D_{-i} : f(r_N) \succeq_i^* f(r_i, \widetilde{r}_{-i})\right]. \quad (\dagger)$$

Additionally, when $\mathcal{P}$ is an uncontestable justification protocol, we require that for all $(a, D) \in \mathcal{P}$, there is some $r_N \in \mathcal{R}_N$ with $f(r_N) = a$ such that D uncontestably justifies a at r_N.

When $\mathcal{P}$ uncontestably justifies f, the designer can always select announcements (a, D) that is consistent with the reports r_N and also reveals enough information on reports to show to each agent i that—given the information i has—no outcome i would have preferred to a would be possible. Hence, under this announcement, no agent i has any reason to "contest" the outcome a, and we say that a is indeed uncontestable.

3.1 Single-item Auctions

In a single-item *auction environment*, agent i's outcome $a_i \in \mathcal{A}_i$ designates either that i wins the item and paying a price p, or that i loses; outcomes $a \in \mathcal{A}$ specify at most winner (and his price). We consider two canonical auctions; both of them have report space $\mathcal{T}_i$, which is referred to as the space of *bids*. In the strategyproof Second-Price Auction (2PA), the agent with the highest bid wins and pays a price equal to the second-highest bid. In the non-strategyproof First-Price Auction (1PA), the agent with the highest bid wins and pays his own bid as his price. In both auctions, only the winner pays, and we denote the outcome as (i, p) where $i \in N$ is the winner and $p \in \mathbb{R}$ is the price he pays.

We start by describe two disclosure protocols justifying 2PA in our framework: Protocols 3 and 4. Protocol 3 uses 2 bid-revealing disclosures, in particular, it reveals the highest-bid b_i and the second-highest bid b_j, where $b_j = p$. To see that this uncontestably justifies outcome (i, p), observe that given only this disclosed information, (1) i knows that, if he wins the auction, his price will be at least p, and (2) all bidders $k \neq i$ know that they

cannot win the auction while paying a price less than b_i.[27]

Protocol 3: Uncontestably justifying 2PA with 2 bid-revealing disclosures
Outcome: Bidder i wins and pays price p.
Disclose that i bid $b_i = q$, where $q > p$.
Disclose, for some $j \neq i$, that j bid $b_j = p$.

Now, the closely-related Protocol 4 also discloses the price-setting bid(er) $b_j = p$, but it differs from Protocol 3 in that it uses more-flexible disclosures, other than just revealing bids. It reveals that $b_k \leq p$ for all non-winning, non-price-setting bidders $k \notin \{i, j\}$. (And, it makes the more-vague disclosure that for the winning bidder i, $b_i \geq p$.) This, of course, uniquely determines the outcome of 2PA, i.e., it unambiguously justifies 2PA.

Protocol 4: Unambiguously justifying 2PA
Outcome: Bidder i wins and pays price p.
Disclose that i bid some value $b_i \geq p$. for some $j \neq i$.
Disclose, for some $j \neq i$, that j bid $b_j = p$.
Disclose, for all $k \notin \{i, j\}$, that k bid some value $b_k \leq p$.

Both these protocols intuitively amount to revealing the price-setting bid.[28] However, we meaningfully distinguish them by what information they contain. This distinction could be made literal, e.g., via some technology such as cryptography that the designer must use to validate the disclosures. Alternatively, the distinction could be conceptual, in that, e.g., real-world participants may ignore vague disclosures such as "$b_k \leq p$," and interpret Protocol 4 as the same as Protocol 3. Hence, it becomes important to delineate what formal notions of justification these protocols obtain; this is precicely what our framework provides.

As discussed in the introduction, unambiguously justifying 2PA by disclosing bids (i.e., combining the simple disclosure types of Protocol 3 with the strong justification type of Protocol 4) requires disclosing *all* the bids. Indeed, this is true even in 1PA. To see this formally, suppose the designer announces $\big((i, p), D\big)$ where D leaves even a single bid undisclosed. Then, it could be the case that the undisclosed bidder placed bid greater than p, and hence should have influence the outcome of the auction, either setting the price (which is possible 2PA) or becoming the winner (which is possible in both 1PA and 2PA). Formally:

[27] Note that Protocol 3 be used to uncontestably justifies an outcome a at b_N, where a is *not* the correct outcome $\text{2PA}(b_N)$. For instance, if $N = \{1, 2, 3\}$, at reports $b_1 = 3, b_2 = 1, b_3 = 5$, the outcome where bidder 3 wins and pays 1 is uncontestably justified by disclosing that b_2 was 1 (and that $b_3 = 5$). This is exactly the sense in which this (uncontestable) protocol is "ambiguous". We provide additional discussion in Section A.4.

[28] Both have communication cost $O(\log n)$, matching (up to constant factors) the cost of communicating a single bid or the identity of a single bidder.

Proposition 3.1. *For both 2PA and 1PA, any bid-revealing, unambiguous disclosure protocol must reveal all n bids.*

On the other hand, applying our notion of uncontestability, it is not hard to see that Protocol 5 justifies 1PA. This particularly-simple protocol discloses one less bid than Protocol 3 for 2PA, and is even outcome-communicated. Hence, while even the transparent mechanism 1PA requires some disclosure in our framework, we corroborate the findings indicating 1PA is more transparent than 2PA. But we also delineate the discussion further, showing that uncontestability is needed (under the restriction to revealing bids) and that 2PA requires disclosing only a single additional bid more than 1PA.

