Matching with Strategic Consistency*

Marzena Rostek† and Nathan Yoder‡

May 18, 2025

Abstract

In many environments, agents form agreements that have externalities or are multilateral, and may view some agreements as substitutable and others as complementary. This paper presents an approach that ensures the existence of stable outcomes in any environment, including those with arbitrary externalities, preferences, and market structures. It does so by endogenizing the agents' choice functions while employing the standard stability concept. Instead of assuming that each agent chooses their favorite set of contracts, we require agents to choose optimally given correct beliefs about the choices of others. We show that stable outcomes are uniquely pinned down by agents' beliefs, which can be microfounded by their relative bargaining power. Our results provide new tools for the counterfactual analysis of stable outcomes and allow the use of matching-theoretic stability in new applications.

Keywords: Externalities, multilateral matching, matching with contracts, stability

* This paper subsumes parts of "Matching with Multilateral Contracts". The authors are grateful to numerous colleagues for their helpful comments and suggestions. We also thank conference attendees at the North American Summer Meetings of the Econometric Society at UCLA, 2023 ASSA in New Orleans, 33rd Stony Brook International Conference on Game Theory, and seminar participants at Virtual Seminars in Economic Theory (VSET), Brown University, and University of Oregon. This material is based upon work supported by the National Science Foundation under Grant No. SES-1357758.

† University of Wisconsin-Madison, Department of Economics; E-mail: mrostek@ssc.wisc.edu.

‡ University of Georgia, Terry College of Business, John Munro Godfrey, Sr. Department of Economics; E-mail: nathan.yoder@uga.edu.

1 Introduction

Matching theory has facilitated the study of many applications where agents negotiate agreements with one another. In particular, the literature has extensively explored settings where these agreements (sometimes referred to as *contracts*) are substitutable, bilateral, and do not have externalities. In these environments, the literature has established the general existence of *stable outcomes* — sets of agreements that are robust to both individual deviations to remove agreements and joint deviations to form new ones — and provided ways to find them (e.g., the seminal work of Gale and Shapley (1962), Kelso and Crawford (1982), Hatfield and Milgrom (2005), and Hatfield et al. (2013)).

But in many environments, agents may form agreements that have externalities, are not substitutable, or are multilateral. Agreements among competing firms to merge or engage in collusion may exert externalities on other market participants, influencing their incentives to enter into alternative agreements. These agreements might involve more than two firms. And depending on the market structure, some could be complementary, while others are substitutable. Analogous features characterize international treaties, legislative negotiations to pass a bill, and agreements to add a healthcare provider to an insurance network, among others. The prevalence of these features has created demand in applied research for matching-theoretic tools that are capable of accommodating them (e.g., Agarwal et al. (2021)).

Accommodating any of these features in general matching environments has been challenging, because each prevents standard approaches from guaranteeing the existence of stable outcomes.[1] This paper introduces an approach that allows matching-theoretic stability to be applied in any setting, including those with general externalities, arbitrary preferences and market structures, and multilateral agreements. Our key observation is that the challenges presented by these features can each be attributed to an implicit assumption about the way that choices are derived. Specifically, when agents choose from a set of available agreements, they select their *favorite* subset, thus behaving as if each available contract will go into effect if they choose it. However, for contracts to go into effect, they must also be chosen by other agents. We show that when agents take others' choices into account — i.e., they choose optimally given beliefs about others' choices, and those beliefs are correct at every set of available agreements and consistent across sets of available agreements — stable outcomes always exist (Theorems 1 and 2). Thus, we ensure existence not by modifying the usual definition of stability, but by endogenizing the agents' choice functions (and the beliefs that generate

[1] When both complementarity and substitutability are present in the same environment, the existence of stable outcomes is generally not guaranteed (Hatfield and Kojima (2008)). Moreover, standard existence results do not always apply in the presence of externalities (Sasaki and Toda (1996)) or in the absence of a key assumption on market structure (acyclicity) that is incompatible with multilateral agreements (Gale and Shapley (1962); Hatfield and Kominers (2012)).

them). We call such a profile of choice functions and beliefs *strategically consistent*. Once we endogenize agents' choices given their beliefs, the standard stability concept pins down an outcome (and, as we show in Theorem 1, does so *uniquely*.)

This result does not require conditions on preferences (e.g., (full) substitutability or complementarity), market structure (e.g., acyclic trading networks), or the agents that agreements can involve (e.g., bilateral agreements) or affect (e.g., no externalities). The literature has demonstrated that these conditions ensure that stable outcomes can be represented as fixed points of monotone operators; by Tarski's theorem, such fixed points always exist.[2] Theorem 2 instead constructs fixed points in profiles of *choice functions and beliefs* for all agents, without relying on a monotonicity condition or a fixed point theorem. Each such profile of choice functions then pins down a unique stable outcome (Theorem 1).

Given the fixed point relationship between optimal choices and correct beliefs, there may be many outcomes that can arise from strategically consistent profiles of choice functions and beliefs. While our framework permits alternative restrictions, our main characterization results (Theorems 3 and 4) focus on profiles that are *Pareto optimal*, in the sense that agents never reject contracts from a Pareto-dominating outcome merely because of coordination failure. As it turns out, one can construct such profiles by solving a set of constrained social planner's problems (Theorem 3).

This provides one direction of our "welfare theorem" for strategic consistency (Theorem 4), which parallels welfare theorem results in the matching literature with transferable utility (e.g., Hatfield et al. (2013)). But since it decentralizes efficient outcomes with a profile of *beliefs* rather than *prices*, Theorem 4 applies even to nontransferable utility settings, and without the conditions on preferences or market structure that are, in general, necessary for competitive equilibrium prices to exist.

This characterization has important implications. First, observe that strategically consistent beliefs play the same role in our framework that bargaining weights play in Nash bargaining: Each profile of beliefs or bargaining weights predicts a *unique* outcome, but different profiles of beliefs or weights predict different outcomes. Theorem 3 makes this analogy precise by showing that a strategically consistent profile — and hence (by Theorem 1) the stable outcome it gives rise to — can be uniquely pinned down by agents' bargaining power, i.e., their weights in a Nash product. Thus, Nash bargaining weights can be treated as a "reduced form" for strategically consistent beliefs — or, if they are taken as primitive, a microfoundation for those beliefs. Intuitively, the agents' beliefs about one another's choices reflect their relative bargaining power, and vice versa. Consequently, even in environments where stable outcomes exist with nonstrategic choice, Theorem 3 gives a foundation for pinning down a single outcome using agents' relative bargaining power. In this sense, the multiplicity of pre-

[2]Since the set of stable outcomes is discrete, fixed points are not guaranteed more generally.

dicted outcomes is a prediction across environments where agents' relative bargaining power differs.

Second, Theorems 3 and 4 provide a tool that makes the problem of finding stable outcomes more tractable. Instead of checking all possible deviations (as is common in network or coalition formation) or running a matching algorithm (as is common in two-sided matching), Theorem 3 shows that one only has to solve a single optimization problem — specifically, a welfare maximization problem — to find a stable outcome. And given the agents' beliefs — or their bargaining weights — this stable outcome is uniquely pinned down in any environment. Theorem 4, on the other hand, allows us to find *all* outcomes that are stable for *some* strategically consistent profile satisfying Pareto optimality, simply by computing a *constrained* Pareto frontier.[3]

Third, Theorems 3 and 4 allow new comparative statics and counterfactual predictions. As is well known, with the standard approach, following a change in the environment (e.g., when the government levies a tax or a regulator disallows a contract), the new set of stable outcomes may be empty, since the existence of a stable outcome is not guaranteed. Even if existence is not a problem, the specific outcome that will result is not pinned down. Endogenizing agents' choices allows one to make a unique prediction about the counterfactual outcome: Theorem 3 allows one to recover agents' relative bargaining power (in the form of bargaining weights) from an observed stable outcome. In a counterfactual scenario, one can apply Theorem 3 with these bargaining weights to uniquely pin down the new stable outcome. We give an example of this procedure in Section 5. More generally, we discuss in Section 4.1 how our approach allows the application of matching- theoretic tools to make unique predictions about the way an observed outcome would change under different kinds of counterfactual scenarios.

Because our approach allows general externalities, market structures, and preferences, and permits multilateral agreements, it enables the use of matching-theoretic stability in applications where it has not traditionally been used, such as network and coalition formation, and bargaining with externalities. As we illustrate in Section 4, this allows one to make predictions in these environments that are robust to *arbitrary* deviations.

This is a larger set of deviations than is considered by existing tools used in these environments. Specifically, relative to perhaps the most common solution concept in network formation, *pairwise stability* (Jackson and Wolinsky (1996)), (matching-theoretic) stability permits agents to swap links (important when they are substitutes or with externalities) or add multiple links (significant when they are complements or with externalities).[4] And relative

[3]More generally, we show that one can construct strategically consistent profiles of choices and beliefs from an order on the set of nonstrategically individually rational outcomes. In the Online Appendix, we consider restrictions on beliefs besides Pareto optimality, and show that each places additional structure on this order, capturing the corresponding restrictions on beliefs about other agents' choices.

[4]Sadler (2023) also allows agents to swap links, rather than merely sever them, and establishes some of the

to *Nash-in-Nash bargaining* (Horn and Wolinsky (1988)), popular in applied work on environments with externalities, stability allows agents to simultaneously alter their agreements with multiple counterparties, endogenizes the agreements counterparties make, and allows agents to exclude counterparties by declining to make any agreements with them. The main obstacle to using stability in these contexts is that with the standard, nonstrategic approach to choice, these additional deviations create an existence problem that is not present with pairwise stability or the Nash-in-Nash solution. Rather than considering robustness to a smaller class of deviations, as those solution concepts do, strategic consistency sidesteps the existence problem by endogenously determining which deviations are relevant.

Similarly, in models where coalitions can form, our results allow existence and uniqueness without imposing any restrictions on deviations allowed (e.g., only to subsets of coalitions) or coalitions that can form (e.g., partitions). In particular, because matching-theoretic stability requires outcomes to be robust to *arbitrary* deviations, its predictions are independent of assumptions about the specific ways that agents form coalitions.

Related Literature

Our paper relates to three strands of the matching literature. The first strand seeks to extend matching theory to accommodate agents' preferences over agreements that do not satisfy the classical substitutability condition. Several studies have demonstrated that the tools of matching theory can be applied to settings in which preferences satisfy more general forms of substitutability, such as full substitutability (Ostrovsky (2008); Hatfield et al. (2013); Fleiner et al. (2019)), or by applying substitutability under a basis change on the set of contracts, as in gross substitutes and complements (Sun and Yang (2006, 2009); Teytelboym (2014)). Another approach considers environments where all contracts are complementary rather than substitutable (Rostek and Yoder (2020)).[5] Other authors have shown that, instead of imposing restrictions on preferences, we can rely on conditions on the market structure (e.g., Bando and Hirai (2021)) or its size (e.g., Jagadeesan and Vocke (2021)), relax feasibility constraints (Nguyen and Vohra (2018)), or consider outcomes that are dynamically stable in markets with patient firms (Liu et al. (2023)). We show that when agents' choices are endogenized by requiring them to be optimal given correct beliefs about the choices of others, rather than being determined by a single-agent optimization problem, the existence of stable outcomes can be established for arbitrary preferences over agreements, market structures, and

classical matching-theoretic results in networks without externalities.

[5]While our results in Rostek and Yoder (2020) also allow for multilateral contracts and externalities, these features are not a central focus there. With nontransferable utility, they do not create any additional challenges for the existence and characterization results from Rostek and Yoder (2020), precisely because complementarity ensures that whenever a block is relevant for stability, the implicit assumptions that nonstrategic agents make about other agents' choices turn out to be correct.

market sizes.

Second, our work also contributes to the literature on matching with externalities. One strand of this literature explores the externalities that arise in settings of applied interest, such as labor market matching with couples (e.g., Kojima et al. (2013)). Other studies consider general environments in matching markets with two sides (e.g., Bando (2012), Fisher and Hafalir (2016), Pycia and Yenmez (2023), and Liu et al. (2023)). Of these, our paper is closest to Pycia and Yenmez (2023), who introduce a matching with contracts framework in two-sided settings with a classical substitutability condition extended to allow for externalities. Our results apply to environments where preferences may not satisfy substitutability and whose market structures may not be two-sided.[6]

Within the literature on two-sided matching markets with externalities, papers like Sasaki and Toda (1996) and Hafalir (2008) consider agents who determine what to take as given about other agents' matchings through the use of an *estimation function* — a concept akin to the beliefs considered in this paper.[7] The agent then evaluates potential partners by taking as given the *least preferred* outcome that is plausible according to her estimation function. While Sasaki and Toda (1996) take these estimation functions as a primitive of the model, Hafalir (2008) allows them to be determined based on a consistency condition: an agent's estimation function treats matchings as plausible if they are stable when the agent and her partner are removed from the market. In contrast, we require an agent's belief to be correct, i.e., match the choices made by other agents from the available set of contracts.[8]

Finally, our paper contributes to the literature on multilateral contracts. There is a large literature on the formation of coalitions or clubs; see, e.g., Pycia (2012); Ellickson et al. (1999). Hatfield and Kominers (2015) initiated the study of multilateral agreements in the matching with contracts framework. They examine settings with continuously divisible contracts and transferable utility, and leverage the concavity of agents' valuations to establish the existence of competitive equilibria (and thus, as they demonstrate, stable outcomes). As is standard in the literature, we work with environments where the set of contracts is discrete, rather than

[6]In Rostek and Yoder (2023), we focus on two-sided markets with externalities, and consider a weaker notion of strategic sophistication: Instead of requiring that agents have correct beliefs about *all* other agents' choices, we only require agents to have correct beliefs about the choices of others *on the same side of the market*. We show that the *standard substitutability* and *monotone externalities* conditions introduced by Pycia and Yenmez (2023) ensure the existence of profiles of choice functions and beliefs that satisfy this notion of strategic consistency, and hence stable outcomes. Intuitively, these conditions ensure that agents on the same side of the market can all form correct beliefs about each other's behavior.

