Three Essays on Microeconomic Theory

Thesis by

SangMok Lee

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2012

(Defended May 19, 2012)

To my family, advisors, and friends.

Abstract

This thesis considers three issues in microeconomic theory — two-sided matching, strategic voting, and revealed preferences.

In the first chapter I discuss the strategic manipulation of stable matching mechanisms commonly used in two-sided matching markets. Stable matching mechanisms are very successful in practice, despite theoretical concerns that they are manipulable by participants. The key finding is that most agents in large markets are close to being indifferent among partners in all stable matchings. It is known that the utility gain by manipulating a stable matching mechanism is bounded by the difference between utilities from the best and the worst stable matching partners. Thus, the main finding implies that the proportion of agents who may obtain a significant utility gain from manipulation vanishes in large markets. This result reconciles the success of stable mechanisms in practice with the theoretical concerns about strategic manipulation. Methodologically, I introduce techniques from the theory of random bipartite graphs for the analysis of large matching markets.

In the second chapter I study the criminal court process, focusing on plea bargaining. Plea bargains screen the types of defendants, guilty or innocent, who go to jury trial, which affects the jurors' voting decision and, in turn, the performance of the entire criminal court. The equilibrium jurors' voting behavior in the case of plea bargaining resembles the equilibrium behavior in the classical jury model in the absence of plea bargaining. By optimizing a plea bargain offer, a prosecutor, however, may induce jurors to act as if they echo the prosecutor's preferences against convicting innocent defendants and acquitting guilty defendants. With reference to Feddersen and Pesendorfer (1998), I study different voting rules in the trial stage and their consequences in the entire court process. Compared to general super-majority rules, we find that a court using the unanimity rule delivers more expected punishment to innocent defendants and less punishment to guilty defendants.

In the third chapter I study collective choices from the revealed preference theory viewpoint. For every product set of individual actions, joint choices are called Nash-rationalizable if there exists a preference relation for each player such that the selected joint actions are Nash equilibria of the corresponding game. I characterize Nash-rationalizable joint choice behavior by zero-sum games, or games of conflicting interests. If the joint choice behavior forms a product subset, the behavior is called interchangeable. I prove that interchangeability is the only additional empirical condition which distinguishes zero-sum games from general noncooperative games.

Contents

List of Figures

List of Tables

Chapter 1

Incentive Compatibility of Large Centralized Matching Markets

1.1 Introduction

1.1.1 Overview

In this paper, we study the most popular class of algorithms, called *stable matching mechanisms*, used in centralized matching markets, such as the National Resident Matching Program (NRMP) and School Choice Programs in NYC and Boston. A matching is regarded as stable if no agent is matched with an unacceptable partner, and there is no pair of agents on opposite sides of the market who prefer each other to their current partners. A stable matching mechanism takes preference reports by participants and produces a stable matching with respect to the submitted preferences. We ask how stable matching mechanisms remain so successful, despite the fact that the mechanisms are easily manipulable by the participants through misrepresenting their preferences. In particular, we analyze whether large markets, i.e. ones consisting of a large number of participants, would mitigate incentives to manipulate a stable matching mechanism.

Two-sided matching markets are markets with two kinds of agents, in which agents of one kind match with agents of the other kind. Examples of such markets include firms and workers in professional labor markets (Roth and Peranson, 1999), schools and students in school choice programs (Abdulkadiroglu and Sönmez, 2003), men and women in the marriage market or dating sites (Choo and Siow, 2006; Hitsch, Hortaçsu, and Ariely, 2010), birth mothers and potential adoptive parents in the market for child adoption (Bernal, Hu, Moriguchi, and Nagypal, 2007; Baccara, Collard-Wexler, Felli, and Yariv, 2010), and cadets and branches in the military (Sönmez and Switzer, 2011). Market designers seeking to achieve desirable outcomes to these matching markets have introduced centralized clearinghouses.

In market design, the concept of “stability” has been considered of central importance. In practice, successful mechanisms often implement a stable matching with respect to submitted preferences (Roth and Xing, 1994; Roth, 2002). The best-known market design examples, such as the NRMP and School Choice Programs in NYC and Boston, also use a particular stable matching mechanism, called the doctor-proposing or student-proposing Gale-Shapley algorithm.[1] Table 1.1 below lists whether each clearinghouse produces a stable matching with respect to submitted preferences, and whether these clearinghouses are still in use or no longer operating. With few exceptions, stable matching mechanisms have been successful for the most part whereas unstable mechanisms have mostly failed.[2]

	Still in use	**No longer in use**
Stable	**The NRMP**: over 40 specialty markets and submarkets for first year postgraduate positions, and 15 for second year positions **Specialty matching services**: over 30 subspecialty markets for advanced medical residencies and fellowships **School choice programs**: NYC, Boston **Canadian lawyers:** multiple regions **British regional medical markets**: Edinburgh ($\geq$‘69), Cardiff **Dental residencies**: 3 specialties **Other healthcare markets**: Osteopaths ($\geq$‘94), Pharmacists, Clinical psychologists ($\geq$‘99)	**Dental residencies**: Periodontists($<$‘97), Prosthodontists ($<$‘00) **Canadian lawyers:** British Columbia($<$‘96)
Unstable	**British regional medical markets**: Cambridge, London Hospital	**British regional medical markets**: Birmingham, Edinburgh ($<$‘67), Newcastle, Sheffield **Other healthcare markets**: Osteopaths ($<$‘94)

Table 1.1: Stable and unstable (centralized) mechanisms.

[1] The algorithm is customized for each application. For details of the actual algorithms applied, see Roth and Peranson (1999); Abdulkadiroglu, Pathak, and Roth (2009); Abdulkadiroglu, Pathak, Roth, and Sönmez (2006).

[2] Table 1.1 is reorganized from tables in Roth (2002) and McKinney, Niederle, and Roth (2003). The clearinghouse for the gastroenterology fellowship market is a rare case in which a stable matching mechanism started to fail in 1996, was abandoned in 2000, and then was reinstated in 2006 (Niederle and Roth, 2005; Roth, 2008).

From a theoretical perspective, however, stable matching mechanisms have a significant shortcoming. While the mechanisms produce stable matchings by assuming that all participants reveal their true preferences, in fact *no stable matching mechanism is strategy-proof* (Roth, 1982).[3] Participants may achieve a better matching by misrepresenting their preferences, either by changing the order of the preference lists or by announcing that some acceptable agents are unacceptable. Even the NRMP and School Choice Programs in NYC and Boston, while widely acknowledged as a model of successful matching programs, cannot rule out such incentives for strategic misrepresentation. Indeed, the possibility of such manipulation is mostly unavoidable. Whenever there is more than one stable matching, at least one agent can profitably misrepresent her preferences (Roth and Sotomayor, 1990), and the conditions under which a preference profile contains a unique stable matching seem to be quite restrictive (Eeckhout, 2000; Clark, 2006).[4] Thus, markets are likely to have agents with an incentive to manipulate a stable matching mechanism. In addition, Pittel (1989) shows that the number of stable matchings tends to increase as the number of participants becomes large. Accordingly, when market designers deal with large markets, concerns regarding strategic manipulation are heightened. As stable matching mechanisms are not incentive compatible, the mechanisms may be manipulated by participants, thereby not implementing the intended matchings. Moreover, each participant's decision may become hard to make since she needs to best respond to other agents' strategic manipulations.

We consider matching markets that each firm hires one worker, a model which is known as a one-to-one matching. We measure incentives to manipulate a stable matching mechanism by assuming that each firm-worker pair receives utilities, one for the firm and the other for the worker, which in turn determine ordinal preferences. In order to study the likelihood of an agent having a significant incentive to manipulate, we assume that utilities are randomly drawn from some underlying distributions. The key finding of this paper is that the proportion of participants who can potentially achieve a significant utility gain from manipulation vanishes as the market becomes large. This result holds both when each agent

[3] In fact, strategy-proofness is incompatible not only with stability but even with weaker conditions of Pareto efficiency and individual rationality (Alcalde and Barberà, 1994; Sönmez, 1999).

[4] It is an open question to characterize the complete set of preference profiles containing a unique stable matching.

knows the preferences of all other agents (complete information), and when an agent may not know the preferences of other agents (incomplete information). Given the tangible and intangible costs of strategic behavior in real life, we believe that this result may reconcile the success of stable matching mechanisms with the theoretical concerns about manipulability. In addition, based on this paper's finding, market designers may more confidently advise participants to submit their true preferences.

1.1.2 A Motivating Example

To understand the logic behind strategic manipulation, consider a simple labor market with three firms and three workers. We illustrate how, in such a situation, an agent can achieve a better partner by misrepresenting her preferences. In addition, we show that the best achievable partner from manipulation must be a partner in a stable matching under her true preferences.

Table 1.2 lists preferences of firms over workers, and of workers over firms which are known to all participants. For instance, firm 1 most prefers worker 3, followed by worker 1 and worker 2. Similarly, worker 1 most prefers firm 2, followed by firm 3 and firm 1. Under these preferences, there are two stable matchings: in one stable matching (marked by $\langle\cdot\rangle$), f_1, f_2, and f_3 are matched with w_1, w_2, and w_3, respectively; in the second stable matching (marked by $[\cdot]$), f_1, f_2, and f_3 are matched with w_2, w_1, and w_3, respectively.

$\mathbf{f_1}$	:	w_3	$\succ$	$\langle w_1\rangle$	$\succ$	$[w_2]$		$\mathbf{w_1}$	:	$[f_2]$	$\succ$	f_3	$\succ$	$\langle f_1\rangle$
$\mathbf{f_2}$	:	$\langle w_2\rangle$	$\succ$	$[w_1]$	$\succ$	w_3	,	$\mathbf{w_2}$	:	$[f_1]$	$\succ$	$\langle f_2\rangle$	$\succ$	f_3
$\mathbf{f_3}$	:	$\langle[w_3]\rangle$	$\succ$	w_1	$\succ$	w_2		$\mathbf{w_3}$	:	f_2	$\succ$	$\langle[f_3]\rangle$	$\succ$	f_1

Table 1.2: An example of a two-sided matching market with 3 firms and 3 workers.

Suppose that all agents submit their true preferences, and a stable matching mechanism produces the second stable matching marked by $[\cdot]$. In that case, suppose firm 1 misrepresent her preferences and announces that workers 3 and 1 are acceptable, but not worker 2. For the submitted preferences, there is a unique stable matching marked by $\langle\cdot\rangle$. The stable matching mechanism, which produces a stable matching for submitted preferences, will produce the matching marked by $\langle\cdot\rangle$. Ultimately, firm 1 is better off because firm 1 is matched with worker 1 rather than worker 2.

However, whichever preference list firm 1 submits, the firm will not be matched with worker 3. The pair (f_3, w_3) would otherwise block the matching. For instance, if f_1 declares that only w_3 is acceptable, then the only stable matching matches f_2 with w_2, and f_3 with w_3, and firm 1 will remain unmatched. More broadly, whenever a stable matching mechanism is applied, participants cannot be matched with a partner who is strictly preferred to all stable matching partners with respect to the initial preferences (Demange, Gale, and Sotomayor, 1987). Since participants are guaranteed to be matched with one of their stable matching partners, the gain from manipulation is bounded by the difference between the most and the least preferred stable matching partners. Based on the above observation, we mainly focus on the difference between the most and the least preferred stable matching partners.

1.1.3 Outline of the Paper

Prior to describing the model in detail, we briefly discuss the outline of the model, our main results, and the key idea behind the proof.

We consider a sequence of one-to-one matching markets, each of which has n firms and n workers. Preferences of firms over workers, or of workers over firms are generated by utilities, which are randomly drawn from some underlying distributions on $\mathbb{R}_+$.[5] We formulate utilities as the weighted sum of a *common-value* component and an *independent private-value* component. That is, when a firm f is matched with a worker w, the firm receives

$$U_{f,w} = \lambda\, U^o_w + (1-\lambda)\, \zeta_{f,w} \qquad (0 \leq \lambda \leq 1),$$

where U^o_w is the intrinsic value of w, which is common to all firms, and $\zeta_{f,w}$ is w's value as independently evaluated by firm f. In other words, any firm that is matched with a worker w receives the same common-value of the worker w, but receives distinct private-value of

[5] The only restrictions on distributions are bounded supports and some continuity conditions.

the worker w. We similarly define the utilities of workers.

The common-value component introduces a commonality of preferences, which is prevalent in real matching markets. In the entry-level labor market for doctors, for instance, the US News and World Report's annual rankings are often referred to as a guideline to the best hospitals. We also consider the pure private-value model ($\lambda = 0$) for theoretical reasons. In matching theory, commonality drives the uniqueness of stable matchings (Eeckhout, 2000; Clark, 2006), a situation in which no agent has an incentive to manipulate a stable matching mechanism (Roth and Sotomayor, 1990). If a preference profile has several stable matchings, commonality of preferences leads to smaller differences in utilities from stable matchings (Samet, 2011), so agents have less of an incentive to manipulate a stable matching mechanism. By including the pure private-value case in our model, we show that commonality may be beneficial, but is not necessary for incentive compatibility of stable matching mechanisms.

The main finding of the paper is that while agents in a large market typically have multiple stable partners, most agents are close to being indifferent among the all stable matching partners (Theorem 1).[6] We observed in the motivating example that when a stable matching mechanism is applied, the best an agent can achieve (by misrepresenting her preferences) is matching with her best stable matching partner with regard to the true preferences (Demange, Gale, and Sotomayor, 1987). As such, our main finding implies that when a stable matching mechanism is applied and all other agents reveal their true preferences, the expected proportion of agents who have an incentive to manipulate the mechanism vanishes as the market becomes large.

Furthermore, we identify an ϵ-Nash equilibrium in which most participants report their true preferences.[7] In a large market, a small proportion of agents may still have large incentives to manipulate a stable matching mechanism. Under the identified equilibrium, we

[6] The main theorem seems quite consistent with observations from real market applications. Pathak and Sönmez (2008) collect the data of students' preferences over schools in the new Boston school choice program, and show that the real market tends to have a very small number of stable matchings. (The preference data is reliable as truthfully revealing their preferences is a dominant strategy for students.) Both suggest that large matching markets tend to have small cores. In our theory of one-to-one matchings, we find small differences in utilities from stable matchings, whereas in the data from a many-to-one matching market, there is a small number of stable matchings.

[7] Under an ϵ-Nash equilibrium, agents are approximately best responding to other agents' strategies such that no one can gain more than ϵ by switching to an alternative strategy.

let those agents with significant incentives to manipulate do misrepresent their preferences. Nevertheless, the rest of participants still have no incentive to respond to such manipulations. More precisely, we show that for any $\epsilon > 0$ with high probability a large market has an ϵ-Nash equilibrium in which most participants reveal their true preferences (Corollary 2).

From a methodological standpoint, our paper is the first to introduce techniques from *random bipartite graph theory* to matching models. To prove the main theorem, we basically need to *count* the number of firms and workers satisfying certain conditions. The theory of random bipartite graphs provides techniques to count the likely numbers of firms and workers satisfying the specified conditions. More precisely, we draw a graph with a set of firms and workers whose common-values are above certain levels. We join each firm-worker pair by an edge if one of their independent private-values is significantly lower than the upper bound of the support. It turns out that every firm-worker pair where both the firm and the worker fail to achieve certain threshold levels of utility in a stable matching must be joined by an edge. Their private-values would otherwise both be so high that they would prefer each other to their current partners, and thus block the stable matching. For each realized graph, we consider the bi-partitioned subset of nodes, i.e. firms and workers, such that every pair of nodes, one from each partition, is joined by an edge. It is known that the possibility of having a relatively large such subset of nodes ultimately becomes infinitesimal as the initial set of nodes becomes large (Dawande, Keskinocak, Swaminathan, and Tayur, 2001). That is, in terms of the matching model, the set of firms and workers, whose common-values are high but who fail to achieve high levels of utility, will remain relatively small as the market becomes large.

This paper mainly focuses on the case of complete information, in which all participants are aware of all other agents' preferences. Nevertheless, we can extrapolate its findings to a market with incomplete information, in which each agent is partially informed about other agents' preferences. Various setups are conceivable: an agent may know (i) only her own utilities from matching with agents on the other side; (ii) her own utilities and common-values from matching with agents on the other side; (iii) her own utilities, common-values from matching with agents on the other side, and her own common-value to agents matching with her; or (iv) her own utilities and all agents' common-values. Regardless of the information structure, the key finding from the complete information case still holds with incomplete information. That is, most agents are ex-ante close to being indifferent among all stable matchings in a large market (Theorem 4). This is because with high probability agents are close to being indifferent among realized stable matching partners, which is this study's key finding in the context of complete information.

However, we do not find an equilibrium corresponding to the ϵ-Nash equilibrium under complete information. If some agents manipulate a stable matching mechanism based on the expected utility gain from manipulation, they may become worse off afterwards. This may, in turn, expand the differences between utilities generated by stable matchings to other agents. Consequently, other agents may misrepresent their preferences as a best response to the manipulation.

1.1.4 Related Literature

Strategic manipulability has been a major concern in market design. Hence, a number of studies have addressed the incentives to manipulate a stable matching mechanism (Roth and Peranson, 1999; Immorlica and Mahdian, 2005; Kojima and Pathak, 2009). These studies consider a particular stable matching mechanism, the worker-proposing Gale-Shapley algorithm, which implements a stable matching favorable to workers. As truthfully revealing their preferences is a dominant strategy for workers in this mechanism (Roth, 1982; Dubins and Freedman, 1981), the papers focus on firms' incentives to misrepresent their preferences.

Unlike the current paper, these studies assume that firms will manipulate a mechanism regardless of how much benefit the firms can obtain by so doing. In particular, a firm has no incentive to misrepresent its preferences if and only if it has a unique stable matching partner (Roth and Sotomayor, 1990). Thus, the primary goal is to find conditions on a preference profile in which most firms have a unique stable matching partner. As Roth and Peranson (1999) also point out, a crucial assumption is that agents on one side (say workers) consider only up to a fixed number of agents on the other side acceptable, even when the market size has become large. Under this assumption, Roth and Peranson, based on a computational analysis, show that the proportion of firms who have more than one stable matching partner converges to zero as the market becomes large. This convergence is theoretically proven by Immorlica and Mahdian and extended to the case of many-to-one matchings by Kojima and Pathak.

The main advantage of our approach is that we obtain non-manipulability of stable matching mechanisms as a pure property of market size, without resorting to the assumption of limited acceptability. In fact, the assumption of limited acceptability may lead to large market models that do not match basic features of real applications. Even with a weak commonality of preferences, the proportion of firms who are accepted by at least some workers may become small as the market becomes large. In this case, most firms do have a unique stable matching partner, but quite often the unique stable matching partner is only the firm itself: i.e. *a large proportion of agents remain unmatched.*

Figure 1.1 presents this phenomenon with simulations in which each worker considers only up to 30 most preferred firms acceptable. The utility of a firm is defined as $U_{f,w} = \lambda\, U^{o}_{w} + (1-\lambda)\, \zeta_{f,w}$, and the utility of a worker is similarly defined. The value of each component is drawn from the uniform distribution over $[0,1]$. Each graph depicts the proportion of firms (or workers) unmatched in stable matchings averaged over 10 repetitions.[8] Even with modest levels of commonality of preferences, the proportion of unmatched agents in stable matchings increases as the market becomes large. It is worth noting that these simulations are based on preferences generated, not by the previous studies' model, but by our own. Thus, the simulations do not directly represent features of the previous studies. However, we observe the similar effects of the limited acceptability assumption in simulations based on the previous studies' model. We provide additional simulation results in Appendix A.5.