Protocol 5: Uncontestably justifying 1PA with 1 bid-revealing disclosure

Outcome: Bidder i wins and pays price p.
Disclose that i bid $b_i = p$.

3.2 Multi-item Auctions

We now briefly investigate our framework for two classic multi-item auction settings: multi-unit auctions with homogeneous goods, and unit-demand auctions with heterogeneous goods, which we introduce below. For simplicity, we focus on our main desiderata of interest for auctions: bid-revealing disclosures and uncontestable justification.

Homogeneous goods. In a homogeneous goods environment, as in e.g. Ausubel (2004), there a number of available units L; bundles $B_i \in \{0, 1, \ldots, L\}$ indicate that i received B_i units; the feasible allocations are those B_N with $\sum_{i \in N} B_i \leq L$; and types are functions $v_i : \{0, 1, \ldots, L\} \to \mathbb{R}$ with diminishing marginals. We consider the VCG auction, and denote this mechanism in the homogeneous goods environment by HG.

To begin to analyze the justification of HG, we observe that at *some* bid profiles, the outcome can be justified with very few bid disclosures indeed. For instance, suppose there are two bidders, a high bidder h and a low bidder ℓ, where $v_h(j) = 100 \cdot j$ for all j and $v_\ell(j) = 10 \cdot j$ for all j. Then h is allocated all L units and pays $10 \cdot L$. If we disclose that: $v_h(L) = 100 \cdot L$, $v_h(L-1) = 100 \cdot (L-1)$, and $v_\ell(L) = 10 \cdot L$, then in fact h knows he could not have paid less than $10 \cdot L$, and ℓ knows he could not have been awarded any items (since the diminishing marginals assumption means that v_h is the only bid consistent with this disclosure). Hence, disclosing 3 bids suffices for uncontestable justification, regardless of the number of units L.[29]

[29] Additionally, the above example is not just an artifact of the fact that there are only two bidders.

On the other hand, suppose that h has valuation as above and is still allocated all the units, but his price is determined from L separate bidders (e.g., since there are L separate other bidders k with $v_k(j) = 50$ for all $j \geq 1$). Then at least these L other bids $v_k(1)$ (or some information about the bid of k) must be disclosed in order to uncontestably justify h's price (which in this case is $50 \cdot L$). Since (as above) we must disclose at least two of h's bids to justify to these other bidders k that they did not receive any units, this example shows we must sometimes disclose $L + 2$ bids.

We now show that HG can, generalizing the single-item auctions above, be uncontestably justified using only a small number of bid-revealing disclosures.

Proposition 3.2. *HG can be unambiguously justified by revealing at most $2L$ bids.*

Proof. To present this proof, define the "marginal bids" of agent i for unit $j > 0$ as $m_i(j) = v_i(j) - v_i(j-1)$. At each bid profile in VCG, the marginal bids fall into three categories: "winning" marginal bids (those i, j where in the outcome of HG, $B_i \geq j$), "price-setting" bids (those i, j such that, if we removed some other agent k, then that marginal unit would become winning), and other "losing" marginal bids. By definition there are at most L winning bids. There are also at most L price-setting bids, since at most one such bid can arise for each winning marginal bid. By revealing all the winning and the price-setting bids (which are always, for each bidder i, a set of units of the form $0, 1, 2, \ldots, k$ for some $k \geq 0$), one can infer all of the corresponding marginal bids, and these suffice to (1) ensure each bidder that he could not have received more units and (2) ensure each bidder that he could not have paid a lower price. Hence, $2L$ prices suffice to uncontestably justify HG. □

Unit Demand. In a homogeneous goods environment, as in e.g. Demange et al. (1986), there is a set of goods M; bundle set $\mathcal{B}_i = M \cup \{\emptyset\}$ indicate what good i (or that i receives nothing); the feasible allocations are matchings assigning each bidder to (at most) one item; and types are functions $v_i : M \to \mathbb{R}$. We consider the VCG auction, and denote this mechanism in the unit demand environment by UD.

In this environment, we prove that (even uncontestable) justification is hard to achieve by disclosing bids:

Proposition 3.3. *Any disclosure protocol that justifies UD (either unambiguously or uncontestably) using bid-revealing disclosures must sometimes reveal $\Omega(n^2)$ bids.*

Proof. Consider an auction with n bidders $\{1, \ldots, n\}$ and $n-1$ items $\{x_1, \ldots, x_{n-1}\}$. Define a report profile as follows. For each bidder $i \in \{1, \ldots, n-1\}$, let i bid 10 on item x_i, and

The example would still apply if there are more than two bidders, so long as their bids are even lower, e.g. $v_k(j) = 3 \cdot j$.

0 otherwise. Then, let bidder n bid 9 on *every* item. On this bid profile, UD allocates each $i \in \{1, \ldots, n-1\}$ to x_i (and they each pay 9), and bidder n goes unassigned. Whatever disclosure D is announced by the designer must uncontestably justify this outcome to bidder n, i.e., to must reveal enough information for bidder n to know that he could not have been assigned (at a price less than 9).