[7]While the beliefs considered in this paper specify the choices that an agent thinks others would make from a proposed set of contracts, estimation functions give a *set* of outcomes that an agent thinks are plausible, given the identity of the individual she is matched to.

[8]One approach to ruling out agents' disagreements about the outcomes of blocking proposals has been to strengthen the solution concept (e.g., setwise stability (Klaus and Walzl (2009))). Strategic consistency eliminates disagreements without strengthening the usual stability concept, while also ensuring existence even with externalities or non-substitutable preferences.

convex, and so the tools they use from convex analysis are not available. On the other hand, Bando and Hirai (2021) investigate a setting with a finite number of multilateral contracts, as we do. Unlike this paper or Hatfield and Kominers (2015), the authors use conditions on the market structure that guarantee the existence of stable outcomes, irrespective of agents' preferences.

Our paper also relates to several papers in the literature on matching with incomplete information that also explicitly incorporate agents' beliefs. In, e.g., Chakraborty et al. (2010), Liu et al. (2014), Liu (2020), and Liu (2022), agents form beliefs about other agents' privately observed *types*, and make choices given those beliefs. We do not consider incomplete information. Instead, beliefs in our paper are deterministic, and pertain to the contracts others will *choose* from each possible set of available contracts, rather than their types.[9]

The structure of the paper is as follows. Section 2 introduces the environment. Section 3 presents an example that illustrates the paper's main ideas, and provides our main existence and characterization results. Section 5 illustrates new tools for counterfactual analysis that are facilitated by our results.

2 Model

2.1 Setting

We work in a matching with contracts framework adapted to accommodate externalities and agreements among more than two agents.[10] Additionally, we do not assume a certain market structure (such as two-sidedness or acyclicity). Our model accommodates, for instance, network formation, many-to-many matching with contracts, and coalition formation.

There is a finite set I of agents and a finite set X of agreements, or *contracts*, that they can sign with one another. Each contract $x \in X$ requires the agreement of a set of agents $N(x) \subseteq I$ in order to be enacted. For sets of contracts $Y \subseteq X$, we write $N(Y) := \bigcup_{x \in Y} N(x)$. We assume that each contract involves at least two agents: For all x, $|N(x)| \geq 2$. A contract x is *multilateral* if $|N(x)| > 2$ and *bilateral* if $|N(x)| = 2$. For each agent $i \in I$, denote the set of contracts requiring i's agreement as $X_i := \{x \mid i \in N(x)\}$. In keeping with the literature, we say that X_i is the set of contracts that *name* i. Similarly, let $X_J := \bigcup_{i \in J} X_i$,

[9] In particular, the beliefs in our model are not equivalent to beliefs in the sense of Liu (2022) about the output of a correlation device: In cooperative games with incomplete information, Liu (2022, Theorems 1 and 6) shows that without payoff-relevant uncertainty, the presence of payoff-irrelevant signals cannot change the set of predictions consistent with stability. As we show, in matching models with complete information, considering beliefs to determine choice can alter the set of outcomes that are consistent with stability, in particular by making it nonempty (Theorem 1).

[10] Agreements between more than two agents cannot be represented by multiple independent bilateral agreements; see Example S.2 in the Online Appendix.

let $X_{-i} := X \setminus X_i$, and for sets of contracts $Y \subseteq X$, write $Y_i := Y \cap X_i$ and $Y_{-i} := Y \cap X_{-i}$.

Each agent i has preferences over sets of implemented contracts, or *outcomes*, which are represented by a utility function $u_i : 2^X \to \mathbb{R}_+$.[11] This allows for *externalities*: Agents' utility can depend on the presence of contracts that do not name them. In settings where it does not — i.e., when $u_i(Y \cup Z) = u_i(Y \cup Z')$ for each $Z, Z' \subseteq X_{-i}$ and $i \in I$ — we say that there are *no externalities*.

A *choice function* for agent i is a function $C_i : 2^{X_i} \times 2^{X_{-i}} \to 2^{X_i}$. Its arguments are the sets of contracts that are *available* — those being discussed in a negotiation — to agent i and to agents other than i. When agent i's choice function is C_i, $C_i(Y_i|Y_{-i})$ gives the set of contracts that agent i chooses from the set of contracts Y_i available to him, given that the set of contracts available to other agents is Y_{-i}. Its second argument allows for the presence of externalities.

In Section 3, we describe two different ways in which these choice functions can be derived, given agents' preferences. In order to ensure that these endogenously derived choice functions are single-valued, we assume that agents' payoff functions have no indifferences, conditional on the set of contracts that do not name them: $u_i(Y \cup X') \neq u_i(Z \cup X')$ for each distinct $Y, Z \subseteq X_i$ and $X' \subseteq X_{-i}$.

2.2 Stability

Our solution concept is the usual matching-theoretic definition of *stability*, generalized to our setting with multilateral contracts and externalities.[12]

Definition (Stability). Given choice functions $\{C_i\}_{i \in I}$, a set of contracts $Y \subseteq X$ is *stable* if it is

i. *Individually rational*: $Y_i = C_i(Y_i|Y_{-i})$ for all $i \in N$.

ii. *Unblocked*: There does not exist a nonempty $Z \subseteq (X \setminus Y)$ such that for all $i \in N(Z)$, $Z_i \subseteq C_i((Z \cup Y)_i|(Z \cup Y)_{-i})$.

In words, a set of contracts Y is stable if (i) when Y is the set of available contracts, no one rejects any contracts from it (individual rationality), and (ii) no group of agents can propose to change the set of contracts in place by adding a new set of contracts Z, or *block*, that they

[11] Throughout, we use 2^B to denote the power set of a set B.

[12] In particular, our solution concept coincides with those of Gale and Shapley (1962) (one-to-one matching), Hatfield and Milgrom (2005) (many-to-one matching with contracts), and Hatfield and Kominers (2012) (matching on networks) in the settings they consider.

are each willing to choose when made available (i.e., discussed in a negotiation) alongside Y.[13]

We accommodate externalities by allowing agents who participate in a block to take into account the contracts available to the agents they negotiate with: the second argument of the choice function in (ii) includes both the existing contracts Y_{-i} and blocking contracts Z_{-i} that do not name agent i.[14]

3 Strategic Consistency

This section presents the paper's main idea, which stems from a simple yet crucial observation about the relationship between the (non-)existence of stable outcomes and the way choice functions are derived from preferences.

By definition, the stability of an outcome is completely determined by agents' choice functions. In matching models like ours where preferences (rather than choice functions) are taken as primitive, the standard approach to deriving those choice functions is to let them be the agents' favorite subsets of the available contracts. With externalities, this usually becomes their favorite subset conditional on the enactment of whatever set of contracts they take as given for everyone else. That is,

$$\hat{C}_i(Y_i|Y_{-i}) := \arg\max_{S \subseteq Y_i} u_i(S \cup Y_{-i}) \text{ for each } Y \subseteq X. \tag{1}$$

When they make choices this way, agents behave as if each available contract will go into effect if they choose it. But for a contract to go into effect, it must *also* be chosen by *other* agents. Hence, when agents are equipped with the choice functions described in (1), they implicitly assume that *all contracts that are available will actually be chosen by the other agents they name.* Our key observation is that the standard approach does not always yield a stable outcome precisely because these assumptions may be incorrect. A familiar example

[13] We could alternatively define a block as the *full proposal* for changing the set of contracts, and replace (ii) with

ii'. There does not exist $Z \subseteq X$ such that for all $i \in N(Z \setminus Y)$, $Z_i = C_i((Z \cup Y)_i|(Z \cup Y)_{-i})$.

If we did, our stability concept would generalize *weak setwise stability* (Klaus and Walzl (2009)), rather than stability, to account for externalities. These definitions are equivalent with strategic consistency: Whenever a block (in the sense of (ii)) is successful, all agents agree about the set of contracts that will obtain after it occurs (as they must in (ii')). But with nonstrategic choice, (ii') is stronger than (ii), and so replacing (ii) with it weakens the definition of stability. We discuss this point in greater detail in Section S.2 of the Online Appendix.

[14] We generalize the usual definition of stability to accommodate externalities in a slightly different way than Pycia and Yenmez (2023) do in their two-sided setting. Under the definition they adopt, agents in a blocking coalition do not anticipate any changes to the set of contracts signed by other agents, even the other members of the blocking coalition. (That is, when evaluating a block Z of Y, the second argument of the choice function in their concept is Y_{-i}, rather than $Z_{-i} \cup Y_{-i}$.) Our stability definition instead assumes that agents in a blocking coalition account for the contracts added by the other agents in the coalition.

illustrates this point.

Example 1 (Roommate Problem). Consider the classical roommate problem from Gale and Shapley (1962). Three friends must come to an agreement about which two of them will rent an apartment together: $I = \{1, 2, 3\}$, $X = \{x_{12}, x_{23}, x_{31}\}$, and $N(x_{ij}) = \{i, j\}$ for each $i, j \in I$. There are no externalities. Each agent prefers having any roommate to being unmatched, and cannot be part of two roommate agreements: $u_i(\{x_{ij}\}) > u_i(\emptyset) > u_i(\{x_{ij}, x_{ik}\})$ for each $i \in I$ and each $j \neq k \neq i$. Moreover, agents' preferences over roommates form a cycle: $u_1(x_{12}) > u_1(x_{31})$, $u_2(x_{23}) > u_2(x_{12})$, and $u_3(x_{31}) > u_3(x_{23})$.

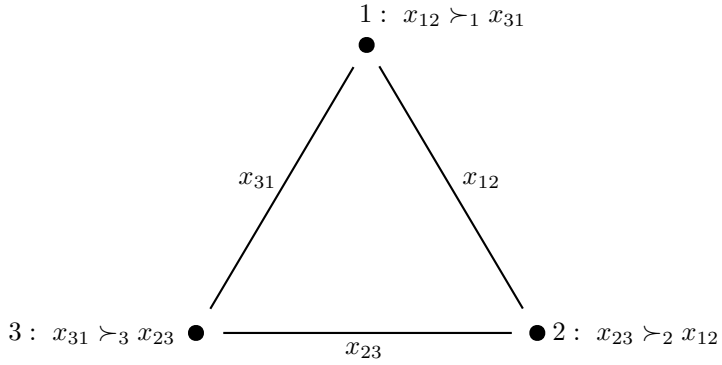

Figure 1: A visual description of the environment in Example 1.

Part 1. Suppose that preferences translate into choices in the standard way, and each agent's choice function is defined by (1). Then there is a *blocking cycle* that makes every outcome unstable: When the set of available contracts is $\{x_{12}, x_{31}\}$, agent 1 chooses x_{12} (and rejects x_{31}), while agent 2 chooses the only agreement available to him, x_{12}. Hence, $\{x_{12}\}$ blocks $\{x_{31}\}$ — and by symmetry, $\{x_{31}\}$ blocks $\{x_{23}\}$ and $\{x_{23}\}$ blocks $\{x_{12}\}$. Since no outcome with more than one contract can be individually rational, and $\emptyset$ is blocked by any $\{x_{ij}\}$, we arrive at the standard conclusion that — with choice functions defined as in (1) —*the roommate problem has no stable outcome.*

This prediction seems at odds with the observation that in the real world, people find roommates. As it turns out, it can be attributed to implicit assumptions by the agents that happen to be mistaken.

To see this, first suppose that *all three* agreements are available. Each of the three friends now has access to their favorite roommate agreement, so with the standard approach to choice described in (1), that is precisely the agreement they choose:

$$\hat{C}_1(X_1|X_{-1}) = \{x_{12}\}; \qquad \hat{C}_2(X_2|X_{-2}) = \{x_{23}\}; \qquad \hat{C}_3(X_3|X_{-3}) = \{x_{31}\}. \tag{2}$$

None of these choices coincide: Every agreement is rejected by someone, and so no pair agrees to room together even though each agent would prefer rooming with anyone to remaining alone.

This is because agents implicitly made incorrect assumptions about each other's choices. For instance, when agent 1 chose $\hat{C}_1(X_1|X_{-1}) = \{x_{12}\}$ according to (1), he implicitly assumed that each x_{12} and x_{31} would take effect if he chose it. In particular, he was also assuming that when all agreements were available, x_{31} would be chosen by agent 3 (which turned out to be correct), and x_{12} would be chosen by agent 2 (which turned out to be incorrect). If he had correctly anticipated the other agents' choices, he would have chosen $\{x_{31}\}$ instead.

