Counterfactual Analysis in Bargaining with Externalities: A Matching-Theoretic Foundation*

Marzena Rostek[†] and Nathan Yoder[‡]

June 9, 2025

Abstract

We show how matching-theoretic stability can be used to perform counterfactual analysis in environments with externalities. Our approach sidesteps the well-known existence problems faced by stability in these environments by endowing agents with correct beliefs about the choices of others. This facilitates a procedure that mirrors the one commonly used with the Nash-in-Nash solution concept, with beliefs playing the same role as bargaining weights: Like profiles of bargaining weights in Nash-in-Nash, each profile of beliefs predicts a unique outcome, but different profiles predict different outcomes.

Keywords: Externalities, matching with contracts, stability, Nash-in-Nash

1 Introduction

A great deal of recent empirical work explores settings where agents form agreements with one another that have externalities on other agents. In industrial

* Some parts of this paper were previously included in "Matching with Strategic Consistency". The authors are grateful to numerous colleagues for their helpful comments and suggestions.

† 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.

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.

In this note, we show by example how our results from Rostek and Yoder (2025) can be applied to perform counterfactual analysis in settings like these. Because the approach that we take is based on the canonical concept of *stability* from matching theory, its predictions are robust to *arbitrary* deviations. For instance, firms can consider swappping an agreement with one supplier for an agreement with another, or simultaneously consider agreements with multiple suppliers for the purchase of complementary inputs. Assumptions about bargaining protocols, delegated agents, or exogenously fixed outside options are unnecessary. Instead, we show in Rostek and Yoder (2025) that the well-known challenges for existence of a stable outcome in these environments can be overcome by ensuring that agents' choices are *optimal* given *beliefs* about others' choices that are *correct* and *consistent* across sets of alternatives.

Our approach to this exercise mirrors the approach often taken with the *Nash-in-Nash* (Horn and Wolinsky (1988)) solution concept — a Nash equilibrium in Nash bargains. This solution concept has found extensive use in both theoretical and empirical work. Nash-in-Nash offers a key advantage that has helped make it popular: It avoids the nonexistence issues faced by stability in applications with externalities by considering a smaller class of deviations. As we show in Rostek and Yoder (2025), endogenizing agents' choice functions as best responses to beliefs about the choices of others allows matching-theoretic stability to overcome these issues, while still allowing agents to renegotiate agreements in arbitrary ways.

Doing so creates a degree of freedom that is similar to the one exploited by Nash-in-Nash counterfactual analysis. Observe that 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. Likewise, we show in Rostek and Yoder (2025) that each profile of correct, consistent beliefs pins down a *unique* stable outcome. Because of the fixed point relationship between choices and beliefs, there may be several such profiles. This suggests exploiting this degree of freedom in the same way, by recovering a profile of beliefs from data and then using it to pin down a new outcome after an exogenous change in the model.

This is what we do in this paper. As Rostek and Yoder (2025) shows, this analogy between beliefs and bargaining weights can be formalized: Correct, consistent beliefs can be constructed by maximizing an asymmetric Nash product, subject to constraints that capture the agents' (endogenous) outside options. We can thus think of these beliefs as a microfoundation for Nash bargaining weights, or vice versa. Since these beliefs each pin down a unique stable outcome, we show that we can use this maximization problem in exactly the same way that the Nash-in-Nash maximization problems are used in practice for counterfactual analysis.

2 Environment

We demonstrate our approach to stability and counterfactual analysis in a model of contracting between upstream firms (i.e., suppliers of inputs) $i \in I$ and downstream firms (i.e., producers of final goods) $h \in H$ à la Collard-Wexler et al. (2019). A *contract* is an agreement between an upstream firm i and a downstream firm f that specifies a price (or more generally, a vector of prices) p_{if}. These prices may be lump sum payments, as in Collard-Wexler et al. (2019), or prices per unit of input sold (e.g., price per subscriber for each channel in Crawford and Yurukoglu (2012), or price per stent in Grennan (2013)). Prices must be whole units of money, and so each p_{if} is an element of a finite set $\mathcal{P} \subset \mathbb{R}_+$. The set of all such contracts is thus $X := \{(i, f, p_{if}) : i \in I, f \in F, p_{if} \in \mathcal{P}\}$. For each upstream firm $i \in I$, we let $X_i := \{(i, f, p_{if}) : i \in I, p_{if} \in \mathcal{P}\}$ denote the set of contracts involving i, and $X_{-i} := X \setminus X_i$ denote the set of contracts that do not; likewise, for each downstream firm $f \in F$, we let $X_f := \{(i, f, p_{if}) : i \in I, \tau_{if} \in \mathcal{P}\}$ denote the set of contracts involving f, and $X_{-f} := X \setminus X_f$ denote the set of contracts that do not. Similarly, for sets of contracts $Y \subseteq X$, we write $Y_j := Y \cap X_j$ and $Y_{-j} := Y \cap X_{-j}$

for each $j \in I \cup F$.