Another strand of literature on large matching markets considers a market where a finite number of firms are matched with a continuum of workers (Azevedo and Leshno, 2011). It is shown that generically each market has a unique stable matching, to which the set of stable matchings in markets with large discrete workers converges. Based on this model,

[8] Given a preference profile, the set of unmatched agents is the same for all stable matchings (McVitie and Wilson, 1970).

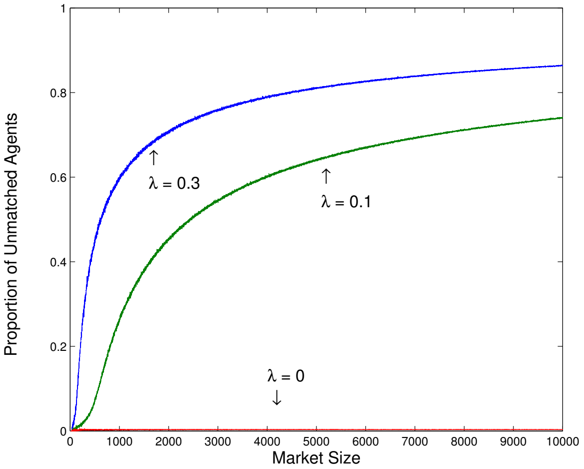


Figure 1.1: Proportion of agents unmatched in stable matchings.

Azevedo (2010) studies firms' incentives to manipulate capacities to hire workers. The paper also compares welfare effects between situations where each firm pays its employees equally (uniform wages) and those where each firm may pay different wages to different workers (personalized wages). While previous studies with fixed capacities suggest that a uniform wage may induce inefficient matching and compress workers' wages (Bulow and Levin, 2006; Crawford, 2008), if firms can manipulate their capacities, the uniform wage may produce higher welfare as they cause less capacity reduction.

The large market approach is not limited to the standard matching model. Ashlagi, Braverman, and Hassidim (2011) and Kojima, Pathak, and Roth (2010), for instance, develop models of large matching markets with couples. When couples are present, notwithstanding the concerns about strategic manipulation, a market does not necessarily have a stable matching (Roth, 1984). These studies show that the probability that a market with couples contains a stable matching converges to one as the market becomes large. Moreover, when a mechanism produces a stable matching with high probability, it is an approximate equilibrium for all participants to submit their true preferences. The results are based on the condition that the number of couples grows slower than the market size, with some additional regularity conditions.[9]

[9] Ashlagi, Braverman, and Hassidim (2011) considers a market where the number of positions offered

In the assignment problem of allocating a set of indivisible objects to agents, Kojima and Manea (2010) study incentives in the probabilistic serial mechanism (Bogomolnaia and Moulin, 2001). The probabilistic serial mechanism is proposed as a mechanism that improves the ex-ante efficiency of the random priority mechanism: All agents have higher chances of obtaining more preferred objects by using the probabilistic serial mechanism. However, while the random priority mechanism is strategy-proof, the probabilistic serial mechanism is not. Kojima and Manea show that for a fixed set of object types and an agent with a given utility function, if there is a sufficiently large number of copies of each object type, then reporting true preferences is a weakly dominant strategy for the agent.[10]

The rest of this paper is organized as follows. In Section 1.2, we introduce our model – a sequence of matching markets with random utilities. In Section 1.3, we state the main theorem informally and then formally, and find an equilibrium behavior which may reconcile the conflicting features of stable matching mechanisms. In Section 1.4, we illustrate the intuition of the proof using a random bipartite graph model. In Section 1.5, we study a market with incomplete information. The conclusion of the paper is provided in Section 1.6. All detailed proofs and simulation results are relegated to the appendix, which also includes definitions and related theorems of asymptotic statistics. Lastly, we extend the model to various directions in Appendix A.6.

1.2 Model

The model is based on the standard one-to-one matching model. We introduce latent utilities, which in turn generate ordinal preferences.

by firms exceeds the number of workers. Kojima, Pathak, and Roth (2010) inherits the assumption from Kojima and Pathak (2009) that agents on one side consider only up to a fixed number of agents on the other side acceptable.

[10] Che and Kojima (2010) show that the random assignments in the two mechanisms converge to each other as the number of copies of each object type goes to infinity. More generally, Liu and Pycia (2011) show that, including the two mechanisms, all sensible and asymptotically symmetric, strategy-proof, and ordinal efficient allocation mechanisms coincide asymptotically.

1.2.1 Standard Two-sided Matching Model (Roth and Sotomayor, 1990)

There are n firms and an equal number of workers. We denote the set of firms by F and the set of workers by W. Each firm has a strict preference list $\succ_f$ such as

$$\succ_f = w_1, w_2, w_3, f, \ldots, w_4.$$

This preference list indicates that w_1 is firm f's first choice, w_2 is the second choice, and that w_3 is the least preferred worker that the firm still wants to hire. We also write $w \succ_f w'$ to mean that f prefers w to w'. We call a worker w **acceptable** to f if $w \succ_f f$, otherwise we call the worker **unacceptable**. We define $\succ_w$ similarly for each $w \in W$, and call $\succ := ((\succ_f)_{f\in F}, (\succ_w)_{w\in W})$ **a preference profile**.

A **matching** μ is a function from the set $F \cup W$ onto itself such that (i) $\mu^2(x) = x$, (ii) if $\mu(f) \neq f$ then $\mu(f) \in W$, and (iii) if $\mu(w) \neq w$ then $\mu(w) \in F$. We say a matching μ is **individually rational** if each firm or worker is matched to an acceptable partner, or otherwise remains unmatched. For a given matching μ, a pair (f, w) is called a **blocking pair** if $w \succ_f \mu(f)$ and $f \succ_w \mu(w)$. We say a matching is **stable** if it is individually rational and has no blocking pair.

For two stable matchings μ and μ', we write $\mu \succeq_i \mu'$ if an agent i weakly prefers μ to μ': i.e. $\mu(i) \succ_i \mu'(i)$ or $\mu(i) = \mu'(i)$. We also write $\mu \succeq_F \mu'$ if every firm weakly prefers μ to μ': i.e $\mu(f) \succeq_f \mu'(f)$ for every $f \in F$. Similarly, we write $\mu \succeq_W \mu'$ if every worker weakly prefers μ to μ': i.e. $\mu(w) \succeq_w \mu'(w)$ for every $w \in W$. A stable matching μ_F is **firm-optimal** if every firm weakly prefers it to any other stable matching μ: i.e. $\mu_F \succeq_F \mu$. Similarly, a stable matching μ_W is **worker-optimal** if every worker weakly prefers it to any other stable matching μ: i.e. $\mu_W \succeq_W \mu$. It is known that every market instance has a firm-optimal stable matching μ_F and a worker-optimal stable matching μ_W (Gale and Shapley, 1962): i.e. for any stable matching μ, we have $\mu_F \succeq_F \mu$ and $\mu_W \succeq_W \mu$. Moreover if μ and μ' are both stable matchings, then $\mu \succeq_F \mu'$ if and only if $\mu' \succeq_W \mu$ (Knuth, 1976). Thus for any stable matching μ, it must be the case that $\mu \succeq_F \mu_W$ and $\mu \succeq_W \mu_F$.

With some abuse of notation, we let μ denote a function $\succ \longmapsto \mu(\succ)$ from the set of all preference profiles to the set of all matchings. We call the function μ a **matching mechanism**, and say that a mechanism μ is **stable** if $\mu(\succ)$ is a stable matching with respect to preference profile $\succ$. We also let μ_F and μ_W denote firm-optimal and worker-optimal stable matching mechanisms. A matching mechanism induces a game in which each agent $i \in F \cup W$ states her preference list $\succ_i$. If for all $\succ_i$ and $\succ_{-i}$,

$$\mu(\succ_i^*, \succ_{-i}) \ \succeq_i \ \mu(\succ_i, \succ_{-i}),$$

then we call $\succ_i^*$ a **dominant strategy** for the agent i. A mechanism μ is called **strategy-proof** if it is a dominant strategy for every agent to state her true preference list.

1.2.2 Random Utilities

In order to measure incentives to manipulate a stable matching mechanism, we assume that preferences are induced by underlying utilities. Moreover, in order to measure likely incentives, we assume that the utilities are drawn from some underlying probability distributions.

We represent utilities by $n \times n$ random matrices $U = [U_{f,w}]$ and $V = [V_{f,w}]$. When a firm f and a worker w match with one another, the firm f receives utility $U_{f,w}$ and the worker w receives utility $V_{f,w}$. We let u and v denote realized matrices of U and V. For each pair (f, w), utilities are defined as

$$\begin{aligned} U_{f,w} &= \lambda \, U_w^o + (1-\lambda)\, \zeta_{f,w} \qquad \text{and} \\ V_{f,w} &= \lambda \, V_f^o + (1-\lambda)\, \eta_{f,w} \qquad (0 \le \lambda \le 1). \end{aligned}$$

We call U_w^o and V_f^o *common-values*, and $\zeta_{f,w}$ and $\eta_{f,w}$ *independent private-values.*

Common-values are defined as random vectors

$$U^o := \langle U_w^o \rangle_{w \in W} \quad \text{and} \quad V^o := \langle V_f^o \rangle_{f \in F}.$$

Each U_w^o and V_f^o are drawn from distributions with positive density functions and with bounded supports in $\mathbb{R}_+$. Independent private-values are defined as $n \times n$ random matrices

$$\zeta := [\zeta_{f,w}] \quad \text{and} \quad \eta := [\eta_{f,w}].$$

Each $\zeta_{f,w}$ and $\eta_{f,w}$ are randomly drawn from continuous distributions with bounded supports in $\mathbb{R}_+$.[11] We assume that the utility of remaining unmatched is equal to 0.[12]

A random market is defined as a tuple $\langle F, W, U, V \rangle$, and a market instance is denoted by $\langle F, W, u, v \rangle$. Each firm f receives distinct utilities from different workers with probability 1. Thus for each $\langle F, W, u, v \rangle$, we can derive a strict preference list $\succ_f$ as

$$\succ_f = w, w', \ldots, w''$$

if and only if

$$u_{f,w} > u_{f,w'} > \cdots > u_{f,w''}.$$

We study properties of stable matchings in a sequence of random markets $\langle F_n, W_n, U_n, V_n \rangle_{n=1}^{\infty}$. The index n will be omitted whenever to do so does not lead to confusion.

The model includes both cases of a commonality of preferences ($\lambda > 0$) and pure private-values ($\lambda = 0$). The common-values introduce a commonality of preferences among firms over workers, and among workers over firms. When $\lambda > 0$, firms with high level of common-values tend to be ranked higher by workers, and vice versa. If $\lambda = 0$, all utilities are i.i.d, so a firm's ordering of workers are equally likely to be any permutation from the set of all permutations of n workers. Similarly, a worker's ordering of firms are equally likely to be any permutation from the set of all permutations of n firms.

In practice, commonality of preferences is prevalent. In the NRMP, some hospitals are considered prestigious and some doctors are considered very well-qualified. The common-value component provides a way of taking into account such commonality of preferences, while retaining the tractability of the model.

[11] In general ($\lambda > 0$), we can relax this assumption so that each pair of $\zeta_{f,w}$ and $\eta_{f,w}$ is jointly drawn from a continuous joint distribution with a bounded support in $\mathbb{R}_+^2$. In this setup, we can introduce a correlation between firms' preferences over workers and workers' preferences over firms.

[12] In terms of preferences induced by utilities, this assumption implies that all workers are acceptable to firms, and all firms are acceptable to workers.

Although the pure private-value case ($\lambda = 0$) hardly represents any real application, it is theoretically valuable to include it in our model. Commonality drives the uniqueness of stable matchings (Eeckhout, 2000; Clark, 2006), a condition in which no agent has an incentive to misrepresent her preferences in a stable matching mechanism (Roth and Sotomayor, 1990). Samet (2011) also proposes commonality as a source establishing a small core: the small difference between utilities from the stable matchings favorable to firms, and to workers. By including the pure private-value case in our model, we can highlight that non-manipulability of stable matching mechanisms is a property solely derived from market size. Commonality may contribute to, but is not necessary for, incentive compatibility of stable matching mechanisms.[13]

1.3 Main Results

We informally state the main theorem, and then restate it with formal expressions. Later, we find an equilibrium behavior of a game induced by a stable matching mechanism in which most agents reveal their true preferences.

1.3.1 Stable Matchings in Large Markets

We first show that, while agents in a large market typically have multiple stable matching partners, most agents are close to being indifferent among the stable matching partners.

Theorem *For every $\epsilon > 0$, the expected proportion of firms (and workers) who have less than ϵ differences between utilities from μ_F and μ_W, converges to one as the market becomes large.*

Corollary *For any positive cost of misrepresenting preferences, if other agents truthfully reveal their preferences, the expected proportion of agents who have no incentive to manipulate a stable matching mechanism converges to one as the market becomes large.*

[13] When preferences have a strong commonality, a stable matching mechanism may have a higher chance to fail by unraveling instead of strategic preference misrepresentation (Halaburda, 2010). In any case, our model includes all degrees of commonality of preferences.

It has been known that no stable matching mechanism is strategy-proof (Roth, 1982). For instance, when the worker-optimal matching mechanism (e.g. worker-proposing Gale-Shapley algorithm) is applied, although it is a dominant strategy for every worker to state her true preference list (Roth, 1982; Dubins and Freedman, 1981), there might be a firm which can become better off by misrepresenting its preference list. Noting that a matching mechanism is defined over all possible preference profiles, we may expect that a stable matching mechanism is not manipulable in most cases of preference profiles. Unfortunately, though, it turns out that whenever there is more than one stable matching, at least one agent can profitably misrepresent her preferences (Roth and Sotomayor, 1990), and the condition of a preference profile containing a unique stable matching seems to be quite restrictive (Eeckhout, 2000; Clark, 2006).

However, the gain by manipulation is bounded even when agents form a coalition and coordinate the members' strategic behavior. Not all firms in the coalition will prefer the new matching outcome to the firm-optimal stable matching with respect to the true preferences, and not all workers in the coalition will prefer the new matching outcome to the worker-optimal stable matching with respect to the true preferences (Demange, Gale, and Sotomayor, 1987). Formally, let $\succ$ be a true preference profile, and let $\succ'$ differ from $\succ$ in that some coalition S of firms and workers misstate their preferences. Then, there is no matching, stable under $\succ'$, which is strictly preferred to every stable matching under $\succ$ by all members of S. If a coalition consists of a single firm, then the best the firm can achieve is matching with the firm-optimal stable matching partner with respect to the true preferences. Likewise, the best a worker can achieve is matching with the worker-optimal stable matching partner. Since every firm and worker is guaranteed to be matched with a stable matching partner without any strategic manipulation, the gain by manipulation is bounded by the difference between utilities from the firm-optimal and the worker-optimal stable matching partners.

As such, the main theorem implies that agents in a large market are most likely to have only a slight utility gain by misrepresenting their preferences, given that all other agents reveal their true preferences. For any given cost of misrepresenting preferences, if a market is large, participants are most likely to find no incentive to manipulate a stable matching mechanism.

In order to see whether a real market is large enough to mitigate incentives to manipulate stable matching mechanisms, we simulate our model with a market size of 26,000, roughly the same size of the NRMP in 2011.[14] We generate firms' and workers' utilities from common-values and independent private-values, each of which is randomly drawn from the uniform distribution over $[0, 1]$. Table 1.3 presents the proportion of firms whose differences in utilities generated by stable matchings are less than 0.05 (upper table) and 0.01 (lower table). The results show that for reasonable degrees of commonality of preferences, the size of the NRMP is large enough such that most agents would not have a significant incentive to manipulate a stable matching mechanism.

λ	0.2	0.4	0.6	0.8
Result 1	97.41%	98.83%	99.39%	99.93%
Result 2	97.44%	98.79%	99.42%	99.92%
Result 3	97.43%	98.67%	99.47%	99.95%

(Differences in utilities < 0.05)

λ	0.2	0.4	0.6	0.8
Result 1	92.84%	96.64%	98.00%	99.44%
Result 2	93.04%	96.70%	98.10%	99.32%
Result 3	92.91%	96.52%	98.28%	99.48%

(Differences in utilities < 0.01)

Table 1.3: Proportions of firms with small differences in utilities (n=26,000)

Formal Statement Given a market instance $\langle F, W, u, v\rangle$ and a matching μ, we let $u_\mu(\cdot)$ and $v_\mu(\cdot)$ denote utilities from the matching outcome: i.e. $u_\mu(f) := u_{f,\mu(f)}$ and $v_\mu(w) := v_{\mu(w),w}$. For each $f \in F$, we define $\Delta(f; u, v)$ as the difference between utilities from firm-optimal and worker-optimal stable matching outcomes: i.e.

$$\Delta(f; u, v) := u_{\mu_F}(f) - u_{\mu_W}(f).$$

[14] In 2011, there were 30,589 active applicants and 26,158 positions offered by 4,235 programs. See `http://www.nrmp.org/data/resultsanddata2011.pdf` and `http://www.nrmp.org/res_match/about_res/impact.html`.

Then, for every $\epsilon > 0$, we have the set of firms whose utilities are within ϵ of one another for all stable matchings, which is denoted by

$$A^F(\epsilon; u, v) := \{f \in F \mid \Delta(f; u, v) < \epsilon\}.$$

The previous theorem is an informal statement of the following theorem. We have similar notations and a theorem for workers, which are omitted here.

Theorem 1. *For every $\epsilon > 0$,*

$$E\left[\frac{\left|A^F(\epsilon; U, V)\right|}{n}\right] \to 1, \quad \textit{as} \quad n \to \infty.$$

1.3.2 Equilibrium Analysis

Previously, we showed that most agents have no incentive to manipulate a stable matching mechanism as a market becomes large. However, the result requires the condition that all other participants reveal their true preferences. This condition is problematic since a small proportion of agents may still have large incentives to misrepresent their preferences. We may want to derive incentive compatibility as equilibrium behavior of a game induced by a stable matching mechanism.

In fact, the main theorem implies that with high probability a large market has a natural equilibrium in which most agents reveal their true preferences. We first state this finding as a corollary, and then describe appealing aspects of the equilibrium behavior and the intuition behind the proof.