However, consider the following modification to the bids. For any $i, j \in \{1, \ldots, n-1\}$, suppose that i bid 50 on x_j, instead of 0, and suppose that no other bid changes. Then, UD allocates i to x_j (at price 10). Furthermore, bidder j goes unassigned, and bidder n is assigned to item x_i (at price 0). Hence, the only way for a bid-revealing disclosure D to uncontestably justify the correct outcome (under the original report profile) is to disclose all such bids $v_i(x_j)$, of which there are $(n-1)^2 = \Omega(n^2)$. □

4 Matching

In this section, we study priority-based matching. We focus on *prefers-outcome* disclosures. We show that a relaxed notion called *uncontestable* justification can often be achieved using only a modest number of bid-revealing disclosures.

Preliminaries on Matching. In a (one-sided) matching environment, agents $i \in N = \{1, \ldots, n\}$ are *students* matched to (high) schools $h \in H$. Student i's type is a preference ranking (i.e., a strict total order) $\succ_i$ order over a subset of schools. The subset of schools h ranked by $\succ_i$ are *acceptable* to i (written $h \succ_i \emptyset$); unranked schools are *unacceptable* (written $\emptyset \succ_i h$). Each outcome $\mu \in \mathcal{A}$ is a *matching* $\mu : N \to H \cup \{\emptyset\}$ assigning at most one student per school. Students may go unmatched (when $\mu(i) = \emptyset$). By abuse of notation, for $h \in H$ we write $\mu(h) = i$ when $\mu(i) = h$, and $\mu(h) = \emptyset$ if no such $i \in N$ exists. When student i has type $\succ_i$, he prefers μ to ν when $\mu(i) \succ_i \nu(i)$.

In matching, and in mechanism design more-generally, we are interested in certain desirable *outcome properties*. We model these as functions $P : \mathcal{T}_N \to 2^{\mathcal{A}}$, where $P(t_N)$ is the set of outcomes satisfying the property at types t_N.[30] For an example in a general environment, Pareto Optimality is the set of outcomes that are not Pareto dominated at t_N, i.e.,

$$P(t_N) = \left\{ a \in \mathcal{A} \;\middle|\; \text{there is no } a' \in \mathcal{A} \text{ such that } (\forall i \in N : a' \succeq_i^{t_i} a) \text{ and } (\exists i \in N : a' \succ_i^{t_i} a) \right\}.$$

[30] When $P(t_N) \neq \emptyset$ for all $t_N \in \mathcal{T}_N$, outcome properties are often referred to as social choice correspondences. We adopt our terminology to emphasize that in our model, the designer will always announce a mechanism that specifies a single outcome $a = f(r_N)$, but may justify only the claim that a satisfies property P, i.e., $a \in P(r_N)$. Note that many properties of mechanisms, such as strategyproofness, are not *outcome* properties.

Prefers-Outcome Disclosures and Justifying Properties. Since matching mechanisms lack the cardinal notion of the value of a bid, and instead elicit students' ranked-order preferences, we consider a different notion of an intuitive disclosure type (compared to our bid-revealing notion in Section 3). We say that a disclosure D is a *prefers-outcome* disclosure at outcome μ if, for some student i and a subset of schools $R \subseteq H$, it reveals that i prefers $\mu(i)$ over all other schools in R.[31] That is, the announcement (μ, D) must be such that

$$D = \{\succ_i \in \mathcal{T}_i \mid \mu(i) \succeq_i r \text{ for all } r \in R\} \times \mathcal{T}_{-i}.$$

We also analyze a weaker notion of justification: (unambiguously) justifying properties such as Pareto Optimality and Stability. This definition is perfectly analogous to unambiguous justification, but shows that the announced outcome has the property. Formally, given an outcome property P, reported types $t_N \in \mathcal{T}_N$ and an outcome $a \in \mathcal{A}$, we say that disclosure D *justifies* a as having the property P at t_N if D is consistent at t_N, and for all $\widetilde{r}_N \in D$, we have $a \in P(\widetilde{r}_N)$. Given a mechanism f, protocol $\mathcal{P}$ *justifies* that f has property P if, for all $t_N \in \mathcal{T}_N$, there exists an $(f(t_N), D) \in \mathcal{P}$ such that D justifies $f(t_N)$ as having property P at t_N, i.e.,

$$\forall\, t_N \in \mathcal{T}_N : \exists \left(f(t_N), D\right) \in \mathcal{P} : t_N \in D \wedge \left[\forall\, \widetilde{t}_N \in D : f(t_N) \in P(\widetilde{t}_N)\right]. \tag{§}$$

Additionally, when $\mathcal{P}$ is an protocol justifying a property P of mechanism f, we require that for all $(a, D) \in \mathcal{P}$, there is some $t_N \in \mathcal{T}_N$ with $f(t_N) = a$ where D justifies a as having property P at t_N.

4.1 Efficient Matching

We consider two canonical Pareto efficient matching mechanisms. First, Serial Dictatorship (SD) is defined be a single priority order $\succ$ over student $i \in N$. The matching can be calculated by, sequentially in the order specified by $\succ$, permanently matching each student to his highest-ranked not-yet-matched institution. For concreteness, throughout our entire paper we assume that priorities are fixed and publicly known (e.g., since they are determined by pre-set school policies based on publicly-available information such as students' address).[32]

[31] This notion restricts disclosures in a way that depends on the selected outcome a; hence, it is a restriction on the announcements (a, D) of a protocol and not only on the disclosures D the protocol uses (as in the case of bid-revealing disclosures).