Part 2. But what if we replaced agents' incorrect assumptions with correct beliefs about others' choices, and required those beliefs to be consistent *across* all sets of available contracts? Then the blocking cycle vanishes, and a stable outcome exists.

For instance, suppose that when all agreements are available, agent 1 believes that, as in (2), x_{31} will be chosen by agent 3, but x_{12} will be rejected by agent 2. Then his optimal choice is $C_1(X_1|X_{-1}) = \{x_{31}\}$. If agent 3 correctly believes that x_{31} will be chosen by agent 1, then since it is his favorite contract, he will optimally choose it as well: $C_3(X_3|X_{-3}) = \{x_{31}\}$. And if agent 2 correctly believes that neither of his friends will choose an agreement with him, it is optimal for him to choose nothing: $C_2(X_2|X_{-2}) = \emptyset$. Hence, the beliefs agent 1 had about the contracts that agents 2 and 3 would choose are correct.

Since they are correct, these beliefs eliminate the myopic behavior observed in Part 1 when all three contracts were available. If they are consistent with beliefs about choices from *other* sets of contracts, they also restore the existence of a stable outcome. Recall that with all three contracts available, none of the agents believed that any of the others would choose x_{23}. If beliefs are consistent across sets of available contracts, making x_{23} unavailable should not change agents' beliefs about the *remaining* contracts $\{x_{12}, x_{31}\}$. Hence, choices should not change either: we should have $C_1(\{x_{12}, x_{31}\}|\emptyset) = C_3(\{x_{31}\}|\{x_{12}\}) = \{x_{31}\}$, and $C_2(\{x_{12}\}|\{x_{31}\}) = \emptyset$, and so $\{x_{31}\}$ blocks $\{x_{12}\}$, rather than the other way around.

These choices break the blocking cycle that ruled out the existence of a stable outcome. In fact, if we continue to construct agents' choice functions in this manner — as optimal choices given correct beliefs that are consistent across sets of available contracts — we arrive at a profile for which $\{x_{31}\}$ is the *unique* stable outcome.[15] Even though agents 1 and 2 would

[15]The choices and beliefs pinned down above (those when all contracts are available and when $\{x_{12}, x_{31}\}$ are available) pin down the stable outcome as $\{x_{31}\}$. They are consistent with multiple profiles of optimal choices and correct beliefs at *other* sets of available contracts, and while each of these profiles has the same stable outcome, they may lead to different comparative statics (e.g., if $\{x_{31}\}$ were removed). One such profile is given by

$$\begin{array}{lll} C_i(\emptyset|\emptyset) = \emptyset; & C_i(\{x_{ij}\}|\emptyset) = \{x_{ij}\}; & C_i(\emptyset|\{x_{jk}\}) = \emptyset, \text{ for each } i \neq j \neq k; \\ C_1(\{x_{31}\}|\{x_{23}\}) = \{x_{31}\}; & C_2(\{x_{23}\}|\{x_{31}\}) = \emptyset; & C_3(\{x_{31}, x_{23}\}|\emptyset) = \{x_{31}\}; \\ C_1(\{x_{12}\}|\{x_{23}\}) = \emptyset; & C_2(\{x_{12}, x_{23}\}|\emptyset) = \{x_{23}\}; & C_3(\{x_{23}\}|\{x_{12}\}) = \{x_{23}\}. \end{array}$$

Other strategically consistent profiles can be found for which the stable outcome is different (e.g., $\{x_{23}\}$ or $\{x_{12}\}$); see Example 2. Intuitively, the players' beliefs about one another's choices in each of these profiles

prefer the outcome to be $\{x_{12}\}$ rather than $\{x_{31}\}$, they fail to coordinate on a block because both correctly believe that the other agent would not follow through with it. ■

In this section, we show that Example 1's conclusion holds more generally. Our main results show that if we account for agents' beliefs about the contracts that other agents will choose, and derive their choice functions given those beliefs, then a unique stable outcome exists in *any* matching environment, whenever beliefs are *correct* (match other agents' actual choices) and *cross-set consistent* (match beliefs at sets from which irrelevant contracts are removed). We call such a profile of choice functions and beliefs *strategically consistent*.

Definition (Strategic Consistency and Nonstrategic Choice). Given agents' payoffs $\{u_i : 2^X \to \mathbb{R}\}_{i \in I}$,

- A profile of choice functions $\{C_i : 2^{X_i} \times 2^{X_{-i}} \to 2^{X_i}\}_{i \in I}$ and *beliefs* $\{\mu_i : 2^X \to 2^X\}_{i \in I}$ is *strategically consistent* if for each $i \in I$,
  - i. μ_i is *correct* given $\{C_j\}_{j \neq i}$: For each $Y \subseteq X$, $\mu_i(Y) = C^{-i}(Y) := \bigcap_{j \neq i}(C_j(Y_j|Y_{-j}) \cup Y_{-j})$.
  - ii. C_i is *optimal* given μ_i: For each $Y \subseteq X$, $C_i(Y_i|Y_{-i}) = \arg\max_{S \subseteq \mu_i(Y)_i} u_i(S \cup \mu_i(Y)_{-i})$.
  - iii. μ_i is *cross-set consistent* given $\{C_i\}_{i \in I}$: For each $Y, Z \subseteq X$, if $Y \supseteq Z \supseteq C_j(Y_j|Y_{-j})$ for all $j \in I$, then $\mu_i(Z) = \mu_i(Y)$.
- Each agent i's *nonstrategic choice function* $\hat{C}_i$ is defined by (1).

Strategic consistency is motivated by two assumptions about agents' epistemic sophistication. First, when faced with any set of contracts that might be proposed, they are able to form correct beliefs about which contracts the other agents will choose. (We model these beliefs as sets of contracts $\mu_i(Y)$ that *no other agent rejects* from Y, since that is what is relevant to an agent's choice.) Second, when contracts that are not chosen by *anyone* are removed, agents do not believe that others would change their behavior (cross-set consistency). (E.g., in Example 1, since no agent chose x_{23} when all contracts were available, we required that none of them would change their choices when we made x_{23} unavailable.) That is, when agents negotiate, those negotiations are independent of irrelevant alternatives. This criterion has bite because strategic consistency requires agents to form beliefs at *each set of available contracts, not just those that are involved in blocks of a potentially stable outcome.*[16]

reflect the players' relative bargaining power. We explore this connection in greater detail in Section 3.3.

[16]Observe that in Example 1, $\{x_{31}\}$ can only be blocked by $\{x_{12}\}$ or $\{x_{23}\}$ *alone* since agent 2 cannot sign both agreements. (The important part here is that agent 2 *would never* choose both agreements; the interpretation in the roommate example just so happens to be that doing so is infeasible.) But cross-set consistency ruled out blocking cycles — thus ensuring a stable outcome — precisely because agents had correct beliefs about each other's choices when all of the contracts were available *together*.

When agents instead assume that each contract they choose will go into effect, we call the resulting choice functions — those derived from preferences using the standard approach — *nonstrategic*.

3.1 Stable Outcomes

Our first main result shows that each strategically consistent profile of choice functions and beliefs pins down a stable outcome. Because agents make correct assumptions about each other's behavior, none of the conditions used to show that stable outcomes exist with nonstrategic choice — e.g., substitutable preferences, no (or well-behaved) externalities, acyclic or two-sided market structure — are necessary to ensure that stable outcomes exist. We say an outcome $Y \subseteq X$ is *stable for* a profile $\{C_i, \mu_i\}_{i \in I}$ if it is stable given choice functions $\{C_i\}_{i \in I}$.

Theorem 1 (Strategic Consistency and Stability). *For each strategically consistent profile of choice functions and beliefs $\{C_i, \mu_i\}_{i \in I}$, there is a unique outcome that is stable for that profile.*

All three parts of strategic consistency play a role in this result. First, correctness and optimality ensure that agents have common beliefs that match each other's choices. Formally:

Lemma 1. *Suppose choice functions $\{C_i\}_{i \in I}$ are optimal given beliefs $\{\mu_i\}_{i \in I}$, and beliefs $\{\mu_i\}_{i \in I}$ are correct given choice functions $\{C_i\}_{i \in I}$. Then for each $i, j \in I$ and $Y \subseteq X$,*

i. Agents' beliefs must coincide: $\mu_i(Y) = \mu_j(Y) := \mu(Y)$.

ii. Agents' choices match common beliefs: $C_i(Y_i | Y_{-i}) = \mu(Y) \cap X_i$.

Intuitively, when choice is optimal given correct beliefs, no agent ever chooses a contract from a set of available contracts that another agent rejects from that set; i.e., unlike in Part 1 of Example 1, contracts must either be rejected by *everyone* they name or rejected by *no one*. Consequently, since all agents' beliefs are correct, they must (i) coincide and (ii) match the choices of *each* individual agent, not just the set of contracts that *none* of them reject (as in the definition of correct beliefs).

Second, since choices match a common belief at *each* set of available contracts, consistency of those beliefs *across* sets of available contracts rules out the kind of blocking cycles that can lead to nonexistence when choice is nonstrategic (as in Part 1 of Example 1). In particular, it ensures that whatever set $\mu(X)$ agents believe others will choose from the set of *all* contracts X, they also believe others will choose $\mu(X)$ from any set $Y \supseteq \mu(X)$ that contains it.

This guarantees that each agent chooses precisely the contracts in $\mu(X)$ that name them when $Y = \mu(X)$ (individual rationality), and chooses no new contracts Z that might be

available alongside it when $Y = \mu(X) \cup Z$ (unblocked). Hence, $\mu(X)$ is stable for the profile $\{C_i, \mu_i\}_{i \in I}$. It also guarantees that $\mu(X)$ is the *unique* stable outcome for that profile: any $S \neq \mu(X)$ either isn't individually rational (if $S \supset \mu(X)$) or is blocked by $\mu(X) \setminus S$ (otherwise).

Thus, Theorem 1 is constructive: once we have found a strategically consistent profile, it is straightforward to find the unique outcome that is stable for that profile, since it coincides with any agent's beliefs $\mu_i(X)$ when all contracts are available. Consequently, with strategic consistency, instead of finding stable outcomes, we can direct our efforts toward finding and characterizing profiles of choice functions and beliefs.

Corollary 1 (Stability and Beliefs). *Given a strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$, Y is the unique stable outcome for that profile of choice functions and beliefs if and only if $\mu_i(X) = Y$ for each $i \in I$.*

The process by which these outcomes are formed is simple. Each agent forms correct and consistent beliefs about the way that other agents deviate both noncooperatively (i.e., by dropping contracts unilaterally) and cooperatively (i.e., by responding to proposed blocks). Then, they each agree to the contracts that are part of the unique set that is robust to these deviations.

3.2 Strategically Consistent Profiles

Theorem 1 shows that each strategically consistent profile of choice functions and beliefs pins down a unique stable outcome. However, the existence of these profiles is not immediate: Strategically consistent profiles are equilibrium objects, in the sense that given a set of contracts, each agent makes optimal choices, given the choices of the others.

Our second main result shows that strategically consistent profiles always exist in any matching environment.

Theorem 2 (Strategically Consistent Profiles: Existence). *Strategically consistent profiles exist.*

Theorem 2 establishes the existence of fixed points in *choice functions* (i.e., strategically consistent profiles) rather than fixed points in *outcomes* (e.g., the outcomes of a deferred acceptance algorithm). To explain it, we describe the algorithm that we introduce to construct strategically consistent profiles of choice functions and beliefs.

We start by considering the outcomes Y that are *nonstrategically individually rational*: $\hat{C}_i(Y_i|Y_{-i}) = Y_i$ for each $i \in I$. These outcomes play an important role in the construction of strategically consistent profiles: they are precisely the sets of contracts that agents can

believe others will choose from an available set of contracts.[17] This fact facilitates a converse to Lemma 1 that powers our construction algorithm.

Lemma 2 (Converse of Lemma 1). *Suppose that $\{C_i, \mu_i\}_{i \in I}$ is a profile of choice functions and beliefs such that beliefs are common across agents, and choices match beliefs: For each $i, j \in I$ and $Y \subseteq X$, (i) $\mu_i(Y) = \mu_j(Y) := \mu(Y)$, and (ii) $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$. Then*

(a) The beliefs $\{\mu_i\}_{i \in I}$ are correct given the choice functions $\{C_i\}_{i \in I}$.

(b) The choice functions $\{C_i\}_{i \in I}$ are optimal given the beliefs $\{\mu_i\}_{i \in I}$ if and only if for each $Y \subseteq X$, $\mu(Y)$ is nonstrategically individually rational.

Our algorithm is initialized by picking some strict total order $\succ$ on the collection of nonstrategically individually rational outcomes. Then, at each each set of available contracts, have each agent choose the contracts in the highest-ranked outcome available, and correctly believe that the other agents will do the same:

$$\underbrace{\mu_i(Y) = \mu(Y) =}_{\text{beliefs are common}} \underbrace{\max_{\succ}\{Y' | Y' \subseteq Y\}}_{\substack{\succ\text{-highest nonstrategically} \\ \text{IR outcome available}}}, \qquad \underbrace{C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i.}_{\text{choices match beliefs}} \tag{3}$$

Intuitively, the order $\succ$ used in this algorithm captures the agents' common assumptions about which outcomes will result from any joint deviation that might be proposed. Because this order pins down beliefs at *every* set of available contracts, these beliefs are cross-set consistent. Since these beliefs are common across agents and match choices, the algorithm always generates a strategically consistent profile of choice functions and beliefs (Lemma 2).