Corollary 2. *For any $\epsilon, \delta, \theta > 0$, there exists N such that with probability at least $(1 - \delta)$ a market of size $n > N$ has an ϵ-Nash equilibrium in which $(1 - \theta)$ proportion of agents reveal their true preferences.*

This corollary is based on *simple* equilibrium behavior. Most agents simply reveal their true preferences. Agents misrepresenting their preferences use *truncation strategies*: an agent submits a preference list of the first k $(k < n)$ in the same order as her true preference list. Truncations are natural strategies. Agents do not need to carefully devise the order of the preference list. In addition, truncation strategies are undominated, or, in other words, have "a best response property" (Roth and Vande Vate, 1991). If a stable matching mechanism is applied, for any given submitted preferences by other agents, an agent always has a best response that is a truncation of her true preference list.[15]

For each market instance $\langle F, W, u, v \rangle$, we consider an ϵ-Nash equilibrium in which some (not necessarily all) agents, who have potential gains from manipulations larger than ϵ, submit truncations of their true preferences. If there exists a stable matching under the true preferences remaining individually rational under the announced preferences, then for all participants the difference between utilities from firm-optimal and worker-optimal stable matchings decreases. Specifically, let $\succ$ be a true preference profile and $\succ'$ differ from $\succ$ in that some coalition of firms and workers misstate their preferences using truncations. If there exists at least one matching μ stable under $\succ$ remaining individually rational under $\succ'$, then all stable matchings for $\succ'$ are also stable under $\succ$. Thus, truncations by some agents result in smaller differences in utilities from stable matchings for all participants.

This property follows because truncations do not create additional blocking pairs. If a matching μ, which is stable under $\succ$, remains individually rational under $\succ'$, then μ is indeed stable under $\succ'$ since no blocking pair has been generated by truncations. Noting that the set of unmatched agents is the same for all stable matchings (McVitie and Wilson (1970)), all participants are matched in stable matchings under $\succ'$.[16] Then, any stable matching μ' with regard to $\succ'$ is also stable under $\succ$. If (f, w) is a blocking pair of μ' with respect to $\succ$, then it would have been a blocking pair of μ' with respect to $\succ'$, which contradicts that μ' is stable under $\succ'$.

For any preference profile and for any coalition of participants, there exist truncations by members of the coalition such that at least one stable matching under true preferences

[15] Furthermore, when agents do not have complete information about the preference profile, truncation strategies require less information to manipulate a stable mechanism (Roth and Rothblum, 1999).

[16] Here, we use the condition that all participants are matched in stable matchings under $\succ$. If some agents are unmatched in stable matchings due to, for instance, unequal populations or unacceptable agents, we need an additional condition that agents would truncate their preferences only when truncations are strictly profitable. In particular, if an agent is unmatched in stable matchings under $\succ$, the agent will remain unmatched when she truncates her preference list. If these unmatched agents do not truncate their preference lists, then we obtain the same result: all stable matchings under $\succ'$ are stable under $\succ$, provided that there exists a stable matching under $\succ$ remaining individually rational under $\succ'$. The proof is easy to derive, and thus we omit it here.

remains individually rational, and those who truncate their preferences have no incentive to truncate further. Then, participants who initially have smaller than ϵ differences in utilities from stable matchings will have even less differences in utilities from stable matchings under the announced preferences. Thus, these participants have no incentive to respond to others' truncations, thereby submitting their true preferences. Lastly, Theorem 1 guarantees that most participants are the ones revealing their true preferences.[17]

1.4 Intuition Behind the Proof of Theorem 1

To prove Theorem 1, we take distinct approaches for the pure common-value case ($\lambda = 1$), the pure private-value case ($\lambda = 0$), and the general cases ($0 < \lambda < 1$).

For the pure common-value case ($\lambda = 1$), there exists a unique stable matching, so the theorem follows immediately. A stable matching sorts firms and workers such that a firm and a worker in the same rank will be matched with one another. Consider the firm-worker pair with the highest common-values. The pair must be matched in a stable matching. If it were otherwise, the firm would prefer the worker to its partner and the worker would prefer the firm to her partner, and thus they would form a blocking pair. By sequentially applying the same argument to pairs with the next highest common-values, we find that assortative matching is a unique stable matching.

For the pure private-value case ($\lambda = 0$), we still derive the theorem relatively easily from Pittel (1989). Pittel considers a model that is essentially the same as our pure private-value model ($\lambda = 0$), and analyzes the sum of each firm's partner's rank number in the worker-optimal stable matching.[18] When each firm ranks workers in order of preferences (i.e. the most preferred worker is ranked 1, the next worker is ranked 2, and so on), Pittel shows that the sum of the rank numbers of firms' partners in the worker-optimal stable matching is

[17] We use an equivalent statement of Theorem 1. Note that $|A^F(\epsilon; U, V)|/n$ is bounded above by 1 with probability 1. By using Theorem A.1.1 and Theorem A.1.2, we shall rewrite Theorem 1, written as convergence in mean, as the following convergence in probability: for any $\epsilon, \delta, \theta > 0$, there exists N such that

$$P\left(\frac{|A^F(\epsilon; U, V)|}{n} > 1 - \theta\right) > 1 - \delta, \quad \text{for every} \quad n > N.$$

[18] Pittel does not consider utilities, but a model with random preference profiles. As all preference profiles are equally likely to occur, though, the model is essentially the same as our pure private-value model ($\lambda = 0$).

asymptotically equal to $n^2 \log^{-1} n$. Then, the rank number of each firm is roughly $n \log^{-1} n$ on average. In turn, as we normalize the rank number by the market size n, the normalized average rank number is roughly equal to $\log^{-1} n$, converging to 0. As the private-values are randomly drawn from distributions with bounded supports, even the worst stable matching gives utilities asymptotically close to the upper bound. Therefore, all stable matchings yield only slightly different utilities.

For the general cases $(0 < \lambda < 1)$, however, the probability distribution over preference profiles becomes complicated and intractable. Accordingly, we directly analyze the asymptotic utilities rather than referring to the corresponding preference rank numbers. Basically, we want to count participants whose utilities from all stable matchings are slightly different from each other. We therefore need techniques of counting for which we use the bipartite graph theory. We interpret the set of firms and workers as a bi-partitioned set of nodes and draw a graph based on the realized utilities. Then, since the utilities are random, the theory of random bipartite graphs provides us with techniques to count the likely numbers of nodes, i.e. firms and workers, meeting specified conditions. Since the theory of random bipartite graphs has not been used before in the matching literature, we describe the techniques in greater depth in the following subsection.

We relegate detailed proofs for the cases of $\lambda = 0$ and $0 < \lambda < 1$ to Appendix A.2 and Appendix A.7, respectively.

1.4.1 A Random Bipartite Graph Model

A **graph** G is a pair (V, E), where V is a set called **nodes** and E is a set of unordered pairs (i, j) or (j, i) of $i, j \in V$ called **edges**. The nodes i and j are called the **endpoints** of (i, j). We say that a graph $G = (V, E)$ is **bipartite** if its node set V can be partitioned into two disjoint subsets V_1 and V_2 such that each of its edges has one endpoint in V_1 and the other in V_2. A **biclique** of a bipartite graph $G = (V_1 \cup V_2, E)$ is a set of nodes $U_1 \cup U_2$ such that $U_1 \subset V_1$, $U_2 \subset V_2$, and for all $u_1 \in U_1$ and $u_2 \in U_2$, $(u_1, u_2) \in E$. In other words, a biclique is a complete bipartite subgraph of G. We say that a biclique is **balanced** if $|U_1| = |U_2|$, and refer to a balanced biclique with the maximum number of nodes as **a**

maximum balanced biclique.

Given a partitioned set $V_1 \cup V_2$, we consider a random bipartite graph $G(V_1 \cup V_2, p)$. A bipartite graph $G = (V_1 \cup V_2, E)$ is constructed so that each pair of nodes, one in V_1 and the other in V_2, is included in E independently with probability p. We use the following theorem in the proof.

Theorem 3 (Dawande, Keskinocak, Swaminathan, and Tayur (2001))**.** *Consider a random bipartite graph $G(V_1 \cup V_2, p)$, where $0 < p < 1$ is a constant, $|V_1| = |V_2| = n$, and $\beta(n) = \log n / \log \frac{1}{p}$. If the maximal balanced biclique of this graph has size $B \times B$, then*

$$P\left(\beta(n) \leq B \leq 2\beta(n)\right) \to 1, \quad \textit{as} \quad n \to \infty.$$

1.4.2 Intuition of the Proof $(0 < \lambda < 1)$

Roughly stated, we observe that stable matchings become assortative-like matchings as a market becomes large: firms with higher common-values become more likely to match with workers with higher common-values. We illustrate this assortative-like feature of stable matchings by introducing a 3-tier market. In a 3-tier market, firms and workers are partitioned into three tiers, and endowed with tier-specific common-values. Then, most firms and workers in the same tier are matched with each other in assortative-like stable matchings. In this situation, the expected proportion of firms in tier-1, which fail to achieve high levels of utility converges to 0 as the market becomes large. We demonstrate how to use techniques from the theory of random bipartite graphs as we prove this observation formally.

In a 3-tier market, F is partitioned into F_1, F_2, and F_3; and W is partitioned into W_1, W_2, and W_3. For simplicity, we assume that all tiers are of equal size:

$$|F_k| = |W_k| = n/3 \qquad (k = 1, 2, 3).$$

If $f \in F_k$ and $w \in W_l$ are matched with one another, then they receive utilities

$$U_{f,w} = u_l^o + \zeta_{f,w} \quad \text{and} \quad V_{f,w} = v_k^o + \eta_{f,w}.$$

Common-values are uniquely determined by tiers such that

$$u_1^o > u_2^o > u_3^o \quad \text{and} \quad v_1^o > v_2^o > v_3^o.$$

Private-values, $\zeta_{f,w}$ and $\eta_{f,w}$, are randomly drawn from uniform distributions over $[0, \bar{u}]$ and $[0, \bar{v}]$, respectively. In other words, the firm receives tier-specific common-value corresponding to the worker's tier added to independent private-value, and the worker receives tier-specific common-value corresponding to the firm's tier added to independent private-value. We, without loss of generality, ignore λ and $(1 - \lambda)$ by incorporating the weights into the tier-specific common-values and the distributions of independent private-values.

We find an asymptotic lower bound on utilities that tier-1 firms receive in a stable matching mechanism. The lower bound is defined as the level arbitrarily close to the maximal utility that a firm can achieve by matching with tier-2 workers: i.e. $u_2^o + \bar{u} - \epsilon$. That is, firms in tier-1 achieve high levels of utility by levering on the existence of tier-2 workers. Although not necessarily being matched with tier-2 workers, firms in tier-1 would otherwise make blocking pairs with workers in tier-2. Formally, we define the set of tier-1 firms that fail to achieve the specified utility level in the worker-optimal stable matching as

$$\bar{F} := \{f \in F_1 \mid u_{\mu_W}(f) \leq u_2^o + \bar{u} - \epsilon\},$$

and show that

$$E\left[\frac{|\bar{F}|}{n/3}\right] \to 0, \quad \text{as} \quad n \to \infty.$$

Given realized private-values, we draw a bipartite graph with the set of firms in tier-1, and workers in tiers up to 2 (i.e. tier-1 and tier-2) as a bi-partitioned set of nodes (see the left figure in Figure 1.2). Each pair of $f \in F_1$ and $w \in W_1 \cup W_2$ is joined by an edge if and only if one of their private-values is low:

$$\zeta_{f,w} \leq \bar{u} - \epsilon \quad \text{or} \quad \eta_{f,w} \leq \bar{v} - (v_1^o - v_2^o).$$

We define the set of workers in tiers up to 2 matched with non tier-1 firms as

$$\bar{W} := \{w \in W_1 \cup W_2 \mid \mu_W(w) \notin F_1\}.$$

Then, $\bar{F} \cup \bar{W}$ is a biclique: i.e. every firm-worker pair from $\bar{F}$ and $\bar{W}$ is joined by an edge (as illustrated by the right figure in Figure 1.2).

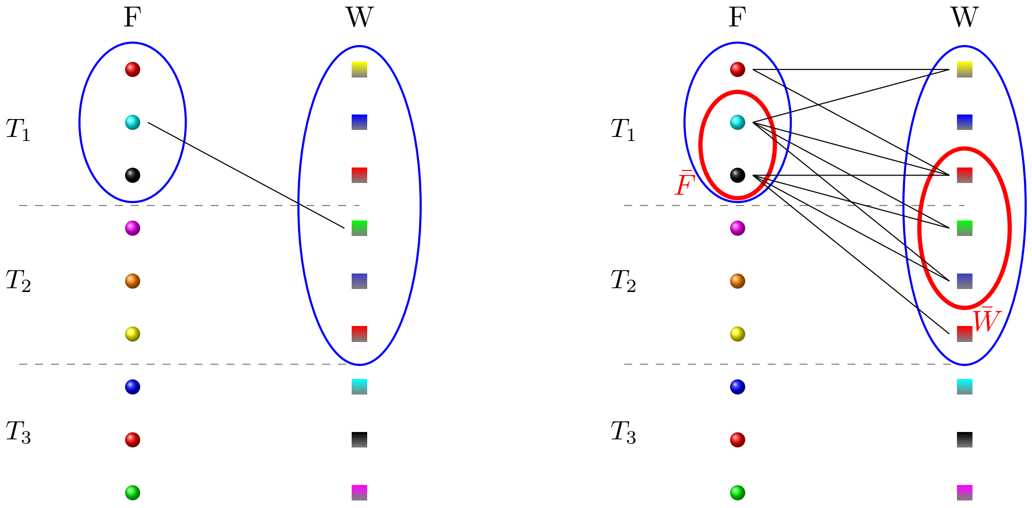


Figure 1.2: For each realized utility, we draw a bipartite graph with firms in tier-1 and workers in tiers up to 2 as the partitioned set of nodes (left). Firms in tier-1 receiving low utilities ($\bar{F}$) and workers in tiers up to 2 matched with non tier-1 firms ($\bar{W}$) form a biclique (right).

To see why $\bar{F} \cup \bar{W}$ is a biclique, suppose that $f \in \bar{F}$ and $w \in \bar{W}$ are not joined. Since $f \in \bar{F}$,

$$u_{\mu_W}(f) \le u_2^o + \bar{u} - \epsilon.$$

Since $w \in \bar{W}$, the worker is not matched with a tier-1 firm, and thus

$$v_{\mu_W}(w) \le v_2^o + \bar{v}.$$

That is, f and w mutually fail to achieve high levels of utility.

On the other hand, since they are not joined by an edge,

$$\zeta_{f,w} > \bar{u} - \epsilon \quad \text{and} \quad \eta_{f,w} > \bar{v} - (v_1^o - v_2^o),$$

and therefore

$$u_{f,w} > u_2^o + \bar{u} - \epsilon \quad \text{and} \quad v_{f,w} > v_1^o + \bar{v} - (v_1^o - v_2^o) = v_2^o + \bar{v}.$$

In other words, the firm-worker pair's private-values are mutually so high that they would have achieved high utilities by making a blocking pair. This contradicts that μ_W is a stable matching.

This construction of a bipartite graph fits into a random bipartite graph model. Given that the tier-structure specifies a bi-partitioned set of nodes, we draw a bipartite graph based on the realized private-values. Since the private-values are i.i.d, each firm-worker pair is joined by an edge independently and with an identical probability. By Theorem 3, if the bi-partitioned set of nodes has a size on the order of n, and each pair of nodes is joined by an edge independently with a fixed probability, then the maximum balanced biclique has a size on the order of $\log(n)$ with a sequence of probabilities converging to 1 as n gets large. In addition, $\bar{W}$ contains at least $n/3$ workers, since there are $2n/3$ workers in tiers up to 2, but only $n/3$ firms in tier-1: i.e. $\bar{W}$ has a size on the order of n. Therefore, $\bar{F}$ must have a size, at most, on the order of $\log(n)$ with a sequence of probabilities converging to 1. The biclique $\bar{F} \cup \bar{W}$ would otherwise contain a balanced biclique with a size bigger than on the order of $\log(n)$, violating the Theorem 3. Lastly, $E\left[\frac{|\bar{F}|}{n/3}\right] \to 0$ follows immediately from $\log(n)/n \to 0$.

For the main theorem (without tier structure), we begin the proof by partitioning the supports of distributions for common-values. Suppose the common-values are drawn from the uniform distribution over $[0, 1]$. We partition the unit interval into K subintervals with equal lengths. Workers and firms are, in turn, grouped into tiers where firms or workers in the same tier have common-values in the same subinterval. Basically, we continue the proof as if we have a model with a finite number K of tiers. The tiers, though, need to be handled with care. This time, because the common-values are random, the tier structure is random. Moreover, agents in adjacent tiers may have arbitrarily close common-values.

As we increase the number of partitions K, the asymptotic lower bound on the utilities of firms in tier-k becomes close to the maximal utility achievable by matching with a worker in tier-k. With a similar exercise, we find an asymptotic lower bound on utilities of workers in each tier. Then, workers in tiers significantly higher than k are most likely to match with firms in tiers higher than k. This assortative-like feature of stable matchings induces an asymptotic upper bound on utilities of tier-k firms. As we finely partition the supports of the distributions of common-values, the differences in the common-values of firms or workers in similar tiers become slightly distinct from each other. Therefore, the asymptotic upper bound on utilities of firms in tier-k also becomes close to the maximal utility achievable by matching with a worker in tier-k. That is, we can find an asymptotic lower bound and an asymptotic upper bound, which are arbitrarily close to each other.

1.5 Market with Incomplete Information

We have so far considered a market with complete information. Agents are assumed to be able to assess the exact gain by misrepresenting preferences. It is a strong assumption, especially when we consider large markets. More realistically, we may want to consider a market with incomplete information, where each agent is only partially informed about the preferences of other participants. Nevertheless, we have mainly focused on the case of complete information since we can extrapolate its findings to show that the incentive to misrepresent preferences vanishes under incomplete information.

In relaxing the complete information assumption, we may consider various information structures. Each agent may know only the probability distributions in addition to either (i) her own utilities; (ii) her own utilities and the common-values of the other side; (iii) her own utilities, the common-values of the other side, and her own common-value evaluated by the other side; or (iv) her own utilities and all agents' common-values. The following results in the context of incomplete information correspond to the main theorem and its direct corollary for the model with complete information. As before, we first state the theorem informally, and then restate it with formal expressions.

Theorem *Regardless of information structure and for every $\epsilon > 0$, the expected proportion of firms (and workers) who have less than ϵ expected differences between utilities from μ_F*

and μ_W, converges to one as the market becomes large.

Corollary *For any positive cost of misrepresenting preferences, if other agents truthfully reveal their preferences, the expected proportion of agents who have no incentive to manipulate a stable matching mechanism converges to one as the market becomes large.*

The intuition behind the theorem is clear. An expectation is a convex combination of all realizations. The expected difference between utilities from firm-optimal and worker-optimal stable matchings under incomplete information is simply a convex combination of the differences between utilities from the two stable matchings in all realized market instances. The differences between utilities are most likely to be insignificant (Theorem 1). Therefore, the expected difference in utilities is most likely to be negligible as well. We relegate the detailed proofs to Appendix A.4.