[32] This assumption can also be relaxed, for each result we present, to the communication model where each student starts out knowing their *priority score* at each school; one simply needs to communicate the relevant information about the scores (that would induce the needed school preferences) in each particular protocol, and students can still validate this information given their knowledge of their own scores. Of course, this

Second, Top Trading Cycle (TTC) is defined by a profile of priority orders $\{\succ_h\}_{h\in\mathcal{I}}$. The matching can be calculated as follows: Every not-yet-matched student "points" to his favorite not-yet-matched institution, and every remaining likewise institution points to its highest-priority remaining student. Choose any cycle in this pointing graph and "eliminate" it by permanently matching the students in the cycle to the institution they are pointing to. Remove eliminated students and institutions, and repeat this process until all students' matches are determined.

To begin, we observe that the rather simple and intuitive Protocol 6 unambiguously justifies SD. This protocol is even outcome-communicated, since students can observe the sets $\{\mu(i), \mu(i+1), \ldots, \mu(n)\}$ given only (the publicly-known priorities and) the matching μ.

Protocol 6: Unambiguously justifying SD with prefers-outcome disclosures

Outcome: The SD matching μ with priority order $1, 2, \ldots, n$ on the students.
Disclose that for all i, student i prefers his match $\mu(i)$ most out of $\{\mu(i), \mu(i+1), \mu(i+2), \ldots, \mu(n)\}$.

Next, we present our unambiguous justification protocol for TTC. This is Protocol 7.

Protocol 7: Unambiguously justifying TTC with prefers-outcome disclosures

Outcome: TTC matching μ.
Announce the ordering $h_1, h_2, h_3 \ldots, h_n$ in which the schools were matched in the TTC algorithm.
Disclose that for all i, agent i prefers his match $\mu(i) = h_k$ most out of $\{h_k, h_{k+1} \ldots, h_n\}$.

It is clear from the disclosure made by the protocol, and the definition of the TTC algorithm, that if the schools indeed match in order $h_1, h_2, h_3 \ldots, h_n$, then the outcome μ must be correct. However, there is a subtlety here: to show that this protocol is truly unambiguous, we must show that no incorrect ordering and incorrect matching can be used, and still be consistent with the reports $\succ_N$. We formalize this next.

Proposition 4.1. *Protocol 7 unambiguously justifies TTC.*

Proof. Given the announced outcome μ, all students can infer from the ordering over schools that the pointing relations when students matched (ignoring, for now, the round of the TTC algorithm when they matched) were:

$$\mu(h_1) \to h_1 \qquad \mu(h_2) \to h_2 \qquad \mu(h_3) \to h_3 \qquad \ldots \qquad \mu(h_n) \to h_n.$$

can means that additional communication (e.g., of others' priorities in Protocol 6) is needed, but for each result it is always at most $\widetilde{O}(n)$ bits with n students and schools.

The main thing is to now observe that, given this information, anyone can infer what student each school will point to at each step: they will point to their highest-priority remaining student, where 'remaining' means that the student was not in an earlier cycle. Since such an ordering always exists by the TTC algorithm, and since the disclosure on student preferences then mean this execution of the algorithm is forced, this implies the matching in unambiguously justified. □

We also note that, in contrast to SD, TTC has no outcome-communicated justification protocol. To see this, suppose student i is top priority at school h_i for $i \in \{1, 2\}$ (corresponding to the Two Trader Mechanism as in Figure 2). Indeed, when 1 and 2 both prefers school 2, or when 1 and 2 both prefer school 1, the outcome is the same—the higher-priority student gets the favorite school. However, the disclosures must be different (as Figure 2 illustrates).

4.2 Stable Matching

We consider the classic Deferred Acceptance (DA) stable matching mechanism. Like TTC, DA is defined by a profile of priority orders $\{\succ_h\}_{h\in\mathcal{I}}$. The matching is produces as follows: A currently-unmatched student i is chosen to "propose" to his favorite institution h which has not yet "rejected" him. Institution h rejects the proposal from i if h is currently matched to an student j with $j \succ_h i$. Otherwise h accepts the proposal, and if currently matched to a j with $i \succ_h j$, then h rejects j. Rejected applicants become (currently) unmatched. Repeat this process until no more students are unmatched and the final matching is determined.

As proven by Figure 1 in the introduction, we cannot justify DA using as-simple of disclosures as for SD and TTC. Formally:

Proposition 4.2. *It is impossible to unambiguously justify the DA outcome with prefers-outcome disclosures.*

However, there are two approaches to circumvent this. One approach, is presented in Protocol 8, is to justify only stability (of DA, or of any stable mechanism). This approach uses only prefers-outcome disclosures, adapting the well-known ideas in Azevedo and Leshno (2016), and also follows easily from the definition of stability.

Protocol 8: Justifying stability using prefers-outcome disclosures

Outcome: A stable matching μ.
Disclose that for all i, student i prefers his match $\mu(i)$ most out of $\{h \mid i \succeq_h \mu(h)\}$.