3.3 Pareto Optimality and Characterization

Strategically consistent profiles are equilibrium objects. Some of them may be Pareto-dominated by others. In particular, some profiles may encode coordination failures in which agents choose a Pareto-dominated set of contracts, even though they could choose a Pareto-improving outcome and still satisfy strategic consistency. Example 2 illustrates.

Example 2 (Roommate Problem Revisited). Consider the roommate problem from Example 1 once more. There, we found a strategically consistent profile for which $\{x_{31}\}$ was

[17] Intuitively, when choice functions are optimal given correct beliefs, those beliefs must be common across agents (Lemma 1). Then at any set of available contracts Z, no one can have an incentive to reject contracts that are part of the common belief $\mu(Z)$, given that the other agents choose precisely the contracts in $\mu(Z)$. In other words, $\mu(Z)$ must be nonstrategically individually rational.

stable: $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$ and $\mu_i(Y) = \mu(Y)$, for each $i \in \{1,2,3\}$ and

$$\mu(\emptyset) = \emptyset; \qquad \mu(\{x_{ij}\}) = \{x_{ij}\} \text{ for each } i,j;$$
$$\mu(\{x_{31}, x_{23}\}) = \mu(\{x_{12}, x_{31}\}) = \{x_{31}\}; \qquad \mu(\{x_{12}, x_{23}\}) = \{x_{23}\}.$$

By symmetry, profiles also exist for which $\{x_{12}\}$ and $\{x_{23}\}$ are stable. In these profiles, people find roommates, in line with the experience of most undergraduate students.

But there is also a strategically consistent profile for which the "autarky" outcome $\emptyset$ is stable: $\{C_i^0, \mu_i^0\}_{i \in I}$, where $C_i^0(Y_i|Y_{-i}) = \mu_i^0(Y) = \emptyset$ for all $Y \subseteq X$ and $i \in \{1,2,3\}$. This seems less plausible: Agents fail to coordinate on choosing a roommate agreement x_{ij} when one is made available alongside $\emptyset$, even though this would be a Pareto improvement.

In contrast, no such coordination failures exist in $\{C_i, \mu_i\}_{i \in I}$ (or the symmetric profiles for which $\{x_{12}\}$ and $\{x_{23}\}$ are stable): At any set of available contracts, there is no outcome that represents a Pareto improvement upon the outcome that the agents believe the others will choose.[18] This suggests a focus on the latter profiles. ■

In this section, we first introduce a criterion on strategically consistent profiles ensuring that agents' beliefs select Pareto-undominated outcomes whenever possible. We then give a "welfare theorem" characterizing the stable outcomes predicted by these profiles. As it turns out, this also makes precise the connection between strategically consistent profiles and Nash bargaining weights.

Recall that when choices are optimal, correct beliefs always select nonstrategically individually rational subsets from each set of available contracts. Our main efficiency criterion requires that they never select one of these subsets when it is Pareto-dominated by another.

Definition (Pareto Optimality). We say that a strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$ satisfies *Pareto optimality* if for any nonstrategically individually rational $Y, Z \subseteq X$ such that $u_i(Y) \geq u_i(Z)$ for all $i \in I$ and $u_i(Y) > u_i(Z)$ for some $i \in I$, we have $\mu_i(Y \cup Z) \neq Z$ for each $i \in I$.

Theorem 3 shows that strategically consistent profiles that satisfy Pareto optimality are easy to find, simply by using our algorithm (3) with an order $\succ$ that is structured so that the agents' common beliefs solve a social planner's problem. Formally, we say that a strict total order $\succ^\phi$ on the nonstrategically individually rational outcomes $\mathcal{M}$ is *induced by* $\phi : \mathbb{R}_+^I \to \mathbb{R}$ if $\phi((u_i(Y))_{i \in I}) > \phi((u_i(Z))_{i \in I})$ implies $Y \succ^\phi Z$. As Lemma 8 in the Appendix shows, each

[18] That said, there is a coordination failure between agents 1 and 2 that makes them unwilling to choose the contract they prefer from $\{x_{31}, x_{12}\}$ in this profile. This is rationalized by a simple story: Agent 1 (correctly) believes that if he breaks his agreement with agent 3 to room with agent 2, agent 2 will then leave to form an agreement with the newly roommateless agent 3. We explore the consequences of requiring profiles to be rationalized by such *forward induction* reasoning in the Online Appendix.

increasing ϕ induces an order $\succ^\phi$. We can interpret ϕ (and the $\succ^\phi$ it induces) as describing the way that agents base their beliefs about other agents' choices on the payoffs all agents will receive from those choices.

Theorem 3 (Pareto-Optimal Profiles). *Let $\phi : \mathbb{R}^I_+ \to \mathbb{R}$ be a strictly increasing function.*

i. For any strict total order $\succ^\phi$ induced by ϕ, the profile $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$ constructed from $\succ^\phi$ using the algorithm (3) is strategically consistent and satisfies Pareto optimality.

ii. The common belief μ^ϕ in $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$ is the solution to a social planner's problem: For all $i \in I$ and $Z \subseteq X$,

$$\mu_i^\phi(Z) \in \arg\max_{S\subseteq Z} \phi((u_i(S))_{i\in I}) \text{ s.t. } \hat{C}_j(S_j|S_{-j}) = S_j \ \forall j \in I. \tag{4}$$

The intuition for Theorem 3 is straightforward. Since ϕ is strictly increasing, $\succ^\phi$ ranks Y ahead of Z whenever Y is a Pareto improvement on Z. Hence, the beliefs constructed by the algorithm never select a Pareto-inferior outcome when a Pareto-superior one is available (i). In fact, the construction of $\succ^\phi$ ensures that they solve the social planner's problem (4) (and do so uniquely if there are no ties) (ii).

Theorem 3 shows that beliefs are part of a strategically consistent profile if, at any set of available agreements, they maximize a social welfare function subject to the constraint that no agent can profit by vetoing contracts. This provides one direction of a "welfare theorem" for strategic consistency (Theorem 4): outcomes that are stable for some strategically consistent profile satisfying Pareto optimality are those on a constrained Pareto frontier.

Theorem 4 (Welfare Theorem for Strategic Consistency). *There is a strategically consistent profile satisfying Pareto optimality for which $Y \subseteq X$ is stable if and only if Y is Pareto efficient among the nonstrategically individually rational outcomes.*

Theorem 4 parallels welfare theorem-like results in the matching literature with *transferable utility* (e.g., with substitutability, Hatfield et al. (2013, Theorems 2-6); with complementarity, Rostek and Yoder (2020, Proposition 3 and Theorem 2)). Rather than decentralizing an efficient outcome through the use of *competitive equilibrium prices*, Theorem 4 shows that such outcomes can be decentralized by a correct and consistent profile of *beliefs*. This allows it to apply even to nontransferable utility settings, and without the conditions on preferences or market structure that are, in general, necessary for competitive equilibrium prices to exist.[19] Example 3 illustrates in a many-to-one matching model with both complementarity and

[19]In particular, the "if" part does not follow immediately from the separating hyperplane theorem and Theorem 3, since the utility possibility set is finite, rather than convex.

substitutability, where the standard, nonstrategic approach to choice does not yield a stable outcome.

Example 3 (Labor Markets with Complementarities). Consider a labor market with two workers, Alice and Bob, and two firms, 1 and 2: $I = \{a, b, 1, 2\}$. Employment contracts are standardized, i.e., each is completely characterized by the worker-firm pair it involves: $X = \{x_{a1}, x_{a2}, x_{b1}, x_{b2}\}$ and $N(x_{ij}) = \{i, j\}$ for each $i \in \{a, b\}$ and $j \in \{1, 2\}$.

Workers can only work for one firm ($u_i(\{x_{i1}, x_{i2}\}) < u_i(\emptyset)$ for each $i \in \{a, b\}$), and there are no externalities. Both workers prefer any employment to unemployment, but Alice prefers firm 1, while Bob prefers firm 2: $u_a(\{x_{a1}\}) > u_a(\{x_{a2}\})$ and $u_b(\{x_{b2}\}) > u_b(\{x_{b1}\})$. Firm 2 wants to hire one worker ($u_2(\{x_{a2}, x_{b2}\}) < u_2(\emptyset)$), and would prefer it to be Alice: $u_2(\{x_{a2}\}) > u_2(\{x_{b2}\})$. Firm 1 could hire both workers, but is only willing to hire Alice if it also hires Bob: $u_1(\{x_{a1}, x_{b1}\}) > u_1(\{x_{b1}\}) > u_1(\emptyset) > u_1(\{x_{a1}\})$.

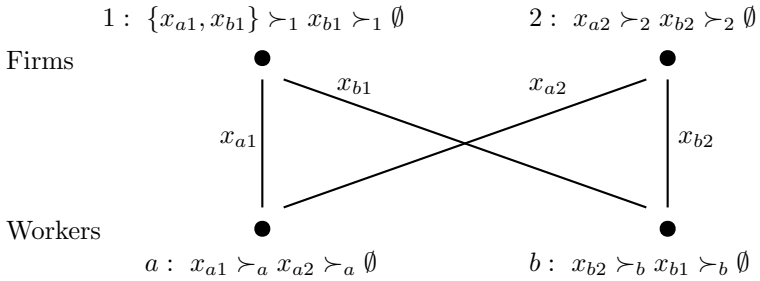

Figure 2: A visual description of the environment in Example 3.

Like many matching environments with both complementarity (between x_{a1} and x_{b1}) and substitutability (among other contracts), this example has no stable outcome when choice is nonstrategic.[20] But when choice is strategically consistent, and agents can overcome coordination failures that result in Pareto-dominated outcomes, Theorem 4 shows that three outcomes can be stable: $\{x_{b2}\}$, $\{x_{a2}, x_{b1}\}$, and $\{x_{a1}, x_{b1}\}$. Each is nonstrategically individually rational, and (unlike $\emptyset$, $\{x_{a2}\}$, and $\{x_{b1}\}$) is not Pareto-dominated by another nonstrategically individually rational outcome.[21]

We describe one such profile, for which $\{x_{a2}, x_{b1}\}$ is stable. To avoid assigning cardinal values to agents' payoffs, we start with a strict total order that never ranks an outcome below another that it Pareto dominates: e.g.,

$$\{x_{a2}, x_{b1}\} \succ \{x_{b2}\} \succ \{x_{a1}, x_{b1}\} \succ \{x_{b1}\} \succ \{x_{a2}\} \succ \emptyset. \tag{5}$$

[20] With nonstrategic choice, each individually rational outcome is blocked: $\emptyset$ is blocked, e.g., by x_{a2}; $\{x_{a2}\}$ is blocked by, e.g., $\{x_{b1}\}$; $\{x_{a2}, x_{b1}\}$ is blocked by $\{x_{a1}, x_{b1}\}$; $\{x_{a1}, x_{b1}\}$ is blocked by $\{x_{b2}\}$; $\{x_{b2}\}$ is blocked by $\{x_{a2}\}$; and $\{x_{a2}\}$ is blocked by $\{x_{a1}, x_{b1}\}$.

[21] In fact, since this setting has no externalities, Theorem 3 shows that each can be decentralized by beliefs that not only avoid coordination failure (i.e., satisfy Pareto optimality), but are robust to forward induction reasoning of the sort considered in the Online Appendix.

Given this order, the algorithm in (3) constructs a strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$ satisfying Pareto optimality and forward induction for which $\{x_{a2}, x_{b1}\}$ is stable.[22] ■

Pareto Optimality and Bargaining Weights

Using Theorem 3 to construct a profile $\{C_i^\phi, \mu_i^\phi\}_{i \in I}$ requires us to first pick a social welfare function ϕ. We may wish to do so in a way that ensures that the order $\succ^\phi$ is not sensitive to specifics of the agents' utility functions that do not affect their incentives. In particular, we might desire the solution to (4) to be invariant under rescaling of the agents' utility functions. As is well known, this pins down the social welfare function in Theorem 3 as the familiar asymmetric Nash product. Formally, we say that $\phi : \mathbb{R}_+^I \to \mathbb{R}$ is *scale invariant* if for any $a, x, y \in \mathbb{R}_+^I$, $\phi(x) > \phi(y) \Leftrightarrow \phi((a_i x_i)_{i \in I}) > \phi((a_i y_i)_{i \in I})$.

Lemma 3. *If $\phi : \mathbb{R}_+^I \to \mathbb{R}$ is continuous, strictly increasing, and scale invariant, then there is some $\alpha \in \Delta(I)$ such that $\phi(x) \geq \phi(y) \Leftrightarrow \prod_{i \in I} x_i^{\alpha_i} \geq \prod_{i \in I} y_i^{\alpha_i}$.*

We emphasize that even when a strategically consistent profile is pinned down by maximizing a Nash product (as we do in Section 5 and Section S.1 in the Online Appendix), the interpretation is *not* that the outcome is determined through a multilateral Nash bargaining among all agents over all contracts. Instead, the outcome is determined by the absence of deviations by *groups* of agents to form new contracts with one another, given beliefs that are refined by Pareto optimality. We discuss this connection in more detail in Section 5.