There are two advantages of showing the result in the context of complete information first, and then deriving the same result in the context of incomplete information. First, the results are robust to the information structure. The intuition of showing the results with incomplete information by using convex combinations remains valid regardless of the details of the information structure. Secondly, we can stress that non-manipulability of stable matching mechanisms is a property of the two-sided matching market itself, rather than stemming from insufficient information to manipulate the mechanism. Even when an agent can obtain complete knowledge of a preference profile at a small cost, it is not worth incurring that cost since the gain from manipulation will be small.

Formal Statement Let Π_f denote what f knows about a preference profile, and let π_f denote its realization. Then, the various incomplete information structures are denoted by (i) $\Pi_f = \langle U_{f,w} \rangle_{w \in W}$; (ii) $\Pi_f = \langle U_{f,w}, U^o_w \rangle_{w \in W}$; (iii) $\Pi_f = \langle U_{f,w}, U^o_w \rangle_{w \in W} \cup \{V^o_f\}$; and (iv) $\Pi_f = \langle U_{f,w}, U^o_w \rangle_{w \in W} \cup \langle V^o_{f'} \rangle_{f' \in F}$. Given a market instance $\langle F, W, u, v \rangle$, we define $\Delta_E(f; u, v)$ as the expected difference between utilities from firm-optimal and worker-optimal stable matchings conditioned on π_f. That is,

$$\Delta_E(f; u, v) := E_{U,V} \left[u_{\mu_F}(f) - u_{\mu_W}(f) \mid \pi_f \right],$$

where the expectation is applied to firm-optimal and worker-optimal stable matchings. For every $\epsilon > 0$, we correspondingly have the set of firms, whose expected differences in utilities from all stable matchings are less than ϵ, which is denoted by

$$A_E^F(\epsilon; u, v) := \{f \in F \mid \Delta_E(f; u, v) < \epsilon\}.$$

The previous theorem is an informal statement of the following theorem. We have similar notations and a theorem for workers, which are omitted here.

Theorem 4. *For any given information structure and for every $\epsilon > 0$,*

$$E\left[\frac{\left|A_E^F(\epsilon; U, V)\right|}{n}\right] \to 1, \quad \textit{as} \quad n \to \infty.$$

Equilibrium Analysis Unfortunately, we do not obtain an equilibrium corresponding to the ϵ-Nash equilibrium in the context of complete information by using convex combinations. The obstacle to obtaining an equilibrium is that truncations by some agents may *increase* the differences in utilities generated by stable matchings for other participants. When preferences are known to all participants, truncations can preserve a stable matching under true preferences as individually rational under the announced preferences. The following example shows that this condition is necessary for truncations by some agents to decrease the differences in utilities from stable matchings for other participants.

$\mathbf{f_1}$:	$\langle w_1 \rangle$	$\succ$	w_2	$\succ$	w_3	$\mathbf{w_1}$:	f_2	$\succ$	$\langle f_1 \rangle$	$\succ$	f_3
$\mathbf{f_2}$:	$\langle w_2 \rangle$	$\succ$	w_3	$\succ$	w_1	$\mathbf{w_2}$:	f_1	$\succ$	$\langle f_2 \rangle$	$\succ$	f_3
$\mathbf{f_3}$:	w_1	$\succ$	w_2	$\succ$	$\langle w_3 \rangle$	$\mathbf{w_3}$:	f_1	$\succ$	f_2	$\succ$	$\langle f_3 \rangle$

,

$\mathbf{f_1}$:	$\langle w_1 \rangle$	$\succ$	$[w_2]$	$\succ$	w_3	$\mathbf{w_1}$:	$[f_2]$	$\succ$	$\langle f_1 \rangle$	$\succ$	f_3
$\mathbf{f_2}$:	$\langle w_2 \rangle$	$\succ$	w_3	$\succ$	$[w_1]$	$\mathbf{w_2}$:	$[f_1]$	$\succ$	$\langle f_2 \rangle$	$\succ$	f_3
$\mathbf{f_3}$:	w_1	$\succ$	w_2			$\mathbf{w_3}$:	f_1				

Table 1.4: True preferences (upper) and their truncations (lower).

Table 1.4 lists true preferences of firms and workers (upper tables) and their truncations (lower tables). In the example, there is a unique stable matching (marked by $\langle\cdot\rangle$) under the true preferences. When f_3 and w_3 truncate their preferences, however, there are two stable matchings (marked by $\langle\cdot\rangle$ and $[\cdot]$). If some agents announce that all stable matching partners are unacceptable, other agents may have larger differences in utilities from all stable matchings.

Given incomplete information of a preference profile, an agent may submit a truncation of her true preference list based on the expected utility gain by manipulation. She may then remain unmatched afterwords depending on the realized preference profile. In this case, truncations may expand differences in utilities from stable matchings of other participants. Although most agents initially have small differences in utilities from stable matchings, participants may want to misrepresent their preferences as a best response to other agents' truncations.

1.6 Conclusions

This paper demonstrates an asymptotic similarity of stable matchings as the number of participants becomes large. Our measure of similarity is based on utilities, by which ordinal preferences are determined. As the utilities are drawn from some underlying probability distributions, one can analyze the likely differences in utilities from all stable matchings. We show that the expected proportion of firms and workers who are close to being indifferent among all stable partners converges to one as the market becomes large.

The result also implies that the expected proportion of agents who have a significant incentive to manipulate the mechanism vanishes in large markets. This is because the gain from manipulation of a stable matching mechanism is bounded above by the difference between utilities from the firm-optimal and the worker-optimal stable matchings. In addition, we show that with high probability a large market has an ϵ-Nash equilibrium in which most agents reveal their true preferences. We prove our results using techniques from the theory of random bipartite graphs, which is a new approach in the matching literature.

This paper is one of many recent studies exploring how the popularly used matching mechanisms really work in practice. It is essential to have a better understanding of stable matching mechanisms as market design applications expand from the NRMP and the School Choice Programs to many other markets, including dental residencies, various medical specialty matching programs, and labor markets for law clerks. Of particular relevance here is the fact that market designers are hoping to investigate the desirability of a clearinghouse in the market for economics Ph.D.s (Coles, Cawley, Levine, Niederle, Roth, and Siegfried, 2010). As such, understanding stable matching mechanisms in real applications becomes not only a market designers' question in theory, but is of concrete interest for economists in general.

Chapter 2

Plea Bargaining: On The Selection of Jury Trials

2.1 Overview

2.1.1 Introduction

Plea bargaining is a pre-trial stage in which a defendant is allowed to plead guilty. Considering what he would receive if he was convicted after a jury trial, a defendant pleads guilty primarily in exchange for a lesser charge.[1] Plea bargaining is prevalent in U.S. criminal court. Amongst the 89.7% convictions out of 83,391 cases in federal courts in 2004, 96% were achieved through plea bargaining, and the rate increased from 87% in 1990 to 96% in 2004 for felony offenses.[2]

The fact that the vast majority of cases end in plea bargaining may lead one to suspect that trials are not important. The current paper certifies that such a conclusion is inaccurate; plea bargaining and jury trials closely interact with each other. Innocent defendants have less incentive to plead guilty, and jurors incorporate this selection bias into their verdict. Conversely, although most cases are settled before jury trials begin, participants in plea bargains anticipate possible outcomes of jury trials in the event that they fail to reach an agreement. In this sense, the primary role of a jury trial is to allocate bargaining power to participants in the plea bargain.[3]

[1] In this paper, prosecutors and defendants are all referred to as male, and jurors are all referred to as female.

[2] See Table 4.2 in Compendium of Federal Justice Statistics, 2004, U.S. Department of Justice, Bureau of Justice Statistics, available online at: *http://bjs.ojp.usdoj.gov/content/pub/pdf/cfjs04.pdf*.

[3] Mnookin and Kornhauser (1979) call this effect, "Bargaining in the shadow of the law."

The interaction between plea bargaining and a jury trial is a challenging issue for legal scholars who want to evaluate various institutions in a criminal court system. A model of either plea bargaining or a jury trial often fails to capture the real dynamics; when defendants and prosecutors actively participate in pre-trial stages, the implications of a jury trial model may not be directly applicable to the entire court process. Similarly, a separate empirical analysis undertakes endogeneity problems. Cases in jury trials, for instance, may tell us how the jury delivers verdicts for those cases, but they are silent on how institutional changes in the trial affect the cases going to trial.[4]

The current paper, building on the standard strategic voting model, develops a model of the criminal court process unifying plea bargaining and a jury trial. We first show that plea bargaining influences the jurors' (identical) belief about the proportion of guilty defendants, and consequently jurors may vote as if they have the prosecutor's preferences. Based on Feddersen and Pesendorfer (1998), we also study different voting institutions in trial stage, and find that inferiority of the unanimity rule persists with the addition of plea bargaining.

In detail, a judicial process starts with a prosecutor indicting a defendant, who is either guilty or innocent with equal ex-ante probabilities. Given the level of just punishment for the charge, the prosecutor initiates a plea bargain by making a take-it-or-leave-it punishment offer to the defendant. If the defendant pleads guilty, then the case terminates with the offered punishment; otherwise, a jury trial follows. In a jury trial, each juror receives either a guilty or an innocent private signal during the testimonies, and votes either for conviction or acquittal. If a super-majority of jurors vote for conviction (such as two-thirds majority), the jury returns a verdict of guilty, and the defendant receives the original just punishment; otherwise, the jury acquits the defendant. The prosecutor and jurors have distinct preferences over mistakenly delivered (or undelivered) punishments to innocent defendants (or guilty defendants).[5]

We first show that, by internalizing plea bargaining into their belief, jurors may vote

[4] Priest and Klein (1984) first raise such challenges in the context of civil court.

[5] In this paper, a prosecutor may not single-mindedly pursue convictions, ignoring possible convictions of innocent defendants. Instead, we consider how different prosecutor's preferences affect court performance. This assumption is justified on realistic grounds. In practice, mismanaged cases may later become public, and such exposure will affect a prosecutor's future career. Even a self-interested prosecutor will be concerned with false prosecutions.

as if they have the prosecutor's preferences. While the prosecutor controls the punishment level of guilty pleas, the optimal level is ultimately determined by how it will influence jurors' behavior. This is because the ex-ante punishment levels (i.e. the expected punishment level upon pleading guilty) are eventually determined in equilibrium by the conviction probabilities in the jury trial.

To see the intuition, consider the following lines of reasoning. If the plea bargain offer is acceptable for the 'guilty' defendants, compared to the jury trial outcome, guilty defendants will plead guilty. Jurors subsequently update their belief, accounting for the lower proportion of guilty defendants arriving at jury trials. Accordingly, conviction probabilities are lowered, and this feeds back to plea bargaining. The previously acceptable offer will become *un*-acceptable for 'guilty' defendants. On the other hand, if the bargain offer is *un*-acceptable, the opposite story follows. 'Guilty' defendants will plead not guilty. As the jurors believe that a higher proportion of defendants who come to trial are guilty, the jurors tend to vote for conviction. When this occurs, the bargain offer, previously unacceptable, becomes now acceptable for the 'guilty' defendants. Thus, in equilibrium guilty defendants will be indifferent between receiving a guilty plea punishment or undergoing a jury trial. As a result, the ex-ante punishment for 'guilty' defendants will be equal to the expected punishment in a jury trial. Meanwhile, 'innocent' defendants are less likely to be convicted in trial than guilty defendants. When guilty defendants are indifferent between pleading guilty and not guilty, 'innocent' defendants are better off pleading not guilty and going to trial. Consequently, the ex-ante punishment for innocent defendants is also determined by the conviction probabilities in the jury trial.

The prosecutor chooses a plea offer such that its effects on jurors' beliefs render the ideal levels of conviction probabilities. The prosecutor cannot force a particular voting behavior on jurors, who will be best responding. Instead, the jurors' voting behavior that is ideal for the prosecutor will be induced when the jurors' preference combined with the altered belief coincide with the prosecutor's preference. For instance, suppose the prosecutor cares more than the jurors about mistakenly delivering punishment to innocent defendants. As the prosecutor lowers guilty plea charges, a higher proportion of guilty defendants plead guilty, and a defendant in a jury trial is more likely to be innocent. Consequently, jurors are more careful when voting to avoid mistakes of convicting innocent defendants, and the influenced jurors' behavior follows the prosecutor's preference.

However, such influence is possible only in one direction: leading jurors to vote more frequently for acquittal. Because guilty defendants are more likely to take the bargain offer, plea bargaining can only decrease the proportion of guilty defendants in trial. When the prosecutor cares less about convicting innocent defendants, and is more averse to acquitting guilty defendants, plea bargaining is of no use to the prosecutor.

The combined model of plea bargaining and a jury trial allows us to re-examine some of the implications derived from the classical strategic voting literature. In particular, we revisit the comparison of two voting mechanisms, the unanimity rule and arbitrary super-majority rules, which are studied in Feddersen and Pesendorfer (1998). Feddersen and Pesendorfer find that the unanimity rule is inferior in terms of the probabilities of convicting innocent defendants and acquitting guilty defendants. If the rule is unanimous, the probabilities do not vanish as the number of jurors grows, whereas the probabilities vanish under any non-unanimous rule. The results in our paper suggest that jurors' voting behavior resembles the voting behavior in the separate jury model, though it may reflect the prosecutor's preference. Therefore, from the viewpoint of expected punishments either by plea bargaining or a jury trial, inferiority of the unanimity rule persists with the addition of plea bargaining.

Note that the game proposed in this paper is effectively that of signaling. While previous literature mainly views plea bargaining as an instrument to save trial costs (see Grossman and Katz (1983); Reinganum (1988)), we intentionally ignore all costs in order to highlight the signaling effect.[6] A defendant, as a *sender*, signals his type by pleading either guilty or not guilty. Afterwards the jurors, as *receivers*, update their belief on the sender's type and determine conviction probabilities. From the prosecutor's viewpoint, plea bargaining allows the court to screen out some guilty defendants before going to a jury trial. Since the accused know whether they are guilty, plea bargaining serves as a self-selection mechanism.

[6] Not only are explicit costs such as time and effort excluded, we also assume that prosecutors and defendants are risk neutral. They bear no cost of uncertainty from a jury verdict.

As such, plea bargaining may contribute to the accuracy of the jury trial, on which the entire court process hinges.

2.1.2 Related Literature

Priest and Klein (1984) is one of the studies closest to our paper, as they clarify the relationship between litigation behavior and jurors' behavior in the jury trial. The set of disputes settled and the set litigated are not necessarily the same. Their important assumption is that the potential litigants produce rational estimates of the likely decision by affecting the belief of the jurors. As in our paper, Priest and Klein consider interactions between the pre-trial process and the jury trial. However, while Priest and Klein informally model how biased jurors' belief affects the jury decision, we explicitly capture the dynamic by employing a strategic voting model.

Collective decision-making under uncertainty is first studied in Condorcet (1785). Assuming two possible true states, Condorcet models a situation in which a group of people, each of whom is imperfectly and privately informed, makes a decision by voting for one alternative. Condorcet shows that the group can more efficiently aggregate private information with simple majority rule than if each member acts as a dictator.

The Condorcet theorem assumes that each juror votes by following her private information. However, a juror's vote affects a group decision only when that juror is pivotal. A strategic juror incorporates this fact in her voting decision, and in some cases her pivotality convinces her to follow other jurors' votes against her private information (see Austen-Smith and Banks (1996); Feddersen and Pesendorfer (1996)). Feddersen and Pesendorfer (1998) apply the strategic voting behavior to jury trials, and find inferiority of the unanimity rule. The current research departs from Feddersen and Pesendorfer (1998) by including plea bargaining.[7]

Much of the literature on plea bargaining approaches the process via a 'bargaining' model

[7] Although we adopt Feddersen and Pesendorfer (1998) as a benchmark, different voting institutions can be applied in the jury trial stage. Some examples from the literature include Coughlan (2000); Austen-Smith and Feddersen (2005, 2006), and Gerardi and Yariv (2007) studying jury deliberation. Accordingly, as the model of jury trial process changes, the results on the voting rule comparison in our model may change. For experimental tests on jury deliberation, see Guarnaschelli, McKelvey, and Palfrey (2000) and Goeree and Yariv (Forthcoming).

(for a brief summary, see, e.g., Cooter and Rubinfeld (1989)). A jury trial contains explicit costs, time, and effort; if participants in a plea bargain do not want to bear additional risks, uncertainty regarding trial outcomes is an additional cost. Given such costs, participants in the plea bargain phase can share a surplus if they reach an agreement. This surplus division is a 'bargaining' problem. A typical model allows either a prosecutor, a defendant, or both to make bargaining offers. Prosecutors know the deliverable punishments of the crime in trial, while the defendant knows whether he is guilty. It is undeniable that plea bargaining initially becomes popular as a way of avoiding jury trial costs.[8] However, what we focus on in this paper are the welfare effects of plea bargaining due to factors other than trial costs, a subject that has received less attention.

Grossman and Katz (1983) show that plea bargaining serves as an insurance and a screening device. As insurance, plea bargaining protects innocent defendants and society against cases where a trial produces incorrect findings and delivers severe punishments. Although innocent defendants may falsely plead guilty due to the threat of conviction, the sentence will be lenient in such cases. As a screening device, plea bargains sort guilty and innocent defendants like a self-selection mechanism. Since the mechanism ensures that violators of the law are indeed punished, it may contribute to the accuracy of the legal system. The first role is irrelevant to our model, since we assume that prosecutors and defendants are risk neutral, and consequently need no insurance. The second role shares the same motivation as ours. In contrast to the current paper, Grossman and Katz (1983) does not consider interactions between plea bargaining and the jury trial. They assume that plea bargaining is a screening device affecting, but never affected by, the jury trial.

2.2 The Model

There are three types of agents in a criminal court process: a prosecutor, a defendant, and jurors. The process begins with a prosecutor indicting a suspect on a charge. We normalize the potential punishment to be equal to 1 and assume that the defendant is either guilty (G) or innocent (I) with equal probabilities. We consider the following timed process, composed

[8] For the historical background of plea bargaining, see, e.g., Rabe and Champion (2002, p. 306 - 308).

of two phases:

At t=1, a plea bargain occurs.

The prosecutor makes a take-it-or-leave-it plea bargain offer, $\theta \in [0,1]$ level of punishment. The defendant pleads either *guilty* or *not guilty.* If the defendant pleads guilty, the case terminates and the punishment θ is delivered. Otherwise, the plea bargain is withdrawn, and the case proceeds to the second phase described below.

At t=2, a jury trial occurs.

A jury consists of n ($n > 1$) jurors and a voting rule $\hat{k}$ ($1 \leq \hat{k} \leq n$). Each juror receives a private signal g or i, which is positively correlated with the true states G or I, as given by

$$Pr[g|G] = Pr[i|I] = p, \quad Pr[i|G] = Pr[g|I] = 1 - p \tag{2.1}$$

where $p \in (.5, 1)$; a juror has a probability p of receiving a correct signal, and a probability $1 - p$ of receiving an incorrect signal.[9]

The jury reaches a decision by casting votes simultaneously. Each juror votes for either conviction or acquittal. If the number of conviction votes is larger than or equal to the voting rule $\hat{k}$, the defendant is convicted (C). Otherwise, the defendant is acquitted (A). We call a rule requiring $\hat{k} = n$ votes for conviction *the unanimity rule*, and others *general super-majority rules.*

Each type of agents has a utility function defined as follows:

- A defendant:

Utility changes negatively by the amount of punishment: -1 if he is convicted, 0 if he is acquitted, and $-\theta$ if he pleads guilty. A defendant is assumed to be risk neutral.[10]

[9] During the testimonies by the witnesses, each juror may have a different interpretation due to her personal background. The private signal (g or i) captures such interpretation.