Given a set of contracts $Y \subseteq X$ — an *outcome* — downstream firms compete with one another in the market for final goods.[1] Their equilibrium pin down the payoff $u_i(Y)$ of each upstream firm i, and the payoff $u_f(Y)$ of each downstream firm f, from that outcome. These preferences may exhibit externalities: in general, $u_j(Y)$ depends on Y_{-j}, not just Y_j. For instance, in Crawford and Yurukoglu (2012), the number of subscribers to a television channel i through a distributor f, and the prices that they pay, depend on the equilibrium bundles and pricing schemes offered by all distributors in that market. Since that equilibrium depends on the *entire* set of contracts between TV producers and distributors, so do producer and distributor profits.

We assume that firms are never indifferent about sets of agreements they might make, conditional on the agreements made by others: for each $j \in I \cup F$, $u_j(Y \cup X') \neq u_j(Z \cup X')$ for each distinct $Y, Z \subseteq X_j$ and each $X' \subseteq X_{-j}$. Each manufacturer-hospital pair can only sign at most one contract; we capture this by setting $u_j(Y \cup \{(i, f, p_{if}), (i, f, p'_{if})\}) < u_j(Y \cup \{(i, f, p_{if})\})$ for each $i \in I$, $f \in F$, and $j \in \{i, f\}$, and each pair of prices $p_{if}, p'_{if} \in \mathcal{P}$. When an outcome contains at most one contract between each pair of firms, a firm j's payoff is just its profit $\pi_j(\mathbf{p}_A; A)$ given the vector $\mathbf{p}_A$ of prices agreed to by the upstream-downstream pairs $A \subseteq I \times F$ that contract with each other.

2.1 Stability

We use stability, the canonical solution concept in matching theory, extended to allow for externalities. Whether an outcome is stable or not depends on preferences indirectly through the agents' *choice functions* $C_j : 2^{X_j} \times 2^{X_{-j}} \to 2^{X_j}$. Firm j's choice function C_j takes two arguments: the set of contracts that are available (i.e., under discussion in a negotiation) that involve firm j, and the set of contracts that are available but do not involve firm j. $C_j(Y_j | Y_{-j}) \subseteq Y_j$ is the set of contracts that firm j chooses when Y is the set of contracts under discussion.

Definition. Given choice functions $\{C_j\}_{j \in I \cup F}$, a set of contracts $Y \subseteq X$ is stable if it is

[1] As in Collard-Wexler et al. (2019), we abstract away from the form of that competition to focus on bargaining between upstream and downstream firms.

i. *Individually rational*: $Y_j = C_j(Y_j|Y_{-j})$ for all $j \in I \cup F$.

ii. *Unblocked*: There does not exist a nonempty $Z \subseteq (X \setminus Y)$ such that for all $j \in I \cup F$ with $Z_j \neq \emptyset$, $Z_j \subseteq C_j((Z \cup Y)_j|(Z \cup Y)_{-j})$.

In words, a set of contracts is stable if (a) when it is available, no one rejects any contracts from it (individual rationality), and (b) no group of agents can propose a new set of contracts, or *block*, that they are each willing to choose when made available (i.e., under negotiation) alongside the existing set of contracts.

2.2 Choices and Beliefs

Because pricing agreements have externalities on other manufacturers, stability could be difficult to apply with the standard, nonstrategic approach to choice, in which firms' choices $C_j(Y_j|Y_{-j})$ are simply given by their *favorite* subset of Y_j, without regard for whether some of those agreements might be rejected by others. This is because of well-understood existence issues that are not present with the Nash-in-Nash solution applied in papers such as Crawford and Yurukoglu (2012), Grennan (2013), or Ho and Lee (2017).