By varying the social welfare function ϕ — e.g., by changing the bargaining weights α in a Nash product — we can construct a profile that is relatively more favorable to some agents, and less favorable to others. This highlights the fact that strategically consistent profiles play the same role in matching-theoretic stability that bargaining weights do in other models of bargaining (e.g., Nash-in-Nash (Collard-Wexler et al. (2019)) bargaining): There are several possible profiles of weights/choice functions and beliefs, and given any such profile, the bargaining solution/stability pins down a unique prediction. Theorem 3 and Lemma 3 formalize this connection in the case of profiles that satisfy Pareto optimality; in Section S.1 of the Online Appendix, we illustrate the connection in the context of Example 3. In Section 5, we show that this facilitates an approach to counterfactual analysis similar to those used with these other models (e.g., Ho and Lee (2017, 2019)).[23]

[22] Specifically, this profile $\{C_i, \mu_i\}_{i \in I}$ is given by $C_i(Y_i | Y_{-i}) = \mu(Y) \cup X_i$ and $\mu_i(Y) = \mu(Y)$, for

$$\begin{array}{cc} \mu(\{x_{a1}, x_{a2}\}) = \mu(\{x_{a2}\}) = \{x_{a2}\}; & \mu(Y) = \{x_{a2}, x_{b1}\} \text{ if } \{x_{a2}, x_{b1}\} \subseteq Y; \\ \mu(\{x_{b1}\}) = \{x_{b1}\}; & \mu(\emptyset) = \mu(\{x_{a1}\}) = \emptyset; \\ \mu(\{x_{a1}, x_{b1}\}) = \{x_{a1}, x_{b1}\}; & \mu(\{x_{a1}, x_{b1}, x_{b2}\}) = \mu(\{x_{a1}, x_{a2}, x_{b2}\}) = \mu(\{x_{b1}, x_{b2}\}) \\ & = \mu(\{x_{a2}, x_{b2}\}) = \mu(\{x_{a1}, x_{b2}\}) = \mu(\{x_{b2}\}) = \{x_{b2}\}. \end{array}$$

[23] Specifically, we can first use the data to identify the agents' preferences, the observed outcome Y^*, and

3.4 Discussion

Existence

Our existence results allow the application of matching-theoretic tools in new environments where agents' preferences may feature both substitutability and complementarity, agreements may have externalities, and/or more than two agents may be involved in the same contract. In particular, they employ the usual matching-theoretic approach, using a standard cooperative solution concept (stability) to pin down *outcomes* given *choice functions*. The key innovation that allows us to guarantee the existence of stable outcomes is to use *noncooperative* reasoning to determine these choice functions, thus ensuring that they are based on correct and consistent beliefs. This noncooperative reasoning cannot make predictions about outcomes on its own, because it only describes the way agents would choose from any set of contracts that might be under negotiation. Instead, applying cooperative reasoning (in the form of stability) to these choice functions makes it possible to predict the actual outcomes.

Profiles vs. Outcomes

The object pinned down by strategic consistency is not an *outcome*, but rather a profile of *choice functions*, each of which implies a unique stable outcome (Theorem 1). In particular, our results do not simply require the *set of contracts that an agent agrees to sign* to be a best response to those agreed to by others *in the stable outcome*. Rather, strategic consistency requires agents' *choice functions* to specify sets of contracts that are best responses to others' behavior *at each possible set of available contracts;* that is, at *every* set of available contracts (i.e., every set that might be discussed in a negotiation), the agents' choices must form a Nash equilibrium of a game where each agent chooses a set of contracts, and contracts go into effect if they are chosen by each agent they name.[24] Thus, a profile of strategically consistent choice functions can be interpreted as collections of equilibria of a contract-announcement game. (We formalize this connection in Section S.5 of the Online Appendix.) Stability then *selects* from among outcomes of this game by pinning down an outcome that is robust to *both individual and joint deviations* given those equilibrium choice functions. In other words, rather than being robust only to noncooperative deviations for each set of contracts, outcomes that are stable given strategically consistent behavior are robust to cooperative deviations across

the ϕ (or, when ϕ is pinned down by scale invariance (Lemma 3), a vector of Nash weights α) such that Y^* is stable for $\{C_i^\phi, \mu_i^\phi\}_{i \in I}$. Then, modify the environment by applying the treatment of interest (e.g., a merger or regulation). Finally, derive the profile associated with ϕ in the modified environment, and compute its unique stable outcome.

[24] This *contract-announcement game* generalizes the link-announcement game discussed in, e.g., Myerson (1991) and Jackson (2010). Strategically consistent profiles always exist precisely because this game always has a Nash equilibrium in pure strategies.

sets of contracts that agents evaluate strategically (i.e., given correct beliefs, rather than nonstrategically).

Beliefs and Perfection/Trembles

The beliefs involved in a strategically consistent profile might appear "brittle", in the sense that *at a given set of available contracts*, the equilibrium is not robust to trembles. In Part 2 of Example 1, for instance, it is optimal for agent 2 to choose the agreement x_{12} with agent 1 if he thinks that there is a small probability that agent 1 will also choose it. And if agent 2 does choose x_{12}, then it is optimal for agent 1 to choose it as well. If we then consider trembles among the other agents, we are left back where we started with nonstrategic choice and nonexistence.

But with strategic consistency, beliefs are a *profile*: They need not be formed in isolation at every set of available contracts the way that the "tremble" thought experiment requires. Instead, the agents can reason that these "trembling" beliefs lead to inconsistency across sets of available contracts, and rule them out.

Strategic vs. Nonstrategic Approach to Choice

Our approach is to work with the standard stability concept — just applied to choice functions that are *determined by a fixed point relationship with agents' beliefs about others' choices* rather than *pinned down as solutions to single-agent decision problems.* The key insight of this paper is that this allows one to make predictions in any environment, regardless of agents' preferences, externalities, or the structure of the market. However, one might also wonder whether there is anything to be gained by taking the strategically consistent approach in settings where the standard, nonstrategic approach can predict stable outcomes. The answer is yes.

First, strategic consistency allows us to attribute the multiplicity of stable outcomes to a multiplicity of beliefs that agents may have about the way others will react to blocking proposals.[25] Each of these profiles of beliefs can be interpreted as a description of the way that bargaining power is distributed among the agents in equilibrium. In fact, when we rule out coordination failure using Pareto optimality, we can regard Nash bargaining weights as sufficient statistics for the bargaining power described by beliefs, as we do in Example 5.

Second, even when stable outcomes do exist with the standard approach, strategic consistency can capture plausible outcomes that are ruled out by nonstrategic choice, a point that we illustrate with Example S.1 in the Online Appendix.[26]

[25] For an alternative interpretation of this multiplicity based on procedural fairness (the order of proposals in a modified deferred acceptance algorithm), see Dworczak (2021).

[26] The possibility of compelling outcomes that are ruled out by stability has been noted before. In particular,

Third, strategic consistency allows us to make new predictions about counterfactual outcomes — a point we explore in the next section.

Stability and Bargaining Theory

Theorem 3 can be thought of as a microfoundation for agents' beliefs by appealing to bargaining theory. If we parameterize agents' bargaining power, as in, e.g., Nash (1950), agents' beliefs, and thus (by Theorem 1) the stable outcome, can be *uniquely* pinned down; see Example 5 and Section S.1 in the Online Appendix. (Moreover, we can identify that outcome without identifying the full profile of choice functions and beliefs.) Alternatively, with Theorem 3, we can think of strategic consistency as a (cooperative) microfoundation for a type of Nash bargaining solution in which outside options are endogenous (since the payoffs that an outcome must provide the agents in order to be individually rational depends on the payoffs they can get by dropping the agreements that they make).

Other Refinements

Our main characterization result (Theorem 4) describes the set of outcomes that are pinned down by strategically consistent profiles that satisfy Pareto optimality. But one can consider alternative restrictions on beliefs that may be attractive in different applications. In the Online Appendix, we consider two of them: evaluating the credibility of deviations by employing forward induction reasoning (Section S.3.1) and requiring robustness to nonstrategic deviations (Section S.3.2). Like Pareto optimality, each places additional structure on the order $\succ$ used to run the algorithm (3) and pin down the agents' (correct and cross-set consistent) beliefs. This additional structure captures the class of blocking deviations that the refinement requires agents to believe are plausible.

4 Applications

Here, we highlight how we can use our results in three applications that have not been traditionally studied using matching-theoretic stability: network formation, environments with externalities from downstream competition, and legislative bargaining.

a variation of stability, *weak setwise stability* (Klaus and Walzl (2009)) does not consider blocks where the participating agents' nonstrategic choices do not coincide (as in Example S.1). Formally, a blocking proposal Z is a weak setwise block of Y if each agent who participates in the block (nonstrategically) chooses the same set of contracts: for all $i \in N(Z \setminus Y)$, $\hat{C}_i(Z_i \cup Y_i | Z_{-i} \cup Y_{-i}) = Z_i$. (Such a proposal corresponds to a block $Z' = Z \setminus Y$.) Strategic consistency instead considers all blocks, but rules out disagreements about the outcome of a block by allowing agents' choices to be endogenously determined given beliefs that are correct.

Network Formation

By modeling links between agents as bilateral contracts, our results can be applied to network formation settings where link formation has externalities on other agents, such as free trade agreement formation (e.g., Furusawa and Konishi (2007)) and joint venture formation among oligopolists (e.g., Goyal and Joshi (2003)). In these settings, arguably the most common solution concept used in the literature is *pairwise stability* (Jackson and Wolinsky (1996)), which selects networks that are robust to deviations by a pair of agents that add a link between them, and deviations by an individual agent that remove one of his links. This is a subset of the changes to the network considered by matching-theoretic stability, which also includes those in which agents substitute between links as well as those where agents add multiple links at the same time. However, pairwise stability is able to make predictions under much more general conditions (e.g., Calvó-Armengol and İlkılıç (2009)) than those known to ensure the existence of matching-theoretically stable outcomes.

But in Theorems 2-4, we show that we can ensure robustness of sets of agreements, including graphs or coexisting coalitions, to the full class of deviations considered by matching-theoretic stability, even without conditions on preferences or externalities. Instead of ruling out some of these deviations exogenously, strategic consistency endogenously determines which of them matter. As we illustrate in Example 4, this also allows us to make predictions in canonical network formation environments in which no pairwise stable outcome exists.

Example 4 (Trading on a Network (Jackson and Watts, 2002)). Consider the following environment from Jackson and Watts (2002). There are two divisible goods, x and y, and N consumers with identical symmetric Cobb-Douglas preferences over those goods. Before they learn their endowments, the consumers form links with one another; each link costs $c > 0$ for the two consumers that it connects. Once the network is formed, they each independently receive endowments $(1, 0)$ and $(0, 1)$ with equal probability, and trade them in separate competitive markets on each connected component of the network. Hence, adding a link benefits an agent by reducing the probability that he faces unfavorable terms of trade because most of the other agents on his connected component have the same endowment as he does.[27]

Jackson and Watts (2002) show that when $c = 5/96$ and $N \geq 4$, no pairwise stable outcome (and hence, when choice is nonstrategic, no stable outcome) exists.[28] If no agent can benefit by dropping a link, each connected component must be a tree: severing a loop

[27] It also reduces the probability that he faces favorable terms of trade because most agents on his connected component have the opposite endowment, but since preferences are convex, this has a smaller effect on his payoffs.

[28] As discussed in Section 4, pairwise stability is weaker than matching-theoretic stability with nonstrategic choice.

does not change the equilibrium on the component, but saves c for the agents who sever their link. Moreover, if an agent has more than one link, each must be to agents that also have multiple links: otherwise, the benefit of keeping the link is lower than 5/96. Thus, a network does not provide agents with incentives to sever links — and hence, in the language of our paper, is nonstrategically individually rational — precisely when it is a collection of connected pairs. But no such network can be pairwise stable: each agent would gain more than 5/96 by forming a link with another connected pair.

Theorem 2 shows that strategic consistency resolves this nonexistence problem by requiring each agent's beliefs about how others would respond to linking proposals to be correct and consistent across different sets of available links. Moreover, Theorem 4 shows that when we focus on Pareto-optimal profiles, strategic consistency predicts exactly those outcomes we would intuitively expect to see in this setting: networks consisting of connected pairs, with no more than one isolated agent. ■

Externalities from Downstream Competition

A large literature in industrial organization analyzes markets in which firms first form agreements with suppliers, and then compete in an imperfectly competitive downstream market. Examples include the formation of agreements between insurers and healthcare providers (e.g., Ho and Lee (2017)); between television networks and distributors (e.g., Crawford and Yurukoglu (2012)); and between medical device manufacturers and hospitals (e.g., Grennan (2013)). Because they affect competition in the downstream market, these supplier relationships have externalities: If one insurer/TV distributor/hospital agrees to a contract with a provider/TV network/device manufacturer, it affects the incentives of other firms to form their own contracts. Clearly, these settings can be embedded into a general matching with contracts model like the one we study.

But the presence of externalities has made it challenging to apply matching theory to analyze the formation of these agreements. Instead, a popular solution concept used in these settings is *Nash-in-Nash bargaining* (Horn and Wolinsky (1988)): a Nash equilibrium in Nash bargains. Since it pins down the division of surplus between firms, and does not require strong conditions on preferences for existence, this framework has proven extremely useful in empirical work.