[10] If a defendant perceives that he will be convicted with probability s, then the ex-ante utility of going to trial is $-s \cdot 1 - (1 - s) \cdot 0$.

- Jurors:

  We normalize the utility of correct judicial decisions such that $u[C|G] = u[A|I] = 0$. Given this normalization, convicting innocent defendants or acquitting guilty defendants incur utility losses, $u[C|I] = -q$ and $u[A|G] = -(1-q)$, respectively. We assume that $q \in [.5, 1)$, and term q as "the threshold level of reasonable doubt." [11, 12]

- A prosecutor:

  The prosecutor has a preference defined on $[0, 1] \times \{\text{G, I}\}$. Much like the jurors' utilities, when a punishment $h \in [0, 1]$ is delivered to a defendant, the prosecutor's utility is given by

  $$v[h|I] = -q'\,h \quad , \quad v[h|G] = -(1-q')(1-h)$$

  where $q' \in [0, 1]$. The prosecutor loses utility if punishments are delivered to innocent defendants, or guilty defendants avoid their just punishments.

Figure 2.1 summarizes the timing of the model: (i) A prosecutor offers θ in a plea bargain and a defendant pleads either guilty or not guilty. (ii) If the defendant pleads guilty, a judge respects the bargain and pronounces sentence θ, and the case terminates. If the defendant pleads not guilty, the case goes to a jury trial. (iii) The jury determines whether to convict or acquit.

We denote by ϕ_G the probability that a guilty defendant pleads guilty; ϕ_I is defined similarly for an innocent defendant. Jurors have an identical belief π that the defendant is guilty conditional on the case proceeding to a jury trial. For each level of belief π, a pair (σ_g^j, σ_i^j) in $[0, 1] \times [0, 1]$ represents a strategy of juror j. Juror j votes for conviction with probability σ_g^j when she receives a signal g, and she votes for conviction with probability

[11] Feddersen and Pesendorfer (1998) term q as "the threshold level of reasonable doubt," from the following motivation. Suppose a juror believes that the defendant is guilty with probability $\tilde{q}$. The expected utility from a guilty verdict, $-q(1-\tilde{q})$, is greater than or equal to the expected utility of an innocent verdict, $-(1-q)\tilde{q}$, if and only if $\tilde{q} \geq q$. Therefore, when jurors vote for conviction, they use q as the threshold level of belief that the defendant is guilty.

[12] We can easily allow $q < 0.5$, and the analysis in this paper is qualitatively intact. However, we focus on the case of $q \geq 0.5$ for simplicity, since $q < 0.5$ requires additional assumptions to ensure that jurors are more likely to vote for conviction when they receive signal g.

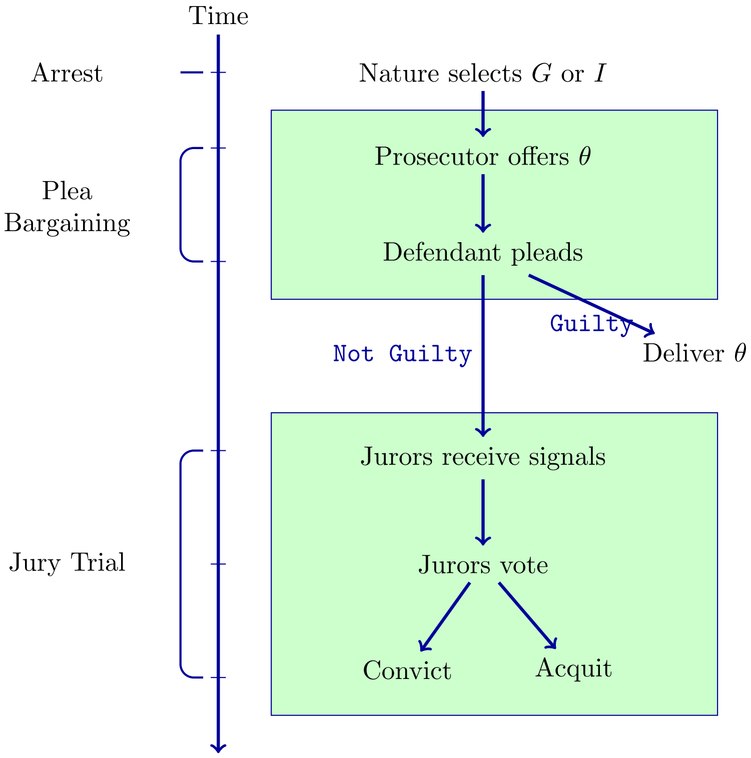


Figure 2.1: A criminal court process.

σ_i^j if the signal is i. Apparently, a defendant's strategy (ϕ_G and ϕ_I) is a function defined on θ, and jurors' strategies (σ_g^j, σ_i^j) are functions defined on π. We omit the arguments of strategies where no confusion arises.

We find a Perfect Bayesian Equilibrium with additional refinements: one in jurors' voting behavior and the other in jurors' belief. For jury trials, we consider *symmetric equilibrium voting behavior* in which all jurors adopt the same strategy. Accordingly, a symmetric strategy profile is denoted as (σ_g, σ_i), without specifying a particular juror.[13] We then find a symmetric voting behavior which gives all jurors the highest expected payoff. Since all jurors have the same preference over judicial decisions, this is a natural way of refining the symmetric voting behavior. We call this refined behavior the *most efficient symmetric equilibrium voting behavior*, or succinctly the *efficient equilibrium voting behavior*.[14] When

[13] Since the jury trial is modeled as a symmetric game, there exists at least one symmetric equilibrium voting behavior. The existence of symmetric equilibrium voting behavior follows very much like the result that a symmetric finite normal form game has a symmetric Nash equilibrium. We formally show the existence in Appendix B.1.

[14] In Appendix B.3, we show that other notions of equilibrium refinement motivated by *trembling hand perfection* in Austen-Smith and Feddersen (2005) or *weakly undominated strategies* in Gerardi and Yariv

no defendant goes to trial, we will refine jurors' belief that a defendant coming to the trial must be innocent. Such refinement is equivalent to imposing D1 by Cho and Kreps (1987) over the signaling game, which is induced by assuming that the jurors follow the most efficient symmetric equilibrium.

In the spirit of backward induction, we first study jury trials and find jurors' efficient equilibrium voting behavior, and then study equilibrium behaviors of a prosecutor and a defendant in plea bargaining. The following section on jury trial is a part of the backward induction, but at the same time the results also serve as a baseline of comparison about the effects of plea bargaining on jury trials.

2.3 A Jury Trial

Jurors' behavior in any jury trial that does take place hinges on the outcome of plea bargaining. Recall that π denotes the jurors' (identical) belief about a defendant's type conditional on the case going to trial. We assume that a guilty defendant is less likely to go to trial than an innocent defendant ($\pi \leq .5$). This assumption turns out to be innocuous, as guilty defendants are more likely to generate guilty signals g, each juror is more likely to vote for conviction when she receives a signal g, and thus, guilty defendants have a higher chance of being convicted.[15] As defendants anticipate such jury behavior, guilty defendants tend to plead guilty, and are therefore less likely to go to trial, relative to innocent defendants.

As is standard in strategic voting models, a juror understands that her vote affects the verdict only when she is pivotal. Thus, in addition to her private signal (g or i), the juror takes into account in her voting decision that she is pivotal (piv) and the defendant in the trial could have pleaded guilty (belief π).

Let $P[G|piv, g, \pi]$ denote the posterior probability that the defendant is guilty, conditional on receiving signal g, belief π, and being pivotal:

$$Pr[G|piv, g, \pi] := \frac{\pi \cdot p \cdot Pr[piv|G]}{\pi \cdot p \cdot Pr[piv|G] + (1-\pi) \cdot (1-p) \cdot Pr[piv|I]}$$

(2007) are insufficient to get a well-behaving equilibrium voting behavior, satisfying properties in Proposition 2.3.2.

[15] We formally prove this reasoning in Proposition 2.3.2.

Convicting the defendant changes her expected utility by $-q \cdot Pr[I|piv, g, \pi]$, and acquitting changes her utility by $-(1-q) \cdot Pr[G|piv, g, \pi]$. The expected utility from a guilty verdict is greater than or equal to the expected utility of an innocent verdict if and only if $Pr[G|piv, g, \pi] \geq q$. In other words, given all the information available, $Pr[G|piv, g, \pi] \geq q$ indicates that evidence of guilt is clear enough to exceed the level of reasonable doubt (q). In such a case, the optimal outcome from the juror's viewpoint is to convict. Whereas, $Pr[G|piv, g, \pi] \leq q$ indicates that the optimal outcome for the juror is to acquit. When these terms are equal, jurors are indifferent between conviction and acquittal.

Thus, jurors' best response is voting for conviction (or acquittal) if and only if

$$\frac{Pr[\,G\,|\,piv, g, \pi\,]}{Pr[\,I\,|\,piv, g, \pi\,]} \quad \geq \text{(or } \leq\text{)} \quad \frac{q}{1-q} \quad \text{if the signal is } g.$$

When they are equal, the juror will use a mixed strategy.

By expanding the above expression, we obtain the following voting criterion that a juror will vote for conviction (or acquittal) if and only if

$$\frac{Pr[\,piv\,|G]}{Pr[\,piv\,|I]}\,\frac{p}{1-p}\,\frac{\pi}{1-\pi} \quad \geq \text{(or } \leq\text{)} \quad \frac{q}{1-q} \quad \text{if the signal is } g. \tag{2.2}$$

A similar argument is applied to a juror receiving signal i, and we obtain

$$\frac{Pr[\,piv\,|G]}{Pr[\,piv\,|I]}\,\frac{1-p}{p}\,\frac{\pi}{1-\pi} \quad \geq \text{(or } \leq\text{)} \quad \frac{q}{1-q} \quad \text{if the signal is } i. \tag{2.3}$$

The left hand side (LHS) is the likelihood ratio of guilty to innocent given that a juror is pivotal, multiplied by the likelihood ratio inferred from private information (g or i), times the ratio of beliefs on the defendant's type; the right hand side (RHS) is the ratio of reasonable doubts.

To state the probabilities of being pivotal precisely, let r_G denote the probability of voting for conviction when the defendant is guilty, and r_I be the same probability when the defendant is, instead, innocent. Since a guilty defendant and an innocent defendant send the signal g with probability p and $1-p$, respectively, we obtain

$$r_G = p\sigma_g + (1-p)\sigma_i, \quad r_I = (1-p)\sigma_g + p\sigma_i. \tag{2.4}$$

When a voting rule requires $\hat{k}$ ($1 \leq \hat{k} \leq n$) number of conviction votes for a guilty verdict, a juror becomes pivotal when $\hat{k}-1$ other jurors vote for conviction. Assuming that $0 < r_I < 1$, we obtain from (2.2) that a juror votes for conviction (or acquittal) if and only if

$$\frac{r_G^{\hat{k}-1}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}-1}(1-r_I)^{n-\hat{k}}} \frac{p}{1-p} \frac{\pi}{1-\pi} \quad \geq \text{(or } \leq\text{)} \quad \frac{q}{1-q} \text{ if the signal is } g, \tag{2.5}$$

and we obtain from (2.3) that a juror votes for conviction (or acquittal) if and only if

$$\frac{r_G^{\hat{k}-1}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}-1}(1-r_I)^{n-\hat{k}}} \frac{1-p}{p} \frac{\pi}{1-\pi} \quad \geq \text{(or } \leq\text{)} \quad \frac{q}{1-q} \text{ if the signal is } i.^{16} \tag{2.6}$$

These expressions show the main restrictions of jurors' equilibrium behavior in the jury trial.

To understand how jurors' belief affects the equilibrium voting behavior, it is convenient to introduce a function $\bar{\pi}$ defined as

$$\bar{\pi}(l\ ; p, q) := \frac{1}{\frac{1-q}{q}\left(\frac{p}{1-p}\right)^l + 1}, \quad \forall l \in \mathbb{N}$$

In order to see the motivation behind the definition of $\bar{\pi}$, we rearrange and obtain

$$\left(\frac{p}{1-p}\right)^l \frac{\bar{\pi}(l)}{1-\bar{\pi}(l)} = \frac{q}{1-q}. \tag{2.7}$$

$\bar{\pi}$ maps a number of guilty signals (l) to the level of belief (π), which gives the minimum amount of evidence for a conviction vote. In other words, if a juror becomes a dictator, $\bar{\pi}(l)$ is the threshold level of the juror's belief, such that once the juror gathers l number of guilty signals, the juror votes for conviction.

[16] When $r_I = 0$ or $r_I = 1$, (2.5) and (2.6) are not defined. When we find the most efficient equilibrium voting behavior in Appendix B.2, we treat these cases separately.

We state the equilibrium voting behavior in Proposition 2.3.1, and relegate details of computing the equilibrium behavior to Appendix B.2. A voting behavior is called *responsive* if the conviction probability with signal g is strictly higher than the probability with signal i.

Proposition 2.3.1. *(Equilibrium voting behavior) If $\pi > \bar{\pi}(\hat{k})$, the most efficient symmetric equilibrium voting behavior is responsive. Otherwise, if $\pi \leq \bar{\pi}(\hat{k})$, the most efficient symmetric equilibrium involves an equilibrium in which no juror votes for conviction.*

In all, Proposition 2.3.1 states that, if the belief is above a certain threshold level, there exists a responsive equilibrium voting behavior. Moreover, if there exists an equilibrium voting behavior which is responsive, it must be more efficient than the equilibrium in which jurors vote either always for conviction or always for acquittal. This is quite intuitive, since jurors *use* the private signals for their voting decisions in a responsive equilibrium voting behavior. The only special case is that, when $\pi = \bar{\pi}(\hat{k})$ under the unanimity rule ($\hat{k} = n$), efficient equilibrium involves both responsive equilibrium voting behavior and non-responsive equilibrium voting behavior, in which no juror votes for conviction.

Equilibrium voting behavior is mainly derived from voting criteria (2.5) and (2.6). Note that LHS of (2.5) is strictly larger than the LHS of (2.6). Unless the denominators are equal to zero, a juror receiving signal g has a greater probability of voting for conviction than a juror receiving a signal i ($\sigma_g > \sigma_i$). Suppose jurors vote for conviction with probabilities r_I and r_G, where $0 < r_I < r_G < 1$. That is, jurors do not always vote for acquittal ($0 < r_I < r_G$) and do not always vote for conviction ($r_I < r_G < 1$). Since $\sigma_g > \sigma_i$, three classes of strategies are consistent with such jury behavior: ($0 < \sigma_g < 1, \sigma_i = 0$), ($\sigma_g = 1, 0 < \sigma_i < 1$), and ($\sigma_g = 1, \sigma_i = 0$).

For instance, under a voting rule requiring $\hat{k}$ ($\hat{k} > \frac{n}{2}$) conviction votes, ($\sigma_g = 1, \sigma_i = 0$) is not an equilibrium behavior for $\pi < \bar{\pi}(2\hat{k} - n)$. To see this, suppose that a juror receives signal g and she turns out to be pivotal; $\hat{k} - 1$ other jurors vote for conviction and $n - \hat{k}$ jurors vote for acquittal. Considering that other jurors act ($\sigma_g = 1, \sigma_i = 0$), $\hat{k} - 1$ conviction votes indicate the same number of guilty signals, and $n - \hat{k}$ acquittal votes indicate the same number of innocent signals. Thus, being pivotal is equivalent to observing $2\hat{k} - n - 1$ guilty signals, which results in $2\hat{k} - n$ guilty signals combining the juror's own guilty signal.[17] When $\pi < \bar{\pi}(2\hat{k} - n)$, $2\hat{k} - n$ guilty signals provide insufficient evidence of guilt. Thus, $\sigma_g = 1$ is not a best response, and $(\sigma_g = 1, \sigma_i = 0)$ must not be an equilibrium voting behavior.

When jurors receiving signal g use a mixed strategy $(0 < \sigma_g < 1, \sigma_i = 0)$, they are necessarily indifferent between conviction and acquittal. In such an instance, the voting criterion (2.5) holds with equality, from which we obtain an expression for σ_g and the consistent range of π. When a juror receiving signal i uses a mixed strategy $(\sigma_g = 1, 0 < \sigma_i < 1)$, we obtain σ_i and the range of π from the equality of voting criterion (2.6). If jurors receiving a signal g vote for conviction and with signal i vote for acquittal $(\sigma_g = 1, \sigma_i = 0)$, the juror receiving a guilty signal has enough evidence to vote for conviction; whereas, a juror receiving an innocent signal lacks evidence, and thus votes for acquittal. The corresponding inequalities of voting criteria (2.5) and (2.6) allow us to find the range of π consistent with such a strategy profile.

We denote conviction probability of a guilty defendant and an innocent defendant by P_G and P_I, respectively. For a pair of conviction voting probabilities, r_G and r_I,

$$P_G = \sum_{k=\hat{k}}^{n} \binom{n}{k} r_G^k \, (1 - r_G)^{n-k} \,, \quad P_I = \sum_{k=\hat{k}}^{n} \binom{n}{k} r_I^k \, (1 - r_I)^{n-k} \,. \tag{2.8}$$

For each level of belief π, when jurors follow the efficient equilibrium voting behavior, we denote the pair of corresponding conviction probabilities of guilty defendants or innocent defendants as $\{(P_G, P_I)|\pi\}$. We also define $f_G(\pi) = \{P'_G|\, \exists P'_I,\, (P'_G, P'_I) \in \{(P_G, P_I)|\,\pi\}\}$ and $f_I(\pi) = \{P'_I|\, \exists P'_G,\, (P'_G, P'_I) \in \{(P_G, P_I)|\,\pi\}\}$: correspondences of the conviction probabilities of guilty defendants and innocent defendants, respectively. Remember that efficient equilibrium voting behavior is almost always unique except when the voting rule is unanimous and $\pi = \bar{\pi}(n)$.[18] Therefore, $f_G(.)$ and $f_I(.)$ are almost always single valued.

Proposition 2.3.2. *(Properties of the efficient equilibrium voting behavior)*

1. *Convicting the guilty is more likely than convicting the innocent: $P_G \geq P_I$ for all π.*

[17] We use the fact that signals have a symmetric structure: $P[g|G]$ and $P[i|I]$ are equal.

[18] This observation was discussed after Proposition 2.3.1.