Rostek and Yoder (2025) introduces an approach that allows us to get around these issues and use stability to make predictions in environments with externalities. Its key innovation is to endogenously determine the agents' choice functions as part of a fixed point together with *correct* beliefs about the choices of others, such that beliefs are *consistent* across different sets of contracts that might be under discussion in a negotiation. Formally, an agent's beliefs are described by a function $\mu_i : 2^X \to 2^X$ that takes the set of available contracts as its argument, and returns the set of contracts that agent i believes will not be rejected by any of the other agents.

Definition. Given firms' payoff functions, the profile of choice functions and beliefs $\{C_j, \mu_j\}_{j \in I \cup F}$ is *strategically consistent* if for each $j \in I \cup F$,

i. μ_j is *correct* given $\{C_k\}_{k \neq j}$: For each $Y \subseteq X$,

$$\mu_j(Y) = \{(i, f, p_{if}) \in Y \mid (i, f, p_{if}) \in C_i(Y_i|Y_{-i}) \cap C_f(Y_f|Y_{-f})\}.$$

ii. C_j is *optimal* given μ_j: For each $Y \subseteq X$,

$$C_j(Y_j|Y_{-j}) = \arg\max_S u_j(S \cup \mu_j(Y)_{-j}) \text{ s.t. } S \subseteq \mu_j(Y)_j.$$

iii. μ_j is *cross-set consistent* given $\{C_k\}_{k \in I \cup F}$: For each $Y, Z \subseteq X$, if $Y \supseteq Z \supseteq C_j(Y_j|Y_{-j})$ for all $j \in I \cup F$, then $\mu_j(Z) = \mu_j(Y)$.

As in Rostek and Yoder (2025), we focus on profiles with beliefs that are *Pareto optimal*. That is, agents do not believe that others will choose a Pareto-dominated set of contracts just because of coordination failure. Formally, suppose that Y is a *nonstrategically individually rational* set of contracts: $u_j(Y) \geq u_j(S \cup Y_{-j})$ for all $j \in I \cup F$ and $S \subseteq Y_j$. μ_j satisfies Pareto optimality if for any such Y, there is no Z that (a) Pareto improves upon Y; (b) is also nonstrategically individually rational; but (c) j believes the other agents will still choose Y when Z has been proposed: $\mu_j(Y \cup Z) = Y$.

Rostek and Yoder (2025) offers two results that are key to our framework for counterfactual analysis. First, these profiles of beliefs each pin down a unique stable outcome. Second, they can be generated by maximizing a social welfare function over the nonstrategically rational outcomes.[2] We summarize these conclusions below in the language of our setting.

Theorem 1. *Let $\phi : \mathbb{R}_+^{I \cup F} \to \mathbb{R}$ be a strictly increasing function such that $\phi((u_j(Y))_{j \in I \cup F}) \neq \phi((u_j(Z))_{j \in I \cup F})$ for any distinct $Y, Z \subseteq X$. For each j and $Y \subseteq X$, define*

$$\mu_j^\phi(Y) = \arg\max_{S \subseteq Y} \phi((u_j(S))_{j \in I \cup F}) \text{ s.t. } u_k(S) \geq u_k(S' \cup S_{-k}) \text{ for all } k \in I \cup F \text{ and } S' \subset S_k; \tag{SPP(Y)}$$

$$C_j^\phi(Y_j|Y_{-j}) = \arg\max_S u_j(S \cup \mu_j^\phi(Y)_{-j}) \text{ s.t. } S \subseteq \mu_j^\phi(Y)_j.$$

$\{C_j^\phi, \mu_j^\phi\}_{j \in I \cup F}$ is strategically consistent, each μ_j^ϕ satisfies Pareto optimality, and the solution to (SPP(X)) is the unique outcome that is stable given the choice functions $\{C_j^\phi\}_{j \in I \cup F}$.

The role played by correct, consistent beliefs in our framework based on stability is analogous to that played by bargaining parameters in frameworks based on Nash bargaining: Each profile of beliefs or vector of bargaining parameters pins

[2] With a consistent tiebreaking rule, where necessary.

down a unique outcome, but different outcomes are predicted by different profiles of beliefs or bargaining parameters. This allows us to use them for our approach to counterfactual analysis analogously to the way that bargaining parameters are used in approaches based on Nash bargaining.