A second approach, presented in Protocol 9,[33] is to unambiguously justify the outcome of DA using more-flexible disclosures, but still with low communication overall.

Protocol 9: Unambiguously justifying DA

Outcome: DA matching μ
Disclose enough to justify the stability of μ, exactly as in Protocol 8.
For all schools h, let $d = d(h)$ be the student with $h \succ_d \mu(d)$
who is highest-ranked student among student below $\mu(h)$ on $\succ_h$.
Disclose, for all h, that $d(h)$ prefers h to $\mu(d)$.
Finally, check that the graph in which each h points to $d(h)$, and each student d points to $\mu(d)$, is acyclic.

While Protocol 9 is likely the most intricate protocol we discuss in this paper, the intuition behind it is simple: if a matching μ was stable but not the outcome of DA, then there has to be another stable μ' that Pareto improves upon μ. In μ', each schools' matching must be (weakly) lower-priority than in μ. Then, the $d(h)$ students identified in Protocol 9 are the minimal worsening of each h's priority such that $d(h)$ would Pareto improve to re-match to h. From here, a simple application of classic truncation arguments in stable matching theory prove that such a μ' exists if and only if there is a cycle in the graph described in Protocol 9.[34] Thus:

Theorem 4.3. *Protocol 9 is an unambiguous justification protocol for DA.*

References

Akbarpour, M., & Li, S. (2020). Credible auctions: A trilemma. *Econometrica*, *88*(2), 425–467 (cit. on p. 4).

Ausubel, L. M. (2004). An efficient ascending-bid auction for multiple objects. *American Economic Review*, *94*(5), 1452–1475 (cit. on p. 15).

Azevedo, E. M., & Leshno, J. D. (2016). A supply and demand framework for two-sided matching markets. *Journal of Political Economy*, *124*(5), 1235–1268 (cit. on pp. 3, 20).

[33] This protocol is inspired by a similar one in the theoretical computer science paper Gonczarowski and Thomas (2024) on the topic of communication complexity. That paper presents the protocol in a different form based on the representation of the stable matching lattice from Gusfield and Irving (1989).

[34] In more detail, the hard direction of this claim is to show that if μ' exists, then the corresponding cycle exists. To see this, consider the *school* proposing DA algorithm in which each h truncates their preference list just before $\mu(h)$, and additionally, some h who improves in μ' truncates $\mu(h)$ off her list as well. Then, since a routine verification checks that μ' must be stable under these modified preferences as well, we can examine the sequence of proposals made in school-proposing DA after these truncations, and recover the existence of the desired cycle.

Bartholdi, L., Hann-Caruthers, W., Josyula, M., Tamuz, O., & Yariv, L. (2021). Equitable voting rules. *Econometrica*, *89*(2), 563–589 (cit. on p. 27).

Bó, I., & Hakimov, R. (2024). Pick-an-object mechanisms. *Management Science*, *70*(7), 4693–4721 (cit. on p. 4).

Börgers, T., & Li, J. (2019). Strategically simple mechanisms. *Econometrica*, *87*(6), 2003–2035 (cit. on p. 4).

BPS. (2005). Recommendation to implement a new BPS assignment algorithm [Presentation prepared by BPS (Boston Public Schools) Strategic Planning Team. Accessed: 2026-02-09]. https://piazza.com/class_profile/get_resource/ij5ugfokmx06ng/imhuju6gqz61z (cit. on p. 3).

Brandt, F., & Sandholm, T. (2005). Unconditional privacy in social choice. *TARK*, 207–218 (cit. on p. 6).

Brandt, F., & Sandholm, T. (2008). On the existence of unconditionally privacy-preserving auction protocols. *ACM Transactions on Information and System Security (TISSEC)*, *11*(2), 1–21 (cit. on p. 6).

Canetti, R., Fiat, A., & Gonczarowski, Y. A. (2025). *Zero-knowledge mechanisms* [Working Paper. Abstract in EC'25]. (Cit. on p. 6).

Demange, G., Gale, D., & Sotomayor, M. (1986). Multi-item auctions. *Journal of Political Economy*, *94*(4), 863–872. https://doi.org/10.1086/261411 (cit. on p. 16).

Despotakis, S., Ravi, R., & Sayedi, A. (2021). First-price auctions in online display advertising. *Journal of Marketing Research*, *58*(5), 888–907 (cit. on p. 1).

Dreyfuss, B., Glicksohn, O., Heffetz, O., & Romm, A. (2025). Deferred acceptance with news utility [Forthcomming]. *Management Science* (cit. on p. 4).

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, 265–284 (cit. on p. 6).

Eilat, R., Eliaz, K., & Mu, X. (2025). *Privacy preserving auctions* [Abstract in EC'25]. Centre for Economic Policy Research. (Cit. on p. 6).

Fernandez, M. A. (2018). *Deferred acceptance and regret-free truth-telling* [Working Paper]. (Cit. on p. 4).

Gonczarowski, Y. A., Heffetz, O., Ishai, G., & Thomas, C. (2024). Describing deferred acceptance and strategyproofness to participants: Experimental analysis [NBER working paper no. 33020]. *Proceedings of the 25th ACM Conference on Economics and Computation*, 416–417 (cit. on p. 4).