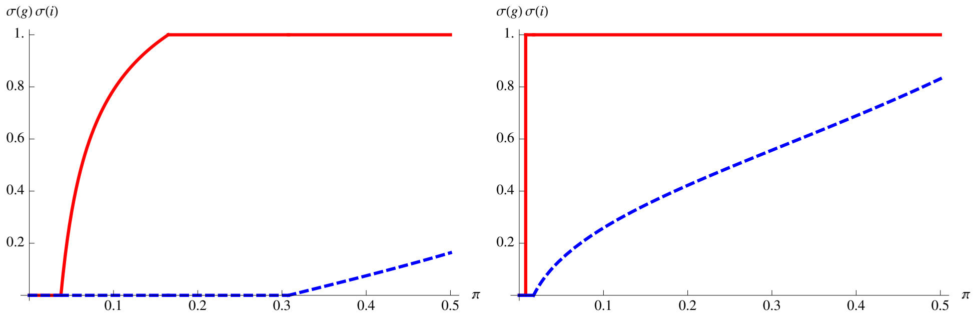
2. *Efficient equilibrium voting behavior* (σ_g, σ_i) *is non-decreasing in* π *and* $\hat{k}$.

3. *Conviction probabilities are non-decreasing in* π : *for all* $\pi < \pi'$, $f_G(\pi) \leq f_G(\pi')$ *and* $f_I(\pi) \leq f_I(\pi')$. [19]

The above properties are intuitively derived from voting criteria (2.5) and (2.6). First, the LHS of (2.5) is larger than the LHS of (2.6); a juror receiving a guilty signal is more likely to vote for conviction ($\sigma_g \geq \sigma_i$). Since guilty defendants tend to send guilty signals, jurors are more likely to vote for conviction when the defendant is guilty: i.e. $r_G \geq r_I$. Thus, guilty defendants have a higher chance of being convicted ($P_G \geq P_I$). Second, for every level of r_G and r_I (i.e. for every given other jurors' voting behavior), the value of LHS of both criteria are increasing in belief π and voting rule $\hat{k}$. Thus, a juror has more incentive to vote for conviction when belief π is higher and voting rule $\hat{k}$ is larger. Lastly, the conviction probabilities are strictly increasing functions of σ_g and σ_i, which are in turn increasing correspondences of π. Thus the conviction probabilities, P_G and P_I are increasing correspondences of π. However, it is worth noting that the conviction probabilities, P_G and P_I, may not be increasing correspondences of $\hat{k}$. Considering (2.8), depending on the level of r_G and r_I, the conviction probabilities may decrease as $\hat{k}$ gets larger.

Figure 2.2 depicts the efficient equilibrium voting behavior under a general supermajority rule ($1 \leq \hat{k} < n$) and the unanimity rule ($\hat{k} = n$). Solid lines represent the probability of voting for conviction with signal g; dashed lines represent the probability of voting for conviction with signal i. Mostly, we have a unique equilibrium voting behavior, except when $\pi = \pi(\hat{k})$ under unanimity rule. The corresponding conviction probabilities are described in Figure 2.3. Solid lines show the conviction probabilities if the defendant is truly guilty; dashed lines show the conviction probabilities of innocent defendants. Again, we certify that conviction probabilities inherit the properties of conviction voting probabilities; guilty defendants have a higher chance of being convicted and the conviction probabilities are non-decreasing in π.

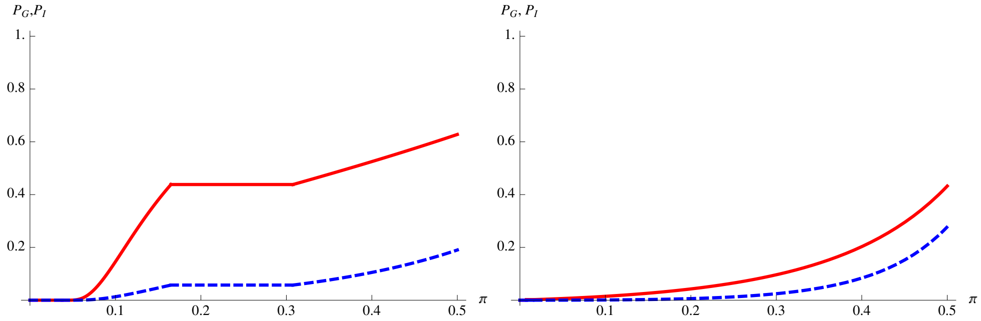
[19] Suppose A and B are sets in $\mathbb{R}$. If $a \geq b$ for every $a \in A$ and $b \in B$, we denote $A \geq B$.



(a) A super-majority rule ($\hat{k} = 8$).

(b) The unanimity rule ($\hat{k} = 12$).

Figure 2.2: Efficient symmetric voting behavior with $n = 12$, $p = \frac{6}{10}$, and $q = \frac{1}{2}$



(a) A super-majority rule ($\hat{k} = 8$).

(b) The unanimity rule ($\hat{k} = 12$).

Figure 2.3: Conviction probabilities with $n = 12$, $p = \frac{6}{10}$, and $q = \frac{1}{2}$

2.4 Plea Bargaining

A prosecutor offers the defendant an opportunity to plead guilty and undergo the penalty $\theta \in [0, 1]$. A guilty defendant compares θ with the conviction probability of guilty defendants P_G; an innocent defendant compares θ with the conviction probability of innocent defendants P_I. If θ is larger than P_G, no guilty defendant pleads guilty; similarly, no innocent defendant pleads guilty when θ is larger than P_I.[20]

Recall that π denotes the jurors' belief that the defendant is guilty conditional on a case proceeding to a trial. When some cases reach jury trials ($\phi_G < 1$ or $\phi_I < 1$), jurors update their belief π by

$$\pi = \frac{1 - \phi_G}{(1 - \phi_G) + (1 - \phi_I)}. \tag{2.9}$$

If all defendants plead guilty, $\phi_G = \phi_I = 1$, we assume that the jurors update their belief by setting it equal to 0.[21]

The relationship between the pleading decisions, ϕ_G and ϕ_I, and the conviction probabilities, P_G and P_I, captures the main interaction between plea bargaining and jury trials. One direction, how pleading decisions affect jury behavior, is explicit. The pleading decisions lead jurors to update their belief about the guilt of the defendant (updating π). As we have shown in the previous section, this belief is taken as part of the evidence of guilt in the jury's behavior, $\{(P_C, P_I)|\pi\}$. The converse direction, how jury behavior affects the pleading decisions, is implicit. The conviction probabilities are taken into account in pleading decisions through the defendants' anticipation: comparing θ and P_G, or θ and P_I. Equilibrium behavior ensures that these interactions must be consistent with each other;

[20] Such pleading decisions presume that defendants *know* the conviction probabilities of guilty or innocent defendants. In practice, defendants get advice from defense attorneys, who are aware of whether their previous clients were truly guilty and who can recall the corresponding judicial decisions. It has been also observed that participants in plea bargaining foresee the outcomes of jury trials, and consequently, previous trial outcomes significantly influence the parties' bargaining power. Among others, see, e.g., Bibas (2004) and Stuntz (2004).

[21] This assumption is equivalent to applying an equilibrium refinement, D1 by Cho and Kreps (1987), to the signaling game, induced by assuming that the jurors follow the most efficient symmetric equilibrium behavior. When jurors follow such equilibrium behavior, guilty defendants are more likely to be convicted for every jurors' belief π. Especially, if $\pi > \bar{\pi}(\hat{k})$, guilty defendants are strictly more likely to be convicted. Therefore, given an equilibrium outcome with $\phi_G = \phi_I = 1$ and for any level of $\theta > 0$, whenever guilty defendants are weakly better off by going to trials, innocent defendants are strictly better off by going to trials. Hence it should be accorded by jurors that a deviator from $\phi_G = \phi_I = 1$ is more likely to be innocent. In such a case, D1 refines jurors belief π equal to 0.

the belief π is consistent with pleading decisions ϕ_G and ϕ_I, and the anticipated conviction probabilities are consistent with π: $(P_G, P_I) \in \{(P'_C, P'_I)|\pi\}$. Proposition 2.4.1 summarizes this equilibrium restriction of the pleading decisions and jurors' voting behavior. We relegate the proof to Appendix B.5.

Proposition 2.4.1. *(Pleading decisions and voting behavior)*

Suppose the jury follows the efficient equilibrium voting behavior. For each prosecutor's offer θ, one, and only one, of the following holds.

1. ***Some guilty pleas**: Guilty defendants are indifferent between pleading guilty and undergoing a jury trial ($P_G = \theta$); innocent defendants prefer to plead not guilty ($P_I \leq \theta$). $\theta = P_G \in f_G(\pi)$ for every equilibrium belief π.*[22, 23]

2. ***No guilty plea**: P_G, and necessarily P_I, are no more than θ. All defendants plead not guilty ($\phi_G = \phi_I = 0$). Thus, $\pi = .5$ and $P_G \in f_G(.5)$.*

In general, guilty defendants are indifferent between pleading guilty and pleading not guilty ($\theta = P_G$), and innocent defendants prefer to go to trial ($P_I \leq \theta$). To see why this holds, suppose we have $\theta < P_G$. Guilty defendants will plead guilty, and depending on θ and P_I, only innocent defendants may go to trial. These pleading decisions will lead jurors to believe that all defendants in trials are innocent, and they will vote for acquittal: $\{(P_G, P_I)|\pi\} = \{(0, 0)\}$. Therefore, $\theta < P_G$ must not be an equilibrium outcome. On the other hand, $\theta > P_G$ can be an equilibrium outcome only when the prosecutor offers a high level of punishment for guilty pleas. In that event, all defendants will go to trial, the induced conviction probabilities (P_G and P_I) are still lower than θ, and such pleading decisions will turn out to be the best response.

The prosecutor wants to offer punishment θ for a guilty plea that yields his highest expected equilibrium payoff. Using the equilibrium restrictions on pleading decisions and jury

[22] The equilibrium belief π may not be unique. For instance, suppose that θ is equal to the conviction probability of a guilty defendant under $\sigma_g = 1$ and $\sigma_i = 0$. Any π inducing $\sigma_g = 1$ and $\sigma_i = 0$ as equilibrium voting behavior can be an equilibrium π. However, all $f_G(\pi)$ contains $\theta = P_G$, and lead to the same level of equilibrium punishment.

[23] Lemma B.5.1 in Appendix B.4 shows that $f_G(\pi)$ is an upper hemicontinuous correspondence with non-empty convex values. Thus for any θ in $[0, \sup f_G(\pi = .5)]$, by Intermediate Value Theorem, there exists π such that $\theta = P_G \in f_G(\pi)$.

behavior, the prosecutor's problem is summarized by the following optimization problem.

$$\max_{\theta\in[0,1]} -\frac{1}{2}q'\Big(\phi_I\theta+(1-\phi_I)P_I\Big)-\frac{1}{2}(1-q')\Big(\phi_G(1-\theta)+(1-\phi_G)(1-P_G)\Big) \tag{2.10}$$

$$\text{s.t.}\quad \begin{array}{ll} (a.1) & \phi_G \in \arg\min_{\phi'\in[0,1]} \phi'\theta+(1-\phi')P_G \\ (a.2) & \phi_I \in \arg\min_{\phi'\in[0,1]} \phi'\theta+(1-\phi')P_I \\ (b) & \pi = \begin{cases} 0 & \text{if } \phi_G=\phi_I=1 \\ \frac{1-\phi_G}{(1-\phi_G)+(1-\phi_I)} & \text{otherwise.} \end{cases} \\ (c) & (P_G,P_I)\in\{(P'_G,P'_I)|\pi\}. \end{array}$$

The objective function is the prosecutor's expected utility. The prosecutor's utility is decreasing with q' if innocent defendants are mistakenly punished. The mistake is either as a result of a guilty plea, with probability ϕ_I and punishment θ, or of conviction in jury trial, with probability $(1-\phi_I)\,P_I$ with punishment 1. When guilty defendants go without being fully punished, the prosecutor's utility is decreased by $(1-q')$. Such a case is either as a result of a guilty plea, with probability ϕ_G and undelivered punishment $(1-\theta)$, or of acquittal in a jury trial, with probability $(1-\phi_G)(1-P_G)$ and undelivered punishment 1.

The defendants will best respond in pleading decisions and the jurors will follow the equilibrium voting behavior. Such equilibrium behavior restricts the prosecutor's optimization: $(a.1)$ and $(a.2)$ represent that guilty and innocent defendants plead in order to minimize their expected punishment, respectively; (b) captures that jurors rationally update their belief π following the defendants' pleading decisions; (c) states that jurors will follow the efficient equilibrium voting behavior. The following proposition presents the prosecutor's optimal behavior, and the consequent jurors' voting behavior. In the proposition, *some guilty pleas* and *no guilty plea* refers to the two classes of equilibrium outcomes in Proposition 2.4.1 the prosecutor can induce. We leave the proof to Appendix B.6.1.

Proposition 2.4.2. *(Equilibrium outcomes of plea bargaining and jury trials)*

1. *If $q'>q$, the prosecutor induces* ***some guilty pleas****. Induced jury behavior resembles the behavior in the jury model without plea bargaining. But, jurors act as if they have*

the prosecutor's preference parameter, q'.

2. *If $q' \leq q$, the prosecutor induces **no guilty plea**. The jury behavior is the same as the behavior in the jury model without plea bargaining.*

The motivation behind the prosecutor's optimal level of θ is quite intuitive. To illustrate the main idea, we first show that the prosecutor is primarily concerned with how plea bargaining affects jurors' belief π.

To begin with, the prosecutor only needs to focus on equilibrium outcomes with *some guilty pleas* in Proposition 2.4.1. Suppose that an equilibrium outcome has *no guilty plea.* That is, the punishment following a guilty plea is so high that all defendants proceed to jury trials. The prosecutor can achieve the utility corresponding to the *no guilty plea* equilibrium outcome by offering $\theta = \bar{\theta}$ where $\bar{\theta} := \sup f_G(.5)$. Although some guilty defendants may change their mind to pleading guilty, the prosecutor achieves the same utility gain or loss, regardless of whether the guilty defendants plead guilty or not guilty.

Without loss of generality, we simplify the prosecutor's objective function in (2.10) using the case of some guilty pleas in Proposition 2.4.1. In general, we have $\theta > 0$, and thus $\theta = P_G > 0$.[24] The equilibrium voting behavior becomes responsive ($P_G > P_I$), and all innocent defendants go to trial ($\phi_I = 0$). Then the prosecutor's objective function becomes

$$-\frac{1}{2}q'P_I - \frac{1}{2}(1-q')(1-P_G). \tag{2.11}$$

We now see that the prosecutor's main concern is to influence jurors' belief π, thereby leading jurors' best responding behavior to be most preferable to the prosecutor. One thing to note here is that the prosecutor is not allowed to 'force' jurors to take a certain voting strategy. That is, he can at best lead them to one of the most efficient equilibrium voting behaviors.

To see *how* the prosecutor should influence the jurors' belief π, we revisit the jurors' voting criteria. By modifying (2.5) and (2.6), we obtain

[24] We will also obtain (2.11) when $\theta = 0$; nevertheless, we treat the case separately in Appendix B.6.1, because the voting criteria (2.5) and (2.6) will not be well-defined.

$$\frac{Pr[\,piv\,|G]}{Pr[\,piv\,|I]}\ \frac{p}{1-p}\ \frac{.5}{1-.5} \quad \geq \text{(or } \leq\text{)} \quad \frac{q}{1-q}\ \frac{1-\pi}{\pi} \quad \text{if the signal is } g,$$

and

$$\frac{Pr[\,piv\,|G]}{Pr[\,piv\,|I]}\ \frac{1-p}{p}\ \frac{.5}{1-.5} \quad \geq \text{(or } \leq\text{)} \quad \frac{q}{1-q}\ \frac{1-\pi}{\pi} \quad \text{if the signal is } i.$$

The voting criteria above lead to the same voting behavior as the voting criteria (2.5) and (2.6); jurors receiving signal g or i vote for conviction if confronted with the former pair of criteria if and only if the jurors receiving signal g or i vote for conviction if confronted with the latter pair of criteria. That is, the jury behavior with a belief π and the ratio of reasonable doubts $\frac{q}{1-q}$ is equal to the jury behavior with belief .5 and the ratio of reasonable doubts equal to $\frac{q}{1-q}\frac{1-\pi}{\pi}$. As a result, we can reinterpret the prosecutor's effort to influence the jurors' belief as an effort to change the level of the jurors' reasonable doubts, while fixing the belief at the prior $\pi_0 = .5$. The question, "How to influence the jurors' belief?" is then the same as, "Which level of the jurors' influenced reasonable doubt is the most preferable to the prosecutor?"

Intuitively, the prosecutor prefers to have the jurors' induced reasonable doubt to perfectly coincide with his weights on mistakenly delivered or undelivered punishments: i.e., $\frac{q'}{1-q'} = \frac{q}{1-q}\frac{1-\pi}{\pi}$. However, the prosecutor can affect the jurors' reasonable doubt in only one direction; he can only increase the reasonable doubt by inducing $\pi \leq .5$. When the jurors, rather than the prosecutor, care more about punishing innocent defendants ($q > q'$), the prosecutor has no incentive to use plea bargaining, and so he induces $\pi = .5$ by offering $\theta \geq \sup f_G(.5)$.

Figure 2.4 illustrates prosecutor's optimal offer of guilty plea punishment, for each level of prosecutor's parameter q' and under various voting rules $\hat{k}$. As Proposition 2.4.2 states, the optimal offer is divided into two classes. Compared to jurors, when the prosecutor is less cautious about punishing innocent defendants ($q' \leq q = \frac{1}{2}$), the prosecutor offers a high level of punishment and induces *no guilty plea*. Otherwise, the prosecutor offers a lower level of punishment and induces *some guilty pleas*. As the guilty plea punishment becomes more lenient, the number of guilty defendants pleading guilty increases. Such pleading

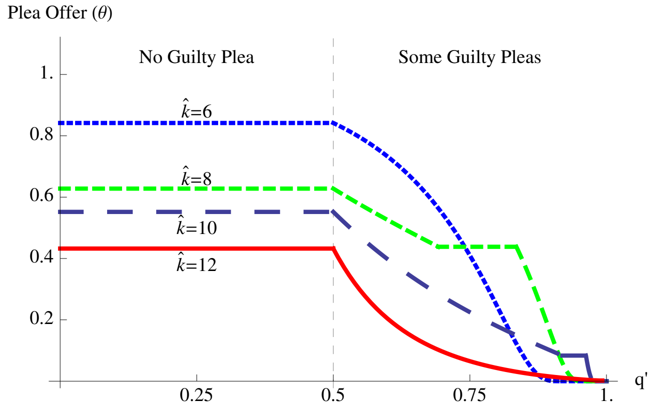


Figure 2.4: Optimal offer of guilty plea punishment given $n = 12$, $p = \frac{6}{10}$, and $q = \frac{1}{2}$

decisions yield a lower level of belief π and consequently lower chances of convicting innocent defendants. Therefore, the optimal offer θ is a decreasing function of prosecutor's utility parameter q'. The optimal plea bargain offer is not a monotone function of the voting rule $\hat{k}$. This is because conviction probabilities are not monotone functions of $\hat{k}$, as mentioned in the discussion of Section 2.3.

2.5 Comparison of Alternative Voting Rules

As a direct application of Proposition 2.4.2, we re-examine a previous finding of the standard jury model (*without* plea bargaining).