Like pairwise stability in the network formation context, the Nash-in-Nash solution can offer predictions in settings where the standard approach to stability does not because it considers a less expansive set of changes to the set of agreements.[29] This has motivated recent

[29]In particular, Nash-in-Nash does not consider, e.g., deviations to substitute between agreements with different other agents, add or remove multiple agreements with different agents at the same time, or remove some agreements with a counterparty while keeping others; matching-theoretic stability considers all of these.

papers such as Ho and Lee (2019) and Liebman (2018) to extend the Nash-in-Nash concept to settings where other types of deviations, such as those to exclude a healthcare provider from an insurer's network, are important. While our results do not explicitly extend the Nash-in-Nash solution, they contribute to this literature by showing how we can consider robustness to all joint deviations by using matching-theoretic stability, even in settings where externalities are prevalent and preferences do not satisfy substitutability or complementarity conditions. Moreover, our analysis in Section 5 suggests techniques for using matching-theoretic stability, along with strategic consistency, to make counterfactual predictions in a similar way to Nash-in-Nash.

Legislative Bargaining

A rich political economy literature considers settings where legislators bargain multilaterally over which of several policies to enact. Following Baron and Ferejohn (1989), this literature generally takes a noncooperative approach, modeling a negotiation as a dynamic game where legislators take turns proposing a division of surplus which is then subjected to a majoritarian vote. As Ali et al. (2019) show, the outcome of this multilateral bargaining protocol is sensitive to its extensive form, which can dramatically change the division of surplus predicted by the Baron and Ferejohn (1989) model.

Our results allow for predictions in environments where agents form multiple multilateral agreements (such as legislative bargaining) without relying on the specifics of the bargaining process. Specifically, suppose that we let contracts represent possible agreements to pass bills among (for concreteness) a majority of legislators. Then our main characterization theorem, Theorem 4, allows us to use matching-theoretic stability to predict which of those agreements will form. Moreover, the counterfactual analysis that we discuss in Section 5 allows us to predict how an observed outcome will change due to changes in, e.g., legislative procedure or the political environment.

4.1 New Comparative Statics with Strategic Consistency

Strategic consistency allows new comparative statics relative to the standard nonstrategic approach. In general, we can consider two types of changes to the environment.

Removal of Contracts

Suppose that one or more contracts is rendered unavailable (e.g., through regulation). This comparative static is not always feasible with the standard, nonstrategic approach to choice: We can compute a new set of stable outcomes after contracts are removed from the environment, but this set may be empty, since the existence of a stable outcome is not

guaranteed. Even when it is not, we do not know which of the new stable outcomes will result, because that depends on the sequence of deviations that follows. But with a strategically consistent profile, there is always a new stable outcome, and it is independent of the path taken to get to it: Since agents' beliefs are cross-set consistent, any sequence of deviations following the contract's removal must lead to the same outcome. In particular, we can simply restrict the domain of beliefs and choice functions to sets that do not include the removed contracts, and recover the new outcome from the agents' common belief when all remaining contracts are available (as in Corollary 1).

More General Changes

Strategic consistency *always* pins down the way that beliefs (and hence outcomes) change when one or more contracts is *removed* from the environment. But as we illustrate in Example 5, if we have a map between profiles (and hence outcomes) in different environments — such as the one that we describe in Section 5 — we can pin down the impact on beliefs and outcomes of more general changes to the environment: changes to the set of contracts X (e.g., replacing intermediated trades with direct ones), changes to the set of agents $N(x)$ named by a contract (e.g., giving a regulator the ability to veto it), or changes to the agents' payoff functions (e.g., through common ownership of two firms, the imposition of a tax, or the introduction of other types of externalities).

5 Counterfactual Analysis with Strategic Consistency

A great deal of recent empirical work explores settings where agents form agreements with one another that have externalities. In industrial organization, researchers have studied agreements between cable television producers and distributors (Crawford and Yurukoglu (2012)) or insurers and healthcare providers (Ho and Lee (2017)); in international economics, trade agreements between countries (Bagwell et al. (2021)). A central focus in this literature is *counterfactual analysis* describing how outcomes would change in response to some exogenous change in the setting. For instance, Ho and Lee (2017) estimate the effects of the removal of an insurer from the healthcare market, while Bagwell et al. (2021) estimate the effects of a change in GATT/WTO rules.

As discussed in Section 4, the most commonly used solution concept in these applications is *Nash-in-Nash* (Horn and Wolinsky (1988)) — a Nash equilibrium in Nash bargains. While it considers a narrower class of deviations than matching-theoretic stability, Nash-in-Nash has a key advantage that has helped make it popular: It avoids the nonexistence issues faced by stability with nonstrategic choice in applications with externalities. The previous sections of

our paper show how strategic consistency allows matching-theoretic stability to overcome these issues, without limiting the ways that agents can renegotiate agreements. This section shows how it can be used in counterfactual analysis analogously to Nash-in-Nash: if we observe an outcome, we can use the matching-theoretic tools that we introduce to make *unique* predictions about how the outcome would change under some counterfactual scenario.

Counterfactual Analysis with Pareto-Optimal Profiles

Nash-in-Nash counterfactual analysis exploits a degree of freedom in the solution concept's predictions: many outcomes are predicted by Nash-in-Nash for some bargaining parameters, but any given vector of bargaining parameters produces a unique Nash-in-Nash outcome. Thus, observed outcomes can be used to recover the object (a vector of bargaining parameters) that selects them, and that object can be used to select a new outcome after an exogenous change. As we discussed in the introduction, strategic consistency creates a similar degree of freedom: many outcomes are stable for some strategically consistent profile, yet any given profile predicts a unique stable outcome. This suggests exploiting it in the same way, by recovering the strategically consistent profile from data and then using it to pin down a new outcome after an exogenous change in the model.

This procedure for making counterfactual predictions may appear more challenging than the one used with the Nash-in-Nash concept. First, although the observed outcome allows us to recover the agents' choices and beliefs at any larger set of agreements, it does not necessarily pin down the full strategically consistent profile. Second, unlike the set of vectors of bargaining parameters, the set of strategically consistent profiles depends on — and thus is affected by exogenous changes to — the environment. Thus, we need a map from the set of profiles recovered in the first step to the set of profiles that are strategically consistent after an exogenous change occurs.

These difficulties can be overcome by considering strategically consistent profiles that satisfy Pareto optimality. Recall from Theorem 3 that such profiles can be constructed by solving a social planner's problem (4). Each such profile $\{C_i^\phi, \mu_i^\phi\}_{i \in I}$ can be identified with the welfare function ϕ used to construct it, which we can interpret as describing the way that agents base their beliefs about other agents' choices on the payoffs the agents will receive from those choices. Then, using the same ϕ, we can find *corresponding* beliefs and choices that are strategically consistent *after* an exogenous change in the environment.

Recovering the information about the strategically consistent profile necessary to make counterfactual predictions thus amounts to recovering the welfare function ϕ — or, when ϕ is pinned down by scale invariance (Lemma 3), a vector of Nash weights α. Specifically, we can first use the data to identify the agents' preferences, the observed outcome Y^*, and the ϕ such

that Y^* is stable for $\{C_i^\phi, \mu_i^\phi\}_{i \in I}$. Then, modify the environment by applying the treatment of interest (e.g., a merger or regulation). Finally, derive the profile associated with ϕ in the modified environment, and compute its unique stable outcome. Example 5 illustrates this procedure.

Example 5 (Counterfactual Analysis (Grennan, 2013)). Suppose we want to apply matching-theoretic stability to a model of contracting between medical device manufacturers $m \in M$ and hospitals $h \in H$ à la Grennan (2013). Each contract is between a manufacturer and hospital, and specifies a price τ_{mh} per patient who is prescribed the manufacturer's device (here, a stent); no manufacturer-hospital pair can sign more than one contract at a time.[30] Given the set of contracts, each hospital's doctors make prescribing decisions that depend on the prices negotiated for each type of stent; these decisions pin down the manufacturers' and hospitals' payoffs from each set of contracts.

Because the quantity of stents that a manufacturer sells to a given hospital depends on the hospital's contracts with *other* manufacturers, the manufacturers' preferences exhibit externalities. Consequently, matching-theoretic stability could be difficult to apply with the standard, nonstrategic approach to choice, because of existence issues that do not appear in the Nash-in-Nash solution applied in Grennan (2013). Strategic consistency, on the other hand, guarantees that we can apply matching-theoretic stability to this environment (Theorems 1 and 2). Applying Pareto optimality as a refinement gives us tools to make counterfactual predictions the way that Nash-in-Nash does, while capturing robustness to the full set of deviations considered by matching-theoretic stability.[31]

As in Grennan (2013), we could estimate doctors' demand and surplus as a function of negotiated prices τ^*. We could then attempt to jointly estimate the firms' hospital-specific marginal costs c_{mh} and the weights α that pin down ϕ using a procedure similar to the Nash-in-Nash estimation procedure used by Grennan (2013), but using a single, multiplayer asymmetric Nash product instead of *separate* Nash products between each manufacturer-hospital pair. Specifically, if we relied on Theorem 3 and scale invariance (Lemma 3), we would use necessary conditions for the following optimization problem (letting $\tau_A := \{\tau_{mh}\}_{(m,h) \in A}$

[30] We can fit our assumptions that the set of contracts is finite and agents are never indifferent between them by suitably discretizing the space of prices.

[31] That is, along with the deviations considered by Nash-in-Nash, matching-theoretic stability accommodates deviations to substitute between agreements with different other agents, add or remove multiple agreements with different agents at the same time, or remove some agreements with a counterparty while keeping others.

denote a set of pricing agreements between manufacturer-hospital pairs in $A \subseteq M \times H$):

$$(A^*, \tau^*) = \arg\max_{A,\tau_A} \prod_{h \in H} u_h(\tau_A)^{\alpha_h} \prod_{m \in M} u_m(\tau_A)^{\alpha_m} \tag{6}$$

$$\text{s.t.} \begin{array}{l} \tau_{mh} \geq c_{mh} \forall (m,h) \in A, \\ u_m(\tau_A) \geq u_m(\tau_A \setminus Y) \forall m \in M, Y \subseteq \{\tau_{hm}\}_{h \in H}; \end{array} \quad \text{(nonstrategic IR)}$$

instead of the first-order conditions for the pairwise problems

$$\tau^*_{mh} = \arg\max_{\tau_{mh}} u_m(\{\tau_{mh}, \{\tau^*_{m'h'}\}_{m',h' \neq m,h}\})^{\alpha_{mh}} (u_m(\{\tau_{mh}, \{\tau^*_{m'h'}\}_{m',h' \neq m,h}\}) - u_m(\tau^* \setminus \{\tau^*_{mh}\}))^{1-\alpha_{mh}}$$

used in the Nash-in-Nash estimation procedure.[32] We can then consider counterfactual analyses, such as (as in Grennan (2013)) a hospital merger or uniform price regulation, by modifying the environment and recomputing the solution to (6) to find the unique stable outcome. ■

We want to emphasize that, even though the planner's problem (6) constructs a stable outcome (and more generally, a strategically consistent profile) as the solution to a multilateral Nash bargaining problem, its interpretation is *not* that all agents bargain with one another á la Nash (1950) to *cooperatively* determine the outcome they will choose from each set of available contracts. Rather, at each set of available contracts, agents' choices are *individually* optimal given their beliefs about others' choices, and those beliefs are refined by some of the same axioms that characterize the Nash (1950) solution (Pareto optimality (Theorem 3) and scale invariance (Lemma 3)).[33] Given that profile of choice functions and beliefs, the stable outcome is then determined cooperatively by the absence of both individual and joint deviations. Hence, we can think of (6) as describing a "stability-in-Nash" procedure: a stable outcome given optimal choices from beliefs that maximize a Nash product at each set of available contracts.[34]

Since this procedure is based on matching-theoretic stability, it endogenizes which contracts will be signed. In particular, it complements the type of analysis allowed by Nash-in-Nash bargaining by endogenizing exclusion: for instance, the planner's problem (6) naturally accommodates outcomes where some manufacturer-hospital pairs do not contract (and so $A^* \neq M \times H$).

[32] Doing so would require additional steps that we do not investigate here; our example just attempts to illustrate how a procedure based on matching-theoretic stability and strategic consistency could, in principle, be used in a canonical application.

[33] Though we do not explicitly invoke independence of irrelevant alternatives, this property of the Nash solution is precisely what ensures that the beliefs it generates satisfy cross-set consistency.

[34] By comparison, Nash-in-Nash is a "Nash Equilibrium in Nash bargains": each pair of agents makes choices cooperatively from the agreements available to them, and their behavior is then determined noncooperatively across different pairs of agents.

6 Conclusion

This paper takes a step towards the study of stable outcomes in applications where agents' preferences over agreements may exhibit both complementarities and substitutabilities, agreements can have externalities and be multilateral, and the market structure described by those agreements can be arbitrary. Our results suggest there might be new possibilities for the use of matching-theoretic models in applied work where the endogeneity of the observed agreements is of interest.

References

Agarwal, N., P. Somaini, et al. (2021): "Empirical models of non-transferable utility matching," *Online and Matching-Based Market Design*.