In fact, this analogy runs deeper: we can think of these beliefs as a microfoundation for Nash bargaining weights, or vice versa. Specifically, we can think of the welfare function ϕ used in Theorem 1 as formalizing the way that agents base their beliefs about other agents' choices on the payoffs the agents will receive from those choices. If this mapping from payoffs to beliefs is invariant under a rescaling of the firms' payoff functions, then it can be described by letting ϕ be an asymmetric Nash product $\phi(\mathbf{x}) = \prod_{j \in I \cup F} x_j^{\alpha_j}$ for some distribution $\alpha \in \Delta(I \cup F)$ (Lemma 3 in Rostek and Yoder (2025)). Thus, given the weights α, (SPP(X)) pins down the stable outcome $Y^*(\alpha)$ as

$$Y^*(\alpha) = \arg\max_{S \subseteq X} \prod_{j \in I \cup F} u_j(S)^{\alpha_j} \text{ s.t. } u_k(S) \geq u_k(S' \cup S_{-k}) \text{ for all } k \in I \cup F \text{ and } S' \subset S_k. \tag{1}$$

It bears emphasizing that even though the planner's problem (1) constructs a stable outcome (and more generally, (SPP(Y)) constructs a strategically consistent profile of choice functions and beliefs) as the solution to a multilateral Nash bargaining problem, its interpretation is *not* that all firms 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, firms' 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 and scale invariance).[3] 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 (1) as describing a "stability-in-Nash" approach: a stable outcome given optimal choices from beliefs that maximize a Nash product at each set of available contracts.[4]

[3]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.

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

3 Counterfactual Analysis

The procedure for Nash-in-Nash counterfactual analysis used in papers like Crawford and Yurukoglu (2012), Grennan (2013), or Ho and Lee (2017) can be described roughly as follows:

1. Use the data to recover (a) firms' preferences over agreements, and (b) the Nash bargaining weights that pin down the agreements actually observed. E.g., in Crawford and Yurukoglu (2012), estimate consumer demand as a function of the menu of television bundles and prices that they face, use equilibrium conditions from the competition in menus among the TV distributors to recover firms' marginal costs, and finally estimate the bargaining weights for which the observed input prices are the Nash-in-Nash solution.

2. Use the recovered preferences and bargaining weights to make predictions in some counterfactual scenario. E.g., in Crawford and Yurukoglu (2012), modify the environment to account for a proposed regulation on the menus that TV distributors can offer, and then recompute the Nash-in-Nash prices.

Our approach follows the same steps, but instead of recovering Nash bargaining weights that pin down a bargaining solution, we recover *correct, consistent beliefs* that pin down a stable outcome. This may appear more challenging than the procedure used with the Nash-in-Nash concept. First, the observed outcome does not necessarily pin down the full profile of beliefs and choice functions for which it is stable. Second, unlike the set of vectors of bargaining weights, 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 using the results from Rostek and Yoder (2025) described in Theorem 1. Recall from Theorem 1 that correct, consistent, Pareto optimal beliefs can be constructed by maximizing a social welfare function. Each such profile $\{\mu_j^{\phi}\}_{j\in I\cup F}$ 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, a vector of Nash weights α.

3.1 Stable Outcomes and Nash Weights

Both steps in the procedure outlined above require us to characterize the relationship between the stable outcome that we observe (or predict), and the welfare function that captures the way that agents' beliefs depend upon each other's preferences. We focus on welfare functions that are scale invariant, and can thus be represented as a Nash product. Thus, at a high level, this relationship is described by (1).

However, the structure of the model allows us to be more concrete. First note that since each upstream-downstream pair can only make one supply agreement, the outcomes that satisfy the individual rationality constraint in (1) cannot contain multiple agreements between the same pair of firms. Hence, we can write (1) as an optimization problem over (a) sets of firm pairs $A \subseteq I \times F$ that form agreements, and (b) price vectors $\mathbf{p}_A = (p_{if})_{(i,f)\in A}$ specified by those agreements, by replacing each firm's payoff function u_i over agreements with their profit function π_i over price vectors, as follows:

$$(A^*, \mathbf{p}^*_{A^*}) = \arg\max_{A, \mathbf{p}_A} \prod_{i\in I} \pi_i(\mathbf{p}_A; A)^{\alpha_i} \prod_{f\in F} \pi_f(\mathbf{p}_A; A)^{\alpha_f} \tag{2}$$

$$\text{s.t. } \pi_i(\mathbf{p}_A; A) \geq \pi_i(\mathbf{p}_{A\setminus B}; A \setminus B) \text{ for all } i \in I \text{ and } B \subseteq \{i\} \times F; \qquad \text{(NSIR}(i))$$

$$\pi_f(\mathbf{p}_A; A) \geq \pi_f(\mathbf{p}_{A\setminus B}; A \setminus B) \text{ for all } f \in F \text{ and } B \subseteq \{f\} \times I. \qquad \text{(NSIR}(f))$$