Gonczarowski, Y. A., Heffetz, O., & Thomas, C. (2025). *Strategyproofness-exposing descriptions of matching mechanisms* [NBER Working Paper No. 31506. Previous version ("Strategyproofness-exposing mechanism descriptions") appeared as an abstract in the Proceedings of the 24th ACM Conference on Economics and Computation (EC'23)]. https://arxiv.org/abs/2209.13148 (cit. on p. 4).

Gonczarowski, Y. A., Nisan, N., Ostrovsky, R., & Rosenbaum, W. (2019). A stable marriage requires communication. *Games and Economic Behavior*, *118*, 626–647 (cit. on p. 5).

Gonczarowski, Y. A., & Thomas, C. (2024). Structural complexities of matching mechanisms. *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, 455–466 (cit. on pp. 4, 5, 21).

Google. (2019). Simplifying programmatic: First price auctions for google ad manager [Announcement Blog Post. Accessed: 2026-02-09]. (Cit. on p. 1).

Grigoryan, A., & Möller, M. (2024). *A theory of auditability for allocation mechanisms* [Working Paper. Abstract appeared in EC 2023.]. (Cit. on pp. 1, 5).

Gusfield, D., & Irving, R. W. (1989). *The stable marriage problem: Structure and algorithms*. MIT Press. (Cit. on p. 21).

Hakimov, R., & Raghavan, M. (2025). Improving transparency and verifiability in school admissions: Theory and experiment. *Management Science* (cit. on p. 5).

Hartman, E., Segal-Halevi, E., & Tao, B. (2025). It's not all black and white: Degree of truthfulness for risk-avoiding agents. *Proceedings of the 26th ACM Conference on Economics and Computation*, 996–1016 (cit. on p. 4).

Haupt, A., & Hitzig, Z. (2025). *Contextually private mechanisms* [Abstract in EC '22.]. (Cit. on p. 6).

He, K., Sandomirskiy, F., & Tamuz, O. (2026). Private private information. *Journal of Political Economy* (cit. on p. 6).

Izmalkov, S., Lepinski, M., & Micali, S. (2011). Perfect implementation. *Games and Economic Behavior*, *71*(1), 121–140 (cit. on p. 6).

Kasman, M., & Valant, J. (2019). The opportunities and risks of k-12 student placement algorithms [Accessed February 2025]. *The Brookings Institute*. https://www.brookings.edu/research/the-opportunities-and-risks-of-k-12-student-placement-algorithms/ (cit. on p. 1).

Katuščák, P., & Kittsteiner, T. (2025). Strategy-proofness made simpler. *Management Science*, *71*(9), 7560–7578 (cit. on p. 4).

Komo, A., Kominers, S. D., & Roughgarden, T. (2024). *Shill-proof auctions* [Abstract in EC'25]. (Cit. on p. 4).

Kushilevitz, E., & Nisan, N. (1997). *Communication complexity*. Cambridge University Press. (Cit. on p. 5).

Leshno, J. D., & Lo, I. (2021). The cutoff structure of top trading cycles in school choice. *The Review of Economic Studies*, *88*(4), 1582–1623 (cit. on p. 3).

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review*, *107*(11), 3257–87 (cit. on p. 4).

Liu, T. (2025). *Worst-case privacy* [Working Paper]. (Cit. on p. 6).

Meisner, V., & von Wangenheim, J. (2023). Loss aversion in strategy-proof school-choice mechanisms. *Journal of Economic Theory*, *207*, 105588 (cit. on p. 4).

Milgrom, P., & Segal, I. (2020). Clock auctions and radio spectrum reallocation. *Journal of Political Economy*, *128*(1), 1–31 (cit. on p. 6).

Möller, M. (2026). Transparent matching mechanisms. *Theoretical Economics*, *21*(1), 22–43 (cit. on pp. 2, 4, 5).

Nagel, L., & Saitto, R. (2024). *As-if dominant strategy mechanisms* [Abstract in EC'24]. (Cit. on p. 4).

Nagel, L., & Saitto, R. (2025). A measure of complexity for strategy-proof mechanisms. [Abstract in EC'23]. *Conditionally accepted at Econometrica* (cit. on p. 4).

Pai, M. M., & Roth, A. (2013). Privacy and mechanism design. *ACM SIGecom Exchanges*, *12*(1), 8–29 (cit. on p. 6).

Pycia, M., & Troyan, P. (2023). A theory of simplicity in games and mechanism design. *Econometrica*, *91*(4), 1495–1526 (cit. on p. 4).

Robertson, S., Nguyen, T., & Salehi, N. (2021). Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (cit. on p. 1).

Segal, I. (2007). The communication requirements of social choice rules and supporting budget sets. *Journal of Economic Theory*, *136*(1), 341–378 (cit. on p. 5).

Shirakawa, R. (2025). Simple manipulations in school choice mechanisms. *American Economic Journal: Microeconomics*, *17*(4), 202–233 (cit. on p. 4).

Troyan, P., & Morrill, T. (2020). Obvious manipulations. *Journal of Economic Theory*, *185*, 104970 (cit. on p. 4).

Woodward, K. (2020). *Self-auditable auctions* [Working paper]. (Cit. on p. 5).