Ali, S. N., B. D. Bernheim, and X. Fan (2019): "Predictability and Power in Legislative Bargaining," *The Review of Economic Studies*, 86, 500–525.

Alva, S. (2018): "WARP and Combinatorial Choice," *Journal of Economic Theory*, 173, 320–333.

Bagwell, K., R. W. Staiger, and A. Yurukoglu (2021): "Quantitative Analysis of Multiparty Tariff Negotiations," *Econometrica*, 89, 1595–1631.

Bando, K. (2012): "Many-to-One Matching Markets With Externalities Among Firms," *Journal of Mathematical Economics*, 48, 14–20.

Bando, K. and T. Hirai (2021): "Stability and Venture Structures in Multilateral Matching," *Journal of Economic Theory*, 105292.

Baron, D. P. and J. A. Ferejohn (1989): "Bargaining in Legislatures," *American Political Science Review*, 83, 1181–1206.

Calvó-Armengol, A. and R. İlkiliç (2009): "Pairwise-Stability and Nash Equilibria in Network Formation," *International Journal of Game Theory*, 1, 51–79.

Chakraborty, A., A. Citanna, and M. Ostrovsky (2010): "Two-sided Matching with Interdependent Values," *Journal of Economic Theory*, 145, 85–105.

Collard-Wexler, A., G. Gowrisankaran, and R. S. Lee (2019): ""Nash-in-Nash" Bargaining: A Microfoundation for Applied Work," *Journal of Political Economy*, 127, 163–195.

Crawford, G. S. and A. Yurukoglu (2012): "The Welfare Effects of Bundling in Multichannel Television Markets," *American Economic Review*, 102, 643–85.

d'Aspremont, C. and L. Gevers (2002): "Social welfare functionals and interpersonal comparability," *Handbook of social choice and welfare*, 1, 459–541.

Dworczak, P. (2021): "Deferred Acceptance with Compensation Chains," *Operations Research*, 69, 456–468.

Ellickson, B., B. Grodal, S. Scotchmer, and W. R. Zame (1999): "Clubs and the Market," *Econometrica*, 67, 1185–1217.

Fisher, J. C. and I. E. Hafalir (2016): "Matching with Aggregate Externalities," *Mathematical Social Sciences*, 81, 1–7.

Fleiner, T., R. Jagadeesan, Z. Jankó, and A. Teytelboym (2019): "Trading Networks with Frictions," *Econometrica*, 87, 1633–1661.

Furusawa, T. and H. Konishi (2007): "Free Trade Networks," *Journal of International Economics*, 72, 310–335.

Gale, D. and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69, 9–15.

Goyal, S. and S. Joshi (2003): "Networks of Collaboration in Oligopoly," *Games and Economic behavior*, 43, 57–85.

Grennan, M. (2013): "Price Discrimination and Bargaining: Empirical Evidence from Medical Devices," *American Economic Review*, 103, 145–77.

Hafalir, I. E. (2008): "Stability of Marriage with Externalities," *International Journal of Game Theory*, 37, 353–369.

Hatfield, J. W. and F. Kojima (2008): "Matching With Contracts: Comment," *The American Economic Review*, 98, 1189–1194.

Hatfield, J. W. and S. D. Kominers (2012): "Matching in Networks with Bilateral Contracts," *American Economic Journal: Microeconomics*, 4, 176–208.

——— (2015): "Multilateral Matching," *Journal of Economic Theory*, 156, 175–206.

Hatfield, J. W., S. D. Kominers, A. Nichifor, M. Ostrovsky, and A. Westkamp (2013): "Stability and Competitive Equilibrium in Trading Networks," *Journal of Political Economy*, 121, 966–1005.

Hatfield, J. W. and P. R. Milgrom (2005): "Matching With Contracts," *American Economic Review*, 913–935.

Ho, K. and R. S. Lee (2017): "Insurer Competition in Health Care Markets," *Econometrica*, 85, 379–417.

——— (2019): "Equilibrium Provider Networks: Bargaining and Exclusion in Health Care Markets," *American Economic Review*, 109, 473–522.

Horn, H. and A. Wolinsky (1988): "Bilateral monopolies and incentives for merger," *The RAND Journal of Economics*, 408–419.

Jackson, M. O. (2010): *Social and Economic Networks*, Princeton University Press.

Jackson, M. O. and A. Watts (2002): "The Evolution of Social and Economic Networks," *Journal of Economic Theory*, 106, 265–295.

Jackson, M. O. and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *journal of economic theory*, 71, 44–74.

Jagadeesan, R. and K. Vocke (2021): "Stability in Large Markets," .

Kelso, A. S. and V. P. Crawford (1982): "Job Matching, Coalition Formation, and Gross Substitutes," *Econometrica: Journal of the Econometric Society*, 1483–1504.

Klaus, B. and M. Walzl (2009): "Stable many-to-many matchings with contracts," *Journal of Mathematical Economics*, 45, 422–434.

Kojima, F., P. A. Pathak, and A. E. Roth (2013): "Matching with Couples: Stability and Incentives in Large Markets," *The Quarterly Journal of Economics*, 128, 1585–1632.

Liebman, E. (2018): "Bargaining in Markets with Exclusion: An Analysis of Health Insurance Networks," *Working Paper, University of Georgia*.

Liu, C., Z. Wang, and H. Zhang (2023): "Self-Enforced Job Matching," *arXiv preprint arXiv:2308.13899*.

Liu, Q. (2020): "Stability and Bayesian Consistency in Two-Sided Markets," *American Economic Review*, 110, 2625–66.

——— (2022): "A Theory of Coalitional Games with Incomplete Information," Working paper.

Liu, Q., G. J. Mailath, A. Postlewaite, and L. Samuelson (2014): "Stable Matching with Incomplete Information," *Econometrica*, 82, 541–587.

Myerson, R. B. (1991): *Game Theory: Analysis of Conflict*, Harvard University Press.

Nash, J. F. (1950): "The Bargaining Problem," *Econometrica: Journal of the Econometric Society*, 155–162.

Nguyen, T. and R. Vohra (2018): "Near-feasible Stable Matchings with Couples," *American Economic Review*, 108, 3154–69.

Ostrovsky, M. (2008): "Stability in Supply Chain Networks," *American Economic Review*, 98, 897–923.

Pycia, M. (2012): "Stability and Preference Alignment in Matching and Coalition Formation," *Econometrica*, 80, 323–362.

Pycia, M. and M. B. Yenmez (2023): "Matching With Externalities," *The Review of Economic Studies*, 90, 948–974.

Rostek, M. and N. Yoder (2020): "Matching with Complementary Contracts," *Econometrica*, 88, 1793–1827.

——— (2023): "Strategic Consistency in Two-Sided Matching Markets," Working Paper.

Sadler, E. (2023): "A Unified Approach to Strategic Network Formation and Classical Matching Theory," *Available at SSRN*.

Sasaki, H. and M. Toda (1996): "Two-Sided Matching Problems with Externalities," *Journal of Economic Theory*, 70, 93–108.

Sun, N. and Z. Yang (2006): "Equilibria and Indivisibilities: Gross Substitutes and Complements," *Econometrica*, 74, 1385–1402.

——— (2009): "A Double-Track Adjustment Process for Discrete Markets With Substitutes and Complements," *Econometrica*, 77, 933–952.

Teytelboym, A. (2014): "Gross Substitutes and Complements: A Simple Generalization," *Economics Letters*, 123, 135–138.

Appendix

Proof of Lemma 1 For each $Y \subseteq X$, let $C(Y) := \cap_{i\in I}(C_i(Y_i|Y_{-i}) \cup Y_{-i})$. For each $i \in I$ and $Y \subseteq X$, since C_i is optimal given μ_i, $C_i(Y_i|Y_{-i}) \subseteq \mu_i(Y)$, and since μ_i is correct given $\{C_j\}_{j\neq i}$, $\mu_i(Y) = C^{-i}(Y)$. Then for each $i \in I$, $C_i(Y_i|Y_{-i}) \subseteq C^{-i}(Y)$, and hence

$$C_i(Y_i|Y_{-i}) = C_i(Y_i|Y_{-i}) \cap C^{-i}(Y) = ((C_i(Y_i|Y_{-i}) \cup Y_{-i}) \cap X_i) \cap C^{-i}(Y) = X_i \cap C(Y). \quad (7)$$

Then for each $j \in I$ and $Y \subseteq X$, $C_{-j}(Y) = C(Y)$: By definition, $C(Y) = (C_j(Y_j|Y_{-j}) \cup Y_{-j}) \cap C_{-j}(Y)$, so $C(Y) \subseteq C_{-j}(Y)$. Now suppose $x \in C_{-j}(Y)$. By assumption, $|N(x)| \geq 2$, so we must have $x \in C_i(Y_i|Y_{-i})$ for some $i \neq j$. Since $C_i(Y_i|Y_{-i}) = X_i \cap C(Y) \subseteq C(Y)$, it follows that $C_{-j}(Y) \subseteq C(Y)$.

Then since beliefs are correct, for all $i \in I$ and $Y \subseteq X$, $\mu_i(Y) = C(Y)$; (i) follows for $\mu(Y) = C(Y)$, and thus (ii) follows from (7). □

Lemma 4 adds to Lemma 1 by showing that given correctness and optimality, cross-set consistency is equivalent to the weak axiom (or equivalently, the irrelevance of rejected contracts condition (Alva (2018))) on the agents' common beliefs.

Lemma 4. *Suppose that the choice functions $\{C_i\}_{i\in I}$ are optimal given beliefs $\{\mu_i\}_{i\in I}$, and the beliefs $\{\mu_i\}_{i\in I}$ are correct given choice functions $\{C_i\}_{i\in I}$. Then $\{\mu_i\}_{i\in I}$ are cross-set consistent given $\{C_i\}_{i\in I}$ if and only if for each $i \in I$, $Y \supseteq Z \supseteq \mu_i(Y)$ implies $\mu_i(Y) = \mu_i(Z)$.*

Proof. By Lemma 1, for each $i, j \in I$ and $Y \subseteq X$, $\mu_i(Y) = \mu_j(Y) = \mu(Y)$ and $C_j(Y_j|Y_{-j}) = \mu(Y) \cap X_j$. Hence, for each $Y, Z \subseteq X$, $Y \supseteq Z \supseteq C_j(Y_j|Y_{-j})$ for each $j \in I \Leftrightarrow Y \supseteq Z \supseteq (\bigcup_{j\in I} \mu(Y) \cap X_j) = \mu(Y)$. Then since $\mu_i(S) = \mu_j(S) = \mu(S)$ for each $i, j \in I$ and $S \subseteq X$, $\{\mu_i\}_{i\in I}$ are cross-set consistent given $\{C_i\}_{i\in I}$ if and only if for each $Y, Z \subseteq X$, $Y \supseteq Z \supseteq \mu(Y) \Rightarrow \mu(Y) = \mu(Z)$. □

Proof of Theorem 1 (Strategic Consistency and Stability) Let $\{C_i, \mu_i\}_{i\in I}$ be a strategically consistent profile. By Lemma 1, for each $i, j \in I$ and $S \subseteq X$, $\mu_i(S) = \mu_j(S) = \mu(S)$ and $C_j(S_j|S_{-j}) = \mu(S) \cap X_j$.

Now for any $Z \subseteq X$, $X \supseteq \mu(X) \cup Z \supseteq \mu(X)$. Then by Lemma 4, $\mu(\mu(X) \cup Z) = \mu(X)$. Then by Lemma 1 (ii), for each $Z \subseteq X \setminus \mu(X)$ and each $i \in I$, $C_i((\mu(X) \cup Z)_i|(\mu(X) \cup Z)_{-i}) = \mu(\mu(X) \cup Z) \cap X_i = \mu(X) \cap X_i$. It follows that $\mu(X)$ is unblocked and (by setting $Z = \emptyset$) individually rational. □

Proof of Corollary 1 (Stability and Beliefs) Follows immediately from the proof of Theorem 1. □

Lemma 5 shows that optimal choices from Y given beliefs μ_i (as in a strategically consistent profile) are the same as nonstrategic choices from the set of contracts $\mu_i(Y)$ that an agent believes the other agents will choose from Y.

Lemma 5. *C_i is optimal given μ_i if and only if $C_i(Y_i|Y_{-i}) = \hat{C}_i(\mu_i(Y) \cap X_i|\mu_i(Y) \cap X_{-i})$ for all $Y \subseteq X$.*

Proof. From (1), we have $\hat{C}_i(\mu_i(Y)_i|\mu_i(Y)_{-i}) = \arg\max_{S \subseteq \mu_i(Y)_i} u_i(S \cup \mu_i(Y)_{-i})$; the statement follows immediately from the definition of optimality of C_i given μ_i. □

Proof of Lemma 2 (Converse of Lemma 1) (a): From condition (ii), for each $i \in I$ and $Y \subseteq X$,

$$C^{-i}(Y) := \bigcap_{j \neq i} (C_j(Y_j|Y_{-j}) \cup Y_{-j}) = \bigcap_{j \neq i} ((\mu_i(Y) \cap X_j) \cup Y_{-j}). \tag{8}$$

Then $\mu_i(Y) = C^{-i}(Y)$: If $x \in C^{-i}(Y)$, then since $|N(x)| \geq 2$, we must have $x \in X_j$ for some $j \neq i$, and hence, by (8), $x \in \mu_i(Y) \cap X_j \subseteq \mu_i(Y)$; it follows that $\mu_i(Y) \supseteq C^{-i}(Y)$. Moreover, since $\mu_i(Y) \subseteq Y$, $\mu_i(Y) \subseteq ((\mu_i(Y) \cap X_j) \cup Y_{-j})$ for each j, and so by (8), $\mu_i(Y) \subseteq C^{-i}(Y)$. Hence, $\{\mu_i\}_{i \in I}$ are correct given $\{C_i\}_{i \in I}$.