(2) describes the relationship between firm profits and beliefs (as characterized by the vector of bargaining weights α in the Nash product) and the supply relationships and negotiated prices that they pin down as part of a stable outcome. We can thus use it in the procedure for counterfactual analysis described in the outset of this section, in precisely the same way as the *set* of optimization problems that

characterize the Nash-in-Nash solution:[5]

$$p^*_{if} = \arg\max_{p_{if}} \begin{array}{l} (\pi_f(p_{if}, \mathbf{p}^*_{I\times F\setminus\{(i,f)\}}; I \times F) - \pi_f(\mathbf{p}^*_{(I\times F)\setminus\{(i,f)\}}; (I \times F) \setminus \{(i,f)\}))^{\alpha_{if}} \\ \times(\pi_i(p_{if}, \mathbf{p}^*_{I\times F\setminus\{(i,f)\}}, I \times F) - \pi_i(\mathbf{p}^*_{(I\times F)\setminus\{(i,f)\}}; (I \times F) \setminus \{(i,f)\}))^{1-\alpha_{if}} \end{array} \quad (3)$$

for each $i \in I, f \in F$.

Using (2) in this way results in our procedure for counterfactual analysis based on stability (in the matching theory sense). We can first use (2) together with the data to recover the Nash bargaining weights that pin down observed agreements and prices. Then, we can make a change to the model and use the same bargaining weights to pin down agreements and prices in some counterfactual scenario.

3.2 Discussion

The contrast between (2) and (3) helps to illustrate the differences between our procedure and the usual one based on the Nash-in-Nash solution, and more generally, the differences between matching-theoretic stability and the Nash-in-Nash concept. First, stability endogenizes the set of firm pairs that form agreements. Consequently, it allows for endogenous exclusion: Unlike (3), (2) accommodates outcomes where some upstream-downstream pairs do not contract (and so $A^* \neq I \times F$). This could be because agreements between such pairs are inefficient. But it could also be because, due to the externalities that agreements have on other firms, a Pareto-improving set of contracts would not be individually rational.

Second, with stability, agreements are not negotiated in isolation from each other. The existence of gains from modifying multiple agreements at once — e.g., a simultaneous decrease in the prices of complementary inputs from distinct upstream firms — impacts the stable outcome pinned down by (2), whereas in **?**, gains from negotiation with different counterparties are considered separately.

Finally, unlike (3), no "outside option" terms appear in the objective function in (2). This is because firms' outside options to a contract (or set of contracts) are already explicitly considered by the underlying stability concept. (2) is just a "reduced form" for stability with correct, consistent beliefs that is provided by

[5] See, e.g., Eq. 7 in Crawford and Yurukoglu (2012); Eq. 7 in Grennan (2013); Eqs. 3 and 4 in Ho and Lee (2017); or p. 174 in Collard-Wexler et al. (2019).

Theorem 1. As a result of Theorem 1, the impact of firms' outside options on (2) is accounted for through the nonstrategic individual rationality constraints (NSIR(i)) and (NSIR(f)).

4 Conclusion

This note shows how matching-theoretic stability can be used for counterfactual analysis in the same way (and in the same environments) as the Nash-in-Nash solution concept is commonly used. This allows an analyst to sidestep the existence issues faced by stability in environments with externalities without restricting the alternative sets of agreements that agents can consider. It also allows for endogenous exclusion.

References

BAGWELL, K., R. W. STAIGER, AND A. YURUKOGLU (2021): "Quantitative Analysis of Multiparty Tariff Negotiations," *Econometrica*, 89, 1595–1631.

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.

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

HO, K. AND R. S. LEE (2017): "Insurer Competition in Health Care Markets," *Econometrica*, 85, 379–417.

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

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

ROSTEK, M. AND N. YODER (2025): "Matching with Strategic Consistency," Working Paper.