Feddersen and Pesendorfer (1998) find that the unanimity rule is inferior to general super-majority rules. As the number of jurors gets large, the chance of convicting innocent defendants and the chance of acquitting guilty defendants do not converge to zero under the unanimity rule; whereas, both converge to zero if the voting rule is non-unanimous.[25] Assuming that the jury trial employs either the unanimity rule or a super-majority rule, we confirm that the previous results are robust to the addition of plea bargaining. We relegate

[25] These are asymptotic properties, rather than results with a finite number of jurors; for example, jury size 12 is common in the U.S. criminal court. In spite of that, when p is not close to $\frac{1}{2}$, the asymptotic properties closely approximate the properties with a finite number of jurors. For instance, when $p = \frac{2}{3}$, $q = \frac{1}{2}$, and $\pi = 12$, the limit of conviction probabilities for a guilty or an innocent defendant is 1 or 0 under any non-unanimous rule, and 0.5 or 0.25 under the unanimity rule, respectively. On the other hand, a jury with 12 jurors convicts a guilty or an innocent defendant with probability 0.90 or 0.03 under a non-unanimous rule $\hat{k} = 8$, and 0.57 or 0.17 under the unanimity rule, respectively. Moreover, asymptotic properties are also mathematically more tractable.

the proof to Appendix B.6.4.

Corollary 5. *(Comparing Voting Rules)*

1. *If a jury trial uses the unanimity rule, the expected punishment of guilty defendants converges to* $1 - \left(\frac{(1-\tilde{q})(1-p)}{\tilde{q}p}\right)^{\frac{1-p}{2p-1}}$ *as* $n \to \infty$*, where* $\tilde{q} = \max\{q, q'\}$*; for innocent defendants, it converges to* $\left(\frac{(1-\tilde{q})(1-p)}{\tilde{q}p}\right)^{\frac{p}{2p-1}}$.

2. *If the jury trial uses a non-unanimous rule, the expected punishment for guilty defendants converges to one; the expected punishment for innocent defendants converges to zero.*

Corollary 5 is from Proposition 2.4.2 and asymptotic properties of the jury's behavior in Feddersen and Pesendorfer (1998).[26] Proposition 2.4.2 states that the induced jury behavior in a court with plea bargaining is similar to the equilibrium behavior in the jury model without plea bargaining. If $q \leq q'$, we can mimic the jury behavior using a jury model without plea bargaining by assuming that jurors echo the prosecutor's preference. If $q > q'$, the behaviors are exactly the same. Concerning jury behavior under the unanimity rule and general super-majority rules, plea bargaining does not change the qualitative findings, but only affects the quantitative analyses: i.e. the probability limits. Therefore, the inferiority result in Feddersen and Pesendorfer (1998) is robust to the addition of plea bargaining.

However, it is worth stressing that while the previous literature considers jury trial outcomes, or conviction probabilities, we treat the outcomes of the entire judicial process: punishment by guilty pleas as well as conviction probabilities. Therefore, Corollary 5 compares expected punishments, rather than conviction probabilities, under either unanimity rule or super-majority rules.

2.6 Discussion

Plea bargaining is the most common method of resolving cases in U.S. criminal court, though studies on collective decision making have largely ignored plea bargaining. Whereas, jury

[26] Propositions 2 and 3 in Feddersen and Pesendorfer (1998) state the asymptotic properties of the jury's behavior under the unanimity rule and general super-majority rules.

trials have been rigorously studied, while in practice only a small portion of criminal cases reach jury trial. The current paper bridges such a gap between the practice and the theory by studying a combined model of plea bargaining and a jury trial. We highlight that plea bargaining and jury trials interact with one another during a criminal court process. By influencing the jurors' belief, plea bargaining may induce the jury's behavior to reflect the prosecutor's preference rather than the jurors'.

The results in this paper raise an important issue, especially for empirical analysis of criminal court process and of its effects on society. Most of our practical knowledge on jury trials is essentially based on the cases handled in trials. Yet, such knowledge lacks fundamental understandings and tells little about the potential effects of institutional changes on society. As jury trials are *chosen* through plea bargaining, the cases in jury trials do not represent the entire population of criminal cases. Moreover, institutional changes will alter the characteristics of the cases coming to trials. As such, it is appropriate to employ a structural model, combining both plea bargaining and jury trials, rather than studying each of them separately.

Chapter 3

The Testable Implications of Zero-sum Games

[1]

3.1 Introduction

Suppose two players choose joint actions from a finite set of alternatives. As outside observers, we witness the joint choice behavior, but we may not know the exact payoffs leading the players to such group choices. By only observing joint choice behavior, we may ask whether people play Nash equilibrium, and, if they do, what type of games they play.

This paper derives falsifiable conditions of joint choice behavior from equilibrium play of a zero-sum game, or a game of conflicting interests. That is, we study additional behavioral implications of a game being zero-sum, in addition to the hypothesis of Nash equilibrium play. Instead of assuming a specific pattern of joint behavior, this study requires only weak rationality axioms: complete and transitive preferences at the individual level, and Nash equilibrium play at the collective level.

The main motivation of this exercise is that we want to be able to refute the notion that two agents are in "direct competition," and detect whether or not there could be "gains from cooperation," without knowing the exact payoffs. However, its applications are not limited to cases where we only observe joint choice behavior. Even when we observe the exact monetary returns (e.g., a laboratory experiment), the observed monetary returns may

[1] This chapter is published as a research paper in *Journal of Mathematical Economics*, Volume 48, Issue 1, January 2012, Pages 39-46.

differ from utility payoffs which players perceive. For example, if each subject cares about her monetary return relative to her opponent's return, the joint behavior may follow Nash equilibrium behavior of a zero-sum game rather than the original game. This is because a two-person game with symmetric monetary returns becomes a symmetric zero-sum game with respect to relative monetary returns.[2] Based on the observed joint choice behavior, we can test whether subjects play the original game or the zero-sum game induced by their relative monetary returns.

Sprumont (2000) assumes that econometricians are given a choice correspondence defined on product sets of individual actions. The question is when the observed joint behavior is consistent with Nash equilibrium play, assuming players are rational and they play games simultaneously. Sprumont proves that the observed joint behavior is Nash rationalizable if and only if it satisfies *Persistent under Restriction* and *Persistent under Expansion* axioms, which are similar to classical axioms of choice theory (see, e.g., Moulin (1985)). We retain Sprumont's basic abstract setup and ask, "Is the choice correspondence Nash-rationalizable with a certain game category, specifically, zero-sum games?"

As an introductory example, Figure 1 shows how Nash-rationalizable choice behavior may not be rationalizable by a zero-sum game. In this example, player 1 conceivably choose either U or D and player 2 may choose L or R. Following classical choice theory, we may observe how players choose when choice sets are restricted. Figure 3.1 shows all the possible product subsets of $\{U, D\} \times \{L, R\}$ from which two players choose their joint actions. For each product subset, $(*)$ is the action profile chosen by the players. We can verify that the joint choice behavior exhibited in Figure 1 is consistent with Nash equilibrium behavior of a coordination game in which coordinating to (U, L) or (D, R) gives a higher payoff to both players.

This choice correspondence, however, does not follow Nash equilibrium behavior for any zero-sum game. We observe that (U, L) is chosen from $\{(U, L), (D, L)\}$ and (D, R) is chosen from $\{(D, L), (D, R)\}$. Assuming that the choices are Nash equilibria of a zero-sum game, these choices imply that for player 1, (U, L) is preferred to (D, L); for player 2, (D, R) is preferred to (D, L), which indicates player 1 prefers (D, L) to (D, R). On the other hand,

[2] See, e.g., Duersch, Oechssler, and Schipper (2011).

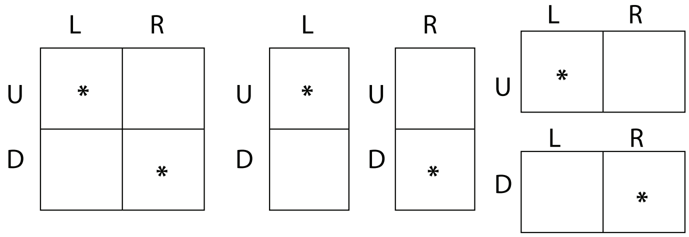


Figure 3.1: Nash-rationalizable but *not* by zero-sum games

(D, R) is chosen from $\{(D, R), (U, R)\}$ and (U, L) is chosen from $\{(U, L), (U, R)\}$. For player 1, (D, R) is preferred to (U, R); for player 2, (U, L) is preferred to (U, R), which implies player 1 prefers (U, R) to (U, L). As a result, the preference of player 1 forms a cycle, which implies that all possible joint actions are indeed indifferent for player 1 (and thus for player 2 by the fact that the game is zero-sum). Therefore, we would expect to see all strategy profiles chosen.

This example shows that once we have two choices on the diagonal in a table of joint actions, the other two pairs of actions must also be chosen in order for the joint choices to follow Nash equilibrium behavior for a zero-sum game. When choice behavior forms a product subset for each game table, we say that the choice behavior is *interchangeable*. Although it is easy to identify that interchangeability is necessary, whether the condition is sufficient is not as straightforward.

Our main theorem shows that this interchangeability of joint choice behavior is indeed the only additional condition that distinguishes the testable implications of zero-sum games from those of general non-cooperative games. It is worth pointing out two assumptions behind the theorem. First, we restrict Nash rationalizability to pure strategy Nash equilibria. Second, we assume complete observations, where choices are observed from all product sets of individual actions.

This paper follows a broad range of revealed preference theory. Since Samuelson (1938), there have been numerous papers on revealed preference theory in various settings. In the context of collective choice, Wilson (1970) and Plott (1974) study cooperative games and find that the Weak Axiom implies the solution concept proposed by von Neumann and Morgenstern. More recently, Echenique and Ivanov (2011) and Chambers and Echenique (2011) study the testable implications of collective decision making such as household behavior and bargaining over money.

The testable implications of game theoretic models have grown only recently relative to the history and popularity of game theory. Peleg and Tijs (1996) and Sprumont (2000) find conditions of joint choice behaviors being consistent with Nash equilibria, as the corresponding games are reduced or expanded. Galambos (2009) weakens the complete observation assumption, and Demuynck and Lauwers (2009) study joint choices over lotteries. The two approaches adopt Richter (1971)'s congruence axiom, and find that the modified versions of the congruence axiom are necessary and sufficient conditions for Nash rationalizability. Ray and Zhou (2001), and Ray and Snyder (2003) consider extensive form games, and find conditions such that sequential choices are rationalizable by a subgame perfect Nash equilibrium. Xu and Zhou (2007) characterize conditions under which choices are rationalizable by game trees when the choice process is not observable.

In the context of more concrete games, Forges and Minelli (2009) apply their main result to market games, in which each player's budget constraint depends on other players' actions. For the model of Cournot competition, Carvajal, Deb, Fenske, and Quah (2010) consider the case of observing a finite set of prices and quantities, and Cherchye, Demuynck, and De Rock (2011) consider the case of observing price and quantity functions defined over exogenous variables. Both studies characterize conditions under which their observed data are consistent with the model of Cournot competition.

3.2 Model and Main Theorem

There are two players, 1 and 2. Let A_1 and A_2 be finite sets of actions that players 1 and 2 may conceivably choose. $A := A_1 \times A_2$ is the set of all possible joint actions. Following the classical revealed preference approach, suppose we observe choices from $B := B_1 \times B_2$ in which $B_1 \subset A_1$ and $B_2 \subset A_2$ are the sets of available actions for player 1 and 2. In this model, all choices from each $B \subset A$ can be summarized as a choice correspondence.

Definition 1. *Let $\mathcal{A} := \{B = B_1 \times B_2 | \emptyset \neq B \subset A\}$ be the set of all nonempty product sets included in A. A* ***joint choice correspondence*** *f assigns to each $B \in \mathcal{A}$ a nonempty set*

$f(B) \subset B$.

In the case where at most one player has more than one available action in B, we say that B is a *line*. Depending on the player, the line is either in a *column* or a *row* - the former when player 1 has choices, the latter when player 2 has choices. In addition, we call a product subset $B \in \mathcal{A}$ a *feasible set*. For any $B^{''} \subset B$ and $B^{''} \in \mathcal{A}$, we call $B^{''}$ a *feasible subset* of B. For any $B, B^{'} \in \mathcal{A}$, define $B \vee B^{'}$ as the set of all possible pairs of actions from B_i and $B_i^{'}$ $(i = 1, 2)$. That is,

$$B \vee B^{'} := \prod_{i=1,2} (B_i \cup B_i^{'})$$

Suppose we wish to test whether a choice correspondence is rational or not. First, we shall assume that each player is individually rational. That is, each player has a preference relation over joint actions, and these relations are complete and transitive.[3] We call such relations **weak orders**. In addition, we wish to test if the players are collectively rational. In terms of collectively rationality, we assume that players play a Nash equilibrium. The following definition is our notion of rationalizability of collective choice behavior.

Definition 2. *A joint choice correspondence f is* ***Nash-rationalizable*** *if there are two weak orders $\succeq_1, \succeq_2$ on A such that for each $B \in \mathcal{A}$, $f(B)$ coincides with the set of all Nash equilibria of the game $(B, \succeq_1, \succeq_2)$.*[4]

Sprumont (2000) introduces the following conditions for Nash-rationalizability. These conditions are extended versions of *Sen's* α, β, and γ in individual choice theory (see, e.g., Moulin (1985)).[5] When a feasible set is restricted to a line, the first condition coincides with *Sen's* α and β, and the second condition coincides with *Sen's* γ.

Definition 3. *A joint choice correspondence over $\mathcal{A}$ is:*

- ***Persistent under Contraction*** *(PC):*

[3] A relation $\succeq$ is called *complete* if for all joint choices $a, b \in A$, it follows that $a \succeq b$ or $b \succeq a$, and is called *transitive* if for all $a, b, c \in A$ for which $a \succeq b$ and $b \succeq c$, it follows that $a \succeq c$.

[4] In other words, if $(b_1, b_2) \in f(B)$, then $(b_1, b_2) \succeq_1 (b_1', b_2)$ and $(b_1, b_2) \succeq_2 (b_1, b_2')$ for every $(b_1', b_2) \in B$ and $(b_1, b_2') \in B$.

[5] Although Moulin (1985) calls these conditions *Chernoff* and *Expansion*, *Sen's* α, β, and γ are more conventional terminologies in individual choice theory. See, for example, Austen-Smith and Banks (1994).

(PC1) : For all $B, B^{'} \in \mathcal{A}$ with $B^{'} \subset B$, $f(B) \cap B^{'} \subset f(B^{'})$.

(PC2) : Moreover, if B is a line, $B' \subset B$ and $f(B) \cap B^{'} \neq \emptyset$ implies $f(B^{'}) \subset f(B)$.

- ***Persistent under Expansion*** *(PE): For all $B, B^{'} \in \mathcal{A}$, $f(B) \cap f(B^{'}) \subset f(B \vee B^{'})$.*

With these two conditions, Sprumont (2000) establishes the following theorem.

Theorem 3.2.1. *A joint choice correspondence f is Nash-rationalizable if and only if it satisfies (PC) and (PE).*

Using this model of Nash-rationalizability, we restrict the set of available rationalizing games from the set of all non-cooperative games to include only zero-sum games, or games of conflicting interests. Under the conditions of zero-sum games, the preferences of two players are opposed. Therefore, while a general non-cooperative game consists of two weak orders, zero-sum games require only a single weak order.

Definition 4. *Let $\succeq$ be a weak order over A, and $\preceq$ is the inverse relation of $\succeq$.*[6] *The game defined by $(A, \succeq, \preceq)$ is called* ***a two-person zero-sum game****. We say that a joint choice correspondence f is* ***Nash-rationalizable by a zero-sum game*** *if there is a weak order $\succeq$ on A such that for each $B \in \mathcal{A}$, $f(B)$ coincides with the set of all Nash equilibria of the game $(B, \succeq, \preceq)$.*

As demonstrated in Example 1, not all Nash-rationalizable joint choice correspondences are Nash-rationalizable by a zero-sum game. In the example, we needed one additional condition to fill the gap in the product space of the two distinct choices. We formally state this condition in the following definition.

Definition 5 (Interchangeability (INT))**.** *A joint choice correspondence f over $\mathcal{A}$ is* ***interchangeable*** *if for all $B \in \mathcal{A}$ and all $b, b^{'} \in f(B)$, $\{b\} \vee \{b^{'}\} \subset f(B)$.*

It is well-known that any pair of equilibrium strategies of a zero-sum game, one for each player, is an equilibrium strategy profile (see, e.g., Luce and Raiffa (1989)). Provided that

[6] Let $\succeq$ be a binary relation over A. We define the inverse relation $\preceq$ as

$$\text{for all } a, b \in A \text{ for which } a \succeq b,\ b \preceq a.$$

The inverse relation of a weak order is also a weak order. The proof is immediate by definition.

players face a zero-sum game, and observed joint actions follow the Nash-equilibria of the corresponding games, the choice correspondence must be interchangeable. Our contribution is showing that interchangeability is indeed the only additional behavioral implication which distinguishes zero-sum games from general non-cooperative games. We summarize this result as the following main theorem.

Theorem 3.2.2. *A joint choice correspondence is Nash-rationalizable by a zero-sum game if and only if it satisfies (PC), (PE), and (INT).*

3.3 Discussion

Our model assumes the existence of a joint choice for all $B \in \mathcal{A}$. Accordingly, verifying whether a joint choice correspondence is Nash-rationalizable means assuming that all feasible sets have a pure strategy Nash equilibrium. Although a small literature provides conditions of zero-sum games having pure strategy Nash equilibria (Shapley, 1964; Radzik, 1991; Duersch, Oechssler, and Schipper, forthcoming), the characterization of conditions that are both necessary and sufficient remains an open question. We may avoid this existence issue by investigating either mixed strategies or correlated strategies. However, these strategies introduce other difficulties since observed joint choices do not directly represent underlying preferences.

Our model also requires observed choices from all feasible sets. We may weaken this requirement by assuming *incomplete observations*, where a choice correspondence is defined on $\mathcal{A}^{'} \subset \mathcal{A}$. In classical choice theory, Richter (1971) shows that a choice correspondence with incomplete observations is rationalizable by a weak order if and only if it is congruent. Galambos (2009) generalizes Richter's congruence condition, and shows that the generalized congruence condition is necessary and sufficient for Nash-rationalizability with incomplete observations.