A Additional Matterial

A.1 Additional Preliminaries

For many notions in mechanism design, it is helpful to consider an outcome as a separate specification of each agent's outcome. To define this, we use the notion of agent i's indifference relation. For $a, b \in \mathcal{A}$, we write $a \sim_i^* b$ when all types of i are indifferent between a and b (i.e., $a \succeq_i^* b \wedge b \succeq_i^* a$), and define $[a]_i = \{b \in \mathcal{A} \mid a \sim_i b\}$. For example, in matching environments, since agent i cares only about his own match, $\mu \sim_i \nu$ if and only if $\mu(i) = \nu(i)$. We say that $[a]_i$ is *i's outcome in a*, and denote the set of all i's possible outcomes by $\mathcal{A}_i$. Using this notion, one can equivalently represent the outcomes $\mathcal{A}$ as elements of $\left\{ \left([a]_i\right)_{i \in N} \;\middle|\; a \in \mathcal{A} \right\} \subseteq \bigtimes_{i \in N} \mathcal{A}_i$, which represents the feasible combinations of individual agents' outcomes.

A.2 Additional Exposition on our Model

Remark A.1 (The reveals-less-than order)**.** We now discuss the properties of the reveals-less-than order, $\lesssim$. For disclosure sets, we define $\mathcal{D}_1 \lesssim \mathcal{D}_2$ to mean $\forall D_1 \in \mathcal{D}_1 : \exists D_2 \in \mathcal{D}_2 : D_1 \supseteq D_2$, and $\forall D_2 \in \mathcal{D}_2 : \exists D_1 \in \mathcal{D}_1 : D_1 \supseteq D_2$. In the special case where $\mathcal{D}_1$ and $\mathcal{D}_2$ are partitions, $\lesssim$ is equivalent to the usual refinement order; we generalize this order to our model, in which disclosures can intersect. As for partitions, not every pair of disclosure sets can be compared under $\lesssim$; e.g., $\{\{1,2\},\{3\}\}$ and $\{\{1\},\{2,3\}\}$ are incomparable. The order $\lesssim$ is easily seen to be reflexive and transitive, hence, it is a preorder. An element x is called minimal (resp., maximal) in a preordered set if for all y in the set, $y \lesssim x$ (resp., $x \lesssim y$) implies that $x \lesssim y$ (resp., $y \lesssim x$).

When $\mathcal{D}_1$ and $\mathcal{D}_2$ are arbitrary, $\lesssim$ need not be antisymmetric. For example, $\mathcal{D}_1 = \{\{1\},\{1,2,3\}\}$ and $\mathcal{D}_2 = \{\{1\},\{1,2\},\{1,2,3\}\}$ are two distinct disclosure sets such that $\mathcal{D}_1 \lesssim \mathcal{D}_2$ and $\mathcal{D}_2 \lesssim \mathcal{D}_1$. However, one can show that if $\mathcal{D}_1$ is an antichain (i.e., $\forall D_1, D_1' \in \mathcal{D}_1 : D_1 \subseteq D_1' \implies D_1 = D_1'$), then $\mathcal{D}_1 \lesssim \mathcal{D}_2 \wedge \mathcal{D}_2 \lesssim \mathcal{D}_1$ implies that $\mathcal{D}_1 = \mathcal{D}_2$. Hence, if $\mathcal{D}_1$ is an antichain, one can prove that $\mathcal{D}_1$ is $\lesssim$-minimal by showing that there is no $\mathcal{D}_2 \neq \mathcal{D}_1$ with $\mathcal{D}_2 \lesssim \mathcal{D}_1$. Moreover, if we restrict attention to antichains, $\lesssim$ is a partial order.

Antichains have a natural interpretation in the context of disclosure sets: they are those with no "dominated" disclosures that reveal strictly more information than some other disclosure. All disclosure sets we construct will be antichains, so we can show that such protocols are $\lesssim$-minimal using the approach above. Moreover, since $\lesssim$ is defined on protocols by separately considering the set of disclosures consistent with each outcome $a \in \mathcal{A}$, all the same remarks hold for protocols.

Remark A.2. Table 1 shows two unambiguous justification protocols for a mechanism f, where f is depicted similarly to in Figure 2 for the Two-Trader Mechanism, and where disclosures justifying outcome 1 are represented by contiguous regions with the same colored shading.

1	0
1	1
0	1

1	0
1	1
0	1

Table 1: Two $\lesssim$-minimal, $\lesssim$-incomparable unambiguous justification protocols.

A.3 Additional Result for Voting

We now consider a simple extension of our voting results to the classic single-peaked environment. In *single-peaked* voting, we write $\mathcal{A} = \{1, 2, \ldots, m\}$ for some $m \geq 1$, and agent i's type is a single-peaked preference over $\mathcal{A}$.[35] In the single-peaked environment, the Median mechanism has agents report their peak candidate, and the outcome is the median candidate (in the usual order on $\mathcal{A} = \{1, 2, \ldots, m\}$). For simplicity assume the number of voters n is odd.

Proposition A.3. *Any disclosure protocol which unambiguously justifies the Median Voting mechanism must sometimes fully reveal all n votes.*

Proof. Suppose the outcomes are $\mathcal{A} = \{1, 2, 3\}$, and consider a report profile r_N such that one agent's peak/vote is at 2, and $(n-1)/2$ agents' have their peaks/votes at both 1 and at 3. See Figure 3 for an illustration of these reports.