(b): (Beliefs are nonstrategically IR ⇒ Choices are optimal) Suppose that for each $Y \subseteq X$, $\mu(Y)$ is nonstrategically individually rational. Then by definition, for each $i \in I$ and $Y \subseteq X$, $\hat{C}_i(\mu(Y) \cap X_i|\mu(Y) \cap X_{-i}) = \mu(Y) \cap X_i = C_i(Y_i|Y_{-i})$. It follows from Lemma 5 that $\{C_i\}_{i \in I}$ are optimal given $\{\mu_i\}_{i \in I}$.

(Choices are optimal ⇒ Beliefs are nonstrategically IR) Suppose that $\{C_i\}_{i \in I}$ are optimal given $\{\mu_i\}_{i \in I}$. For each $Y \subseteq X$ and $i \in I$, we have $\mu_i(Y) \cap X_i = C_i(Y_i|Y_{-i})$ (by condition (ii)) and $C_i(Y_i|Y_{-i}) = \hat{C}_i(\mu(Y) \cap X_i|\mu(Y) \cap X_{-i})$ (by Lemma 5). Then for each $Y \subseteq X$ and $i \in I$, $\mu(Y) \cap X_i = \hat{C}_i(\mu(Y) \cap X_i|\mu(Y) \cap X_{-i})$, and so $\mu(Y)$ is nonstrategically individually rational. □

Lemma 6 shows that our algorithm (3) always constructs a strategically consistent profile.

Lemma 6 (Construction Algorithm). *For any strict total order $\succ$ on the set $\mathcal{M} = \{Y \subseteq X|\hat{C}_i(Y_i|Y_{-i}) = Y_i$ for each $i \in I\}$ of nonstrategically individually rational outcomes, the profile of choice functions and beliefs $\{C_i, \mu_i\}_{i \in I}$ defined in (3) is strategically consistent.*

Proof. First note that $\{C_i, \mu_i\}_{i \in I}$ is well-defined: Since $\{\hat{C}_i\}_{i \in I}$ are nonstrategic, we have $\hat{C}_i(\emptyset|\emptyset) = \emptyset$ for each $i \in I$, and so $\emptyset \in \mathcal{M}$. Then for each $Y \subseteq X$, $\{Y'|Y' \subseteq Y\}$ contains at least one element of $\mathcal{M}$, and so $\mu(Y) = \max_{\succ}\{Y'|Y' \subseteq Y\}$ is well-defined for each Y.

By construction, for each $i, j \in I$ and $Y \subseteq X$, (i) $\mu_i(Y) = \mu_j(Y) := \mu(Y)$, and (ii) $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$. Then by Lemma 2 (a), $\{\mu_i\}_{i \in I}$ are correct given $\{C_i\}_{i \in I}$. And since $\mu_i(Y) \in \mathcal{M}$ for each $Y \subseteq X$, by Lemma 2 (b), $\{C_i\}_{i \in I}$ are optimal given $\{\mu_i\}_{i \in I}$.

To show cross-set consistency, suppose $\mu(Y) \subseteq Z \subseteq Y$ for some $i \in I$ and $Y, Z \subseteq X$. Then by construction, $\mu(Y) \succeq Y'$ for each $Y' \subseteq Y$, and hence for each $Y' \subseteq Z$. Then by construction, $\mu(Y) = \mu(Z)$. It follows from Lemma 4 that $\{\mu_i\}_{i \in I}$ are cross-set consistent given $\{C_i\}_{i \in I}$. □

Lemma 7 (Strategically Consistent Profiles: Characterization). *There is a strategically consistent profile for which the outcome $Y \subseteq X$ is stable if and only if Y is nonstrategically individually rational.*

Proof. (Only if) Suppose S is stable for the strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$. By Corollary 1, $S = \mu_i(X)$ for each $i \in I$. Since $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent, by Lemma 1, for all $i, \in I$ and $Y \subseteq X$, $\mu_i(Y) = \mu_j(Y) = \mu(Y)$ and $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$. Then by Lemma 2 (b), S is nonstrategically individually rational.

(If) Suppose S is nonstrategically individually rational. Choose any strict total order $\succ$ on the set $\mathcal{M} = \{Y \subseteq X | \hat{C}_i(Y_i|Y_{-i}) = Y_i \text{ for each } i \in I\}$ of nonstrategically individually rational outcomes which ranks S highest. Let $\{C_i, \mu_i\}_{i \in I}$ be the profile of choice functions and beliefs constructed according to (3). By Lemma 6, $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent. Since $S \succeq Y$ for all $Y \in \mathcal{M}$, $S = \max_{\succ}\{Y' | Y' \subseteq X\} = \mu_i(X)$ for each $i \in I$. Then by Corollary 1, $\mu(X) = S$ is uniquely stable for $\{C_i, \mu_i\}_{i \in I}$. □

Proof of Theorem 2 (Existence of Strategically Consistent Profiles) By definition, $\hat{C}_i(\emptyset|\emptyset) = \emptyset$ for each $i \in I$. Hence, $\emptyset$ is nonstrategically individually rational; existence follows from the "if" part of Lemma 7. □

Lemma 8. *If $\phi : \mathbb{R}^I_+ \to \mathbb{R}$ is a strictly increasing function, then for any $Y \in \arg\max_{S \in \mathcal{M}} \phi((u_i(S))_{i \in I})$, there is a strict total order $\succ^\phi$ that is induced by ϕ which ranks Y highest.*

Proof. Let $\mathcal{M} = \{Z \subseteq X | \hat{C}_i(Z_i|Z_{-i}) = Z_i \text{ for each } i \in I\}$ denote the set of nonstrategically individually rational outcomes, and label its elements according to the sequence $\{Y^n\}_{n=1}^{|\mathcal{M}|}$, constructed recursively as follows: To begin, let $Y^1 = Y \in \arg\max_{S \in \mathcal{M}} \phi((u_i(S))_{i \in I})$. Then, given elements $\{Y^n\}_{n=1}^m$, choose $Y^{m+1} \in \arg\max_{S \in \mathcal{M} \setminus \{Y^n\}_{n=1}^m} \phi((u_i(S))_{i \in I})$. (The set of maximizers is nonempty since X (and hence $\mathcal{M} \subseteq 2^X$) is finite.) This construction implies that whenever $\phi((u_i(Y^n))_{i \in I}) > \phi((u_i(Y^m))_{i \in I})$, we must have $n < m$: If $n > m$, then $Y^n \in \mathcal{M} \setminus \{Y^k\}_{k=1}^{m-1}$, and so Y^m could not have been chosen as the mth element of the sequence.

Now define the order $\succ^\phi$ on $\mathcal{M}$ as follows: $Y^n \succ^\phi Y^m \Leftrightarrow n < m$. Since $\{Y^n\}_{n=1}^m = \mathcal{M}$, we can label any two elements of $\mathcal{M}$ as Y^n and Y^m for some n, m. If $\phi((u_i(Y^n))_{i \in I}) > \phi((u_i(Y^m))_{i \in I})$, we must have $n < m$, and hence $Y^n \succ^\phi Y^m$. So $\succ^\phi$ is induced by ϕ, as desired. □

Proof of Theorem 3

(i): Strategic consistency follows from Lemma 6. For Pareto optimality, suppose that $Y, Z \subseteq X$ are such that $u_i(Y) \geq u_i(Z)$ for all $i \in I$ and $u_i(Y) > u_i(Z)$ for some $i \in I$. Then since ϕ is strictly increasing, we must have $\phi((u_i(Y))_{i\in I}) > \phi((u_i(Z))_{i\in I})$. Thus, since $\succ^\phi$ is induced by ϕ, we have $Y \succ^\phi Z$. Then the algorithm (3) yields $\mu_i^\phi(Y \cup Z) = \mu^\phi(Y \cup Z) = \max_{\succ^\phi}\{Y' | Y' \subseteq Y \cup Z\} \neq Z$ for each $i \in I$. It follows that $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$ satisfies Pareto optimality.

(ii): For each $i \in I$ and $Z \subseteq X$, we have $\mu_i^\phi(Z) = \mu^\phi(Z) = \max_{\succ^\phi}\{S | S \subseteq Z\}$ from (3). Then $\mu_i^\phi(Z) \succ^\phi s$ for all $S \subseteq Z$. Since $\succ^\phi$ is induced by ϕ, it follows that for all $S \subseteq Z$, $\phi((u_j(\mu_i^\phi(Z)))_{j\in I}) \geq \phi((u_j(S))_{j\in I})$: If not, then $\phi((u_j(\mu_i^\phi(Z)))_{j\in I}) < \phi((u_j(S))_{j\in I})$, and hence $\mu_i^\phi(Z) \prec^\phi S$. It follows that $\mu_i^\phi(Z)$ solves (4). □

Proof of Theorem 4 (Welfare Theorem for Strategic Consistency) (Only if) Suppose that Y is stable for a strategically consistent profile $\{C_i, \mu_i\}_{i\in I}$ that satisfies Pareto optimality. By Lemma 7, Y is nonstrategically individually rational. Suppose there is another nonstrategically individually rational outcome Z that Pareto-dominates Y: $u_i(Z) \geq u_i(Y)$ for all $i \in I$ and $u_i(Z) > u_i(Y)$ for some $i \in I$. Then since $\{C_i, \mu_i\}_{i\in I}$ satisfies Pareto optimality, $\mu_i(Z \cup Y) = Z$ for each $i \in I$. Then we must have $\mu_i(X) \neq Y$ for each $i \in I$; otherwise, cross-set consistency would imply $\mu_i(Z \cup Y) = Y \neq Z$. Then by Corollary 1, Y is not stable for $\{C_i, \mu_i\}_{i\in I}$, a contradiction.

(If) Denote by $\mathcal{M}$ the nonstrategically individually rational outcomes, and suppose that Y is Pareto efficient among these outcomes: there exists no $S \in \mathcal{M}$ such that $u_i(S) \geq u_i(Y)$ for all $i \in I$ and $u_i(S) > u_i(Y)$ for some $i \in I$. Then for every $S \in \mathcal{M}$, either $u_i(S) = u_i(Y)$ for all $i \in I$ or $u_i(S) < u_i(Y)$ for some $i \in I$. For each $\rho < 0$, define

$$\begin{aligned} \phi_\rho : \mathbb{R}_+^I &\to \mathbb{R} \\ x &\mapsto \left(\sum_{i\in I}\left(\frac{x_i}{u_i(Y)}\right)^\rho\right)^{1/\rho}. \end{aligned}$$

Each ϕ_ρ is strictly increasing, since

$$\frac{\partial \phi_\rho}{\partial x_i}(x) = \frac{x_i^{\rho-1}}{u_i(Y)^\rho}\left(\sum_{i\in I}\left(\frac{x_i}{u_i(Y)}\right)^\rho\right)^{1/\rho-1} > 0 \text{ for each } i \in I.$$

Now for any $Z \in \mathcal{M}$, we have

$$\begin{aligned}
\phi_\rho((u_i(Y))_{i\in I}) - \phi_\rho((u_i(S))_{i\in I}) &= 1 - \left(\sum_{i\in I}\left(\frac{u_i(S)}{u_i(Y)}\right)^\rho\right)^{1/\rho}; \\
\lim_{\rho\to-\infty}\left(\phi_\rho((u_i(Y))_{i\in I}) - \phi_\rho((u_i(S))_{i\in I})\right) &= 1 - \min\left\{u_i(S)/u_i(Y)\right\}_{i\in I} \\
&< 0, \text{ if } (u_i(S))_{i\in I} \neq (u_i(Y))_{i\in I}.
\end{aligned}$$

Then for every $S \in \mathcal{M}$ with $(u_i(Y))_{i\in I} \neq (u_i(S))_{i\in I}$, there exists r_S such that for all $\rho < r_S$, $\phi_\rho((u_i(Y))_{i\in I}) > \phi_\rho((u_i(S))_{i\in I})$. Choose $\rho^* = \min_{S\in\mathcal{M}} r_S$ and let $\phi = \phi_{\rho^*}$; it follows that $Y \in \arg\max_{S\in\mathcal{M}} \phi((u_i(S))_{i\in I})$. Then by Lemma 8, there is a strict total order $\succ^\phi$ that is induced by ϕ and ranks Y highest, and by Theorem 3(i), the profile $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$ constructed from $\succ^\phi$ using the algorithm (3) is strategically consistent and satisfies Pareto optimality. By (3), since $\succ^\phi$ ranks Y highest, $\mu_i(X) = Y$ for each $i \in I$; it follows from Corollary 1 that Y is stable for $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$, as desired. □

Proof of Lemma 3 Follows from d'Aspremont and Gevers (2002) Theorem 4.17. □