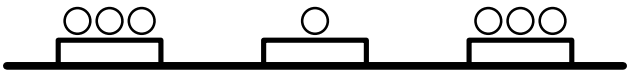


Figure 3: Illustration of the proof of Proposition A.3.

The outcome of Median at r_N is 2. However, observe that if we are allowed to change *any* agent's report, then we can make the outcome either 1 or 3. (Specifically, if the agent who votes for 2 changes to 1, the outcome is 1; if an agent who votes for 1 changes to 3, then the outcome is 3; and symmetrically for an agent who votes for 3.) Hence, if even a single report is not fully disclosed for $D \ni r_N$, then the outcome cannot be unambiguous at r_N. □

[35] Formally, a single-peaked preference is an ordering $\succ$ over $\mathcal{A}$ such that there exists a $t \in \mathcal{A}$, referred to as the *peak*, such that $t \succeq s$ for all $s \in \mathcal{A}$, and moreover, for all $x, y \in \mathcal{A}$ with either $x < y < t$ or $t < y < x$, we have $y \succ x$.

Interestingly, there are other natural voting rules that can be justified by disclosing a small fraction of the votes, such as the generalized representative democracy examples found in Bartholdi et al., 2021.

A.4 Discussion of Unambiguous and Uncontestable Justification

Recall that we define, for a disclosure D, an outcome a, and a report profile r_N, that:

- D *unambiguously justifies* a at r_N if $r_N \in D$, and for all $s_N \in D$, we have $a = f(s_N)$.
- D *uncontestably justifies* a at r_N if $r_N \in D$, and for all agents $i \in N$ and $s_{-i} \in D_{-i}$, we have $a \succeq_i^* f(r_i, s_{-i})$.

We now compare these two notions further, and in order to facilitate additional discussion, we introduces a third type of justification, as follows:

- D *individually-unambiguously justifies* a at r_N if $r_N \in D$, and for all agents $i \in N$ and $s_{-i} \in \mathcal{I}_{-i}(D)$, we have $[a]_i = [f(r_i, s_{-i})]_i$, where $[\cdot]_i$ denotes i-outcomes as defined in Section A.1.

Hence, individually-unambiguous justification adopts the information model of uncontestable justification—in which agents know their own report and are only aiming to check their own outcome—with the stronger requirement that the outcome (of each i) must be unambiguous (given what is known to i).

First, unambiguous justification means that an external third party directly learns enough information about the agents' reports to independently confirm the full outcome is correct. This provides an (arguably) strong notion of transparency in explaining why the outcome is a, for instance, no $\widetilde{a} \neq f(r_N)$ can ever be unambiguously justified. Moreover, if all disclosures $D \in \mathcal{D}$ ambiguously justify some outcome, then $D \in \mathcal{D}$ unambiguously justifies the outcome $f(r_N)$ if and only if D is consistent with r_N.

Second, individually-unambiguous justification means that each agent i—who already knows his own report—learns enough information about *other* agents' reports to confirm their own outcome a_i is correct. Again, no outcome $\widetilde{a} \neq f(r_N)$ can be individually-unambiguously justified, because a_i must be correct for each $i \in N$. However, there can be disclosures D that are consistent with r_N, but do not individually-unambiguously justify the outcome $a = f(r_N)$. For example:

Example A.4. For example, consider a first-price auction with agents $N = \{1, 2\}$. Suppose the disclosure D attests that agent 2 bid 2, and let a denote the outcome where agent 2 wins

and pays 2. Then, when the reports are $r_1 = 1$ and $r_2 = 2$, disclosure D is accurate and individually-unambiguously justifies a. In contrast, when the report r_1 is instead $r_1 = 3$, the disclosure D is still accurate but it no longer individually-unambiguously justifies outcome a.

Third, uncontestable justification means that each agent i learns only enough about others' reports to confirm that his own outcome could not have been *better* for him—regardless of his own type—than the announced one.[36] This is the weakest notion of justification we consider; for instance, in some mechanisms and disclosure rules, the designer can uncontestably justify outcomes *other than* the correct one $M(r_N)$.

Example A.5. For example, consider 2PA, and suppose agents $1, 2, 3$ report bids of $1, 2, 3$, respectively. The the outcome where agent 3 wins and pays 1 can be uncontestably justified by the disclosure attesting that agent 1 bids 1 and agent 3 bids 3.

Nevertheless, uncontestability still provides a compelling notion of justification: each agent is convinced that the announced outcome is at least as good for him as any outcome consistent with the reports, given the knowledge revealed to him through D.

Each of these three notions is successively more permissive than the other, in the sense that every unambiguous justification is (in particular) an individually-unambiguous justification, and every individually-unambiguous justification is (in particular) an uncontestable justification.

[36] One could generalize our notion of uncontestable justification to not require that $a_i \succeq_i^* M_i(r_i, s_{-i})$—meaning that the preference holds for *all* types of i—but rather to require that $a_i \succeq_i^{t_i} M_i(r_i, s_{-i})$ for some rationalizable t_i (for instance, taking $t_i = r_i$ in the context of strategyproof direct-revelation mechanisms, or taking t_i to be any type such that r_i is undominated in non-strategyproof mechanisms). However, this is not needed for any of our results, so we opt for our simpler and stronger notion of uncontestability.