Unfortunately, interchangeability together with individual-level congruent choices is not sufficient for Nash-rationalizability by a zero-sum game. For example, suppose $\mathcal{A}^{'}$ is the set of $B := \{U, M\} \times \{L, C, R\}$, $B' := \{U, M, D\} \times \{C, R\}$, and all lines in B and B'. Suppose $f(B) = \{(M, R)\}$ and $f(B') = \{(U, C)\}$, and assume that choices in each line satisfy (PE) and (PC). The choices are congruent in terms of Galambos (2009), and therefore, Nash-rationalizable. However, the choices are not Nash-rationalizable by a zero-sum game. From $\{U, M\} \times \{C, R\}$, (U, C) and (M, R) are the only choices consistent with the choices in each line, but this observation violates interchangeability.[7]

[7] Alternatively, we may consider a *congruent* joint choice correspondence $f : \mathcal{A}' \rightrightarrows A$, where $\mathcal{A}' \subset \mathcal{A}$ is a set of observed games. We define a binary relation $\succeq$ on A by: for $a = (a_1, a_2), b = (b_1, b_2) \in A$,

$$\begin{aligned} &a \succeq b \text{ if and only if there exists } B \in \mathcal{A}' \text{ such that } a, b \in B, \text{ and} \\ &\qquad\qquad \text{either } a_2 = b_2 \text{ and } a \in f(B), \text{ or } a_1 = b_1 \text{ and } b \in f(B). \end{aligned}$$

If there is a finite sequence $c, d, \ldots, e$ such that $a \succeq c \succeq d \cdots \succeq e \succeq b$, then we write $a \, T_{\succeq} \, b$. We say that a joint choice correspondence f is *congruent*, if for all $a, b \in A$ and all $B \in \mathcal{A}'$,

$$a \, T_{\succeq} \, b, \; a \in B, \text{ and } b \in f(B) \implies a \in f(B).$$

Assuming that a joint choice correspondence is congruent is, however, almost the same as assuming its Nash-rationalizability by a zero-sum game. In particular, when $\mathcal{A}' = \mathcal{A}$, the assumption implies that the relation $\succeq$ is *consistent* (see Definition 6). Most of the proof in this paper is devoted to showing that $\succeq$ is consistent (see Section C.1).

Appendix A

Appendix to Chapter 1: Large Matching

First in Appendix A.1, we summarize definitions and related theorems of asymptotic statistics. We prove Theorem 1 for the case of $\lambda = 0$ in Appendix A.2, and for the case of $0 < \lambda < 1$ in Appendix A.7. The proof of Theorem 4 is given in Appendix A.4. Lastly in Appendix A.5, we provide additional simulation results of effects of limited acceptability assumption on the proportion of unmatched agents.

A.1 Asymptotic Statistics (Serfling, 1980)

Let $X_1, X_2, \ldots$ and X be random variables on a probability space $(\Omega, \mathcal{A}, P)$. We say that X_n **converges in probability** to X if

$$\lim_{n \to \infty} P\left(|X_n - X| < \epsilon\right) = 1, \quad \text{every } \epsilon > 0.$$

This is written $X_n \xrightarrow{p} X$.

For $r > 0$, we say that X_n **converges in the r^{th} mean** (or in the L^r-norm) to X if

$$\lim_{n \to \infty} E\left(|X_n - X|^r\right) = 0.$$

This is written $X_n \xrightarrow{L^r} X$.

Theorem A.1.1. *If $X_n \xrightarrow{L^r} X$, then $X_n \xrightarrow{p} X$.*

Theorem A.1.2. *Suppose that* $X_n \xrightarrow{p} X$, $|X_n| \leq |Y|$ *with probability 1 (for all* n*), and* $E\left(|Y|^r\right) < \infty$*. Then,* $X_n \xrightarrow{L^r} X$*.*

Remark 1. *In this paper, most random variables represent proportions, which are bounded above by 1 with probability 1. As such, convergence in probability and convergence in the* r^{th} *mean are equivalent.*

Theorem A.1.3. *Let* $\mathbf{X_1}, \mathbf{X_2}, \ldots,$ *and* $\mathbf{X}$ *be random* k*-vectors defined on a probability space, and let* g *be a vector-valued Borel function defined on* $\mathbf{R}^k$*. If* g *is continuous with* $P_{\mathbf{X}}$*-probability 1, then*

$$\mathbf{X_n} \xrightarrow{p} \mathbf{X} \implies g(\mathbf{X_n}) \xrightarrow{p} g(\mathbf{X}).$$

In particular, if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$ and $X_n Y_n \xrightarrow{p} XY$.

Given a univariate distribution function F and $0 < q < 1$, we define q^{th} **quantile** ξ_q as

$$\xi_q := \inf\{x : F(x) \geq q\}.$$

Consider an i.i.d sequence $\langle X_i \rangle$ with distribution function F. For each sample of size n, $\{X_1, X_2, \ldots, X_n\}$, a corresponding **empirical distribution function** F_n is constructed as

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{X_i \leq x\right\}, \quad -\infty < x < \infty.$$

The **empirical** q^{th} **quantile** $\hat{\xi}_{q:n}$ is defined as the q^{th} quantile of the empirical distribution function. That is

$$\hat{\xi}_{q:n} := \inf\{x : F_n(x) \geq q\}.$$

For each x, $F_n(x)$ is a random variable, and therefore, $\hat{\xi}_{q:n}$ is also a random variable.

Theorem A.1.4. *Suppose that* q^{th} *quantile* ξ_q *is the unique solution* x *of* $F(x-) \leq q \leq F(x)$*. Then, for every* $0 < q < 1$ *and* $\epsilon > 0$*,*

$$P\left(\left|\hat{\xi}_{q:n} - \xi_q\right| > \epsilon\right) \leq 2e^{-2n\lambda_\epsilon^2}$$

for all n*, where* $\lambda_{1,\epsilon} = F(\xi_q + \epsilon) - q$*,* $\lambda_{2,\epsilon} = q - F(\xi_q - \epsilon)$*, and* $\lambda_\epsilon = \min\{\lambda_{1,\epsilon},\ \lambda_{2,\epsilon}\}$*.*

For each sample of size n, $\{X_1, X_2, \dots, X_n\}$, the ordered sample values

$$X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$$

are called the **order statistics**.

In view of

$$X_{k:n} = \hat{\xi}_{k/n:n}, \quad 1 \leq k \leq n, \tag{A.1}$$

we will carry out proofs in terms of empirical quantiles, even when variables are defined as order statistics.

A.2 Proof of Theorem 1 ($\lambda = 0$)

Let $\zeta = [\zeta_{f,w}]$ be an i.i.d sample from a continuous distribution Γ^W with support $[0, \bar{u}]$, and $\eta = [\eta_{f,w}]$ be an i.i.d sample from a continuous distribution Γ^F with support $[0, \bar{v}]$.[1]

For $\epsilon > 0$ and for each $\langle F, W, u, v \rangle$, we define

$$B^F(\epsilon; u, v) := F \backslash A^F(\epsilon; u, v) = \{f \in F \mid \Delta(f; u, v) \geq \epsilon\},$$

and prove that

$$E\left[\frac{|B^F(\epsilon; U, V)|}{n}\right] \to 0, \quad \text{as} \quad n \to \infty. \tag{A.2}$$

We define the set of firms whose utilities from the worst stable matching are significantly below the upper bound $\bar{u}$, which we shall write as

$$\bar{B}(\epsilon; u, v) := \{f \in F | u_{\mu_W}(f) \leq \bar{u} - \epsilon\}.$$

Note from $u_{\mu_F}(f) \leq \bar{u}$ that

$$u_{\mu_F}(f) - u_{\mu_W}(f) \leq \bar{u} - u_{\mu_W}(f),$$

[1] We use Γ^W, instead of Γ^F, to represent the distribution of utilities of firms, interpreting it as the distribution of private-values of workers. This notation will be consistent with the additional notation G^W representing the distribution of workers' common-values. By the same reason, we use Γ^F to denote the distribution of utilities of workers, or private-values of firms.

and thus

$$B^F(\epsilon; u, v) \subset \bar{B}(\epsilon; u, v).$$

Therefore, (A.2) follows immediately from the following proposition.

Proposition A.2.1. *For every $\epsilon > 0$,*

$$E\left[\frac{\left|\bar{B}(\epsilon; U, V)\right|}{n}\right] \to 0 \quad as \quad n \to \infty.$$

We divide the proof of Proposition A.2.1 into two lemmas. For every market instance $\langle F, W, u, v\rangle$, we let $R_{\mu_W}(f)$ be the rank number of firm f's worker-optimal stable matching partner: e.g. $R_{\mu_W}(f) = 1$ if f matches with its most preferred worker. We first observe that for most firms, the rank number of worker-optimal matching partner normalized by n converges to 0. The second lemma shows that the corresponding utility level must become close to the upper bound $\bar{u}$ as the market becomes large.

Lemma A.2.2. *For $\gamma > 0$, let*

$$\bar{B}_q(\gamma; u, v) := \left\{f \in F \mid \frac{R_{\mu_W}(f)}{n} \geq \gamma\right\} = \left\{f \in F \mid 1 - \frac{R_{\mu_W}(f)}{n} \leq 1 - \gamma\right\}.$$

Then, for every sequence $\langle \gamma_n \rangle$ such that $\gamma_n \to 0$ and $(\log n) \cdot \gamma_n \to \infty$,

$$E\left[\frac{\left|\bar{B}_q(\gamma_n; U, V)\right|}{n}\right] \to 0 \quad as \quad n \to \infty.$$

Proof. For every instance $\langle F, W, u, v\rangle$ and for every sequence $\langle \gamma_n \rangle$ satisfying the conditions,

$$\begin{aligned}\frac{1}{n}\gamma_n \left|\bar{B}_q(\gamma_n; u, v)\right| &\leq \frac{1}{n} \sum_{f \in \bar{B}_q(\gamma_n; u, v)} \frac{R_{\mu_W}(f)}{n} \\ &\leq \frac{1}{n} \sum_{f \in F_n} \frac{R_{\mu_W}(f)}{n}.\end{aligned}$$

We use Theorem 2 in Pittel (1989) showing that

$$\frac{\sum_{f \in F_n} R_{\mu_W}(f)}{n^2 \log^{-1} n} \xrightarrow{p} 1. \tag{A.3}$$

Applying (A.3), we shall write

$$\begin{aligned}\frac{\left|\bar{B}_q(\gamma_n; U, V)\right|}{n} &\leq \frac{\sum_{f \in F_n} R_{\mu_W}(f)}{n^2} \frac{1}{\gamma_n} \\ &= \frac{\sum_{f \in F_n} R_{\mu_W}(f)}{n^2 \log^{-1} n} \frac{1}{\log n \cdot \gamma_n} \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty.\end{aligned}$$

We obtain Lemma A.2.2 since $\frac{\left|\bar{B}_q(\gamma_n; U, V)\right|}{n}$ is bounded above by 1 with probability 1 for all n so that convergence in probability implies convergence in mean (Theorem A.1.2). □

Lemma A.2.3. *For any $\gamma > 0$, let*

$$\bar{B}'(\epsilon, 1-\gamma; u, v) := \left\{ f \in F \mid \hat{\xi}^f_{1-\gamma;n} \leq \bar{u} - \epsilon \right\},$$

where $\hat{\xi}^f_{1-\gamma;n}$ is the realized value of the empirical $(1-\gamma)^{th}$ quantile of $U_f = \langle U_{f,w} \rangle_{w \in W_n}$.

Then, for every $\epsilon > 0$ and sequence $\langle \gamma_n \rangle$ such that $\gamma_n \to 0$ and $(\log n) \cdot \gamma_n \to \infty$,

$$E\left[\frac{\left|\bar{B}'(\epsilon, 1-\gamma_n; U, V)\right|}{n}\right] \to 0 \quad \textit{as} \quad n \to \infty.$$

Proof. For each n, let $f_n \in F_n$ and consider the resulting sequence $\langle f_n \rangle_{n=1}^{\infty}$. Note that

$$\begin{aligned}E\left[\frac{|\bar{B}'(\epsilon, 1-\gamma_n; U, V)|}{n}\right] &= \frac{1}{n} \sum_{f \in F_n} E\left[\mathbf{1}\left\{\hat{\xi}^f_{1-\gamma_n;n} \leq \bar{u} - \epsilon\right\}\right] \\ &= E\left[\mathbf{1}\left\{\hat{\xi}^{f_n}_{1-\gamma_n;n} \leq \bar{u} - \epsilon\right\}\right] \\ &= P\left(\hat{\xi}^{f_n}_{1-\gamma_n;n} \leq \bar{u} - \epsilon\right).\end{aligned}$$

Thus, it is enough to show that

$$P\left(\hat{\xi}^{f_n}_{1-\gamma_n;n} \leq \bar{u} - \epsilon\right) \to 0, \quad \text{as} \quad n \to \infty.$$

Take any q from the interval $\left(\Gamma^W(\bar{u} - \epsilon), 1\right)$ such that q^{th} quantile ξ_q is the unique solution x of $\Gamma^W(x-) \leq q \leq \Gamma^W(x)$.[2] For any large n, we have $1-\gamma_n > q$, and thus

$$\hat{\xi}^{f_n}_{1-\gamma_n;n} \geq \hat{\xi}^{f_n}_{q;n}.$$

Therefore, we shall write

$$\begin{aligned} P\left(\hat{\xi}^{f_n}_{1-\gamma_n;n} \leq \bar{u}-\epsilon\right) &\leq P\left(\hat{\xi}^{f_n}_{q;n} \leq \bar{u}-\epsilon\right) \\ &= P\left(\left|\hat{\xi}^{f_n}_{q;n} - \xi_q\right| \geq \xi_q - (\bar{u}-\epsilon)\right), \end{aligned}$$

which converges to 0 by Theorem A.1.4. □

We complete the proof of Proposition A.2.1 using the following observation. For each $\langle F, W, u, v\rangle$ and for every sequence $\langle \gamma_n \rangle$ such that $\gamma_n \to 0$ and $(\log n)\cdot \gamma_n \to \infty$,

$$\begin{aligned} \bar{B}(\epsilon; u, v) &= \left(\bar{B}(\epsilon; u, v) \cap \bar{B}_q(\gamma_n; u, v)\right) \cup \left(\bar{B}(\epsilon; u, v) \cap (F \backslash \bar{B}_q(\gamma_n; u, v))\right) \\ &\subset \bar{B}_q(\gamma_n; u, v) \cup \left(\bar{B}(\epsilon; u, v) \cap (F \backslash \bar{B}_q(\gamma_n; u, v))\right). \end{aligned}$$

Each f in $F\backslash \bar{B}_q(\gamma_n; u, v)$ matches in μ_W with a worker of a normalized rank less than γ_n. Nevertheless if f obtains utility less than $\bar{u}-\epsilon$ in μ_W (i.e. $f \in \bar{B}(\epsilon; u, v)$), then the realized empirical $(1-\gamma_n)^{th}$ quantile of his utilities is below $\bar{u}-\epsilon$.

That is,

$$\bar{B}(\epsilon; u, v) \cap F\backslash \bar{B}_q(\gamma_n; u, v) \subset \bar{B}'(\epsilon, 1-\gamma_n; u, v),$$

and therefore

$$\bar{B}(\epsilon; u, v) \subset \bar{B}_q(\gamma_n; u, v) \cup \bar{B}'(\epsilon, 1-\gamma_n; u, v).$$

We proved in Lemma A.2.2 and A.2.3 that both $\frac{|\bar{B}_q(\gamma_n; U, V)|}{n}$ and $\frac{|\bar{B}'(\epsilon, 1-\gamma_n; U, V)|}{n}$ converge to 0 in mean, which completes the proof.

[2] There exists such a q. For every q in $\left(\Gamma^W(\bar{u}-\epsilon), 1\right)$, we have x_q in $(\bar{u}-\epsilon, \bar{u})$ such that $\Gamma^W(x_q) = q$ by Intermediate Value Theorem. Suppose toward contradiction that every q has two distinct $\underline{x}_q$ and $\bar{x}_q$ in $(\bar{u}-\epsilon, \bar{u})$ such that $\Gamma^W(\underline{x}_q) = \Gamma^W(\bar{x}_q) = q$. Since $\Gamma^W(\cdot)$ is a distribution, every q then has a closed interval $[\underline{x}_q, \bar{x}_q]$ such that $\Gamma^W(x) = q$ for all $x \in [\underline{x}_q, \bar{x}_q]$. Moreover, if $q \neq q'$, then $[\underline{x}_q, \bar{x}_q]$ and $[\underline{x}_{q'}, \bar{x}_{q'}]$ are disjoint. There are uncountable number of elements in $\left(\Gamma^W(\bar{u}-\epsilon), 1\right)$, whereas there are at most countable number of closed disjoint intervals in $(\bar{u}-\epsilon, \bar{u})$.

A.3 Proof of Theorem 1 $(0 < \lambda < 1)$.

To simplify notations, we compress λ and $1-\lambda$, and consider utilities defined as

$$U_{f,w} = U^o_w + \zeta_{f,w} \quad \text{and} \quad V_{f,w} = V^o_f + \eta_{f,w}.$$

We do not lose generality since we can regard common-values and private-values as the ones already multiplied by λ and $1-\lambda$, respectively.

Let U^o_n and V^o_n be i.i.d samples of size n from distributions G^W and G^F, respectively. G^W and G^F have strictly positive density functions on supports in $\mathbb{R}_+$. $\zeta = [\zeta_{f,w}]$ is an i.i.d sample from a continuous distribution Γ^W with support $[0, \bar{u}]$, and $\eta = [\eta_{f,w}]$ is an i.i.d sample from a continuous distribution Γ^F with support $[0, \bar{v}]$.[3]

We define

$$B^F(\epsilon; u, v) := F \backslash A^F(\epsilon; u, v) = \{f \in F \mid \Delta(f; u, v) \geq \epsilon\}$$

and prove that $\frac{|B^F(\epsilon;U,V)|}{n}$ converges to 0 in probability, which is equivalent to proving convergence to 0 in mean (Theorem A.1.2). That is, we fix $\epsilon > 0$ and $K \in \mathbb{N}$, and prove that

$$P\left(\frac{|B^F(\epsilon; U, V)|}{n} > \frac{9}{K}\right) \to 0, \quad \text{as} \quad n \to \infty.$$

First, we partition the supports of the common-value distributions into K intervals. Then for each market instance, in particular for each realized profile of common-values, we group firms and workers into *two* versions of a finite number of tiers, where agents in the same tier have similar common-values. We first find that tier-k firms are most likely to achieve a utility level higher than an arbitrary ϵ less than the maximal utility achievable from workers in tier-$(k+3)$ (Proposition A.7.1).[4] For the proof, we use techniques from a

[3] When $\lambda > 0$, we can relax this assumption such that each pair of $\zeta_{f,w}$ and $\eta_{f,w}$ is an i.i.d sample from a continuous joint distribution with a bounded support in $\mathbb{R}^2_+$.

[4] In Section 1.4, we showed with a market with tiers that firms in tier-t are most likely to achieve a utility level higher than an arbitrary ϵ less than the maximal utility from a worker in tier-$(k+1)$. In the model with tiers, each tier has a distinct tier-specific common-value, so there is a clear-cut distinction between tier-k and tier-$(k+1)$ specific values. In the general model (without tiers), however, there is no such distinction in common-values between adjacent tiers. The highest common-value of workers in tier-$(k+1)$ can be arbitrarily close to the lowest common-value of workers in tier-k. This leads us to use the maximal utility from a worker in tier-$(k+3)$ rather than tier-$(k+1)$ as an asymptotic lower bound on utilities of tier-k firms.

theory of random bipartite graphs.

Once we find an asymptotic lower bound on utilities of firms in each tier, we find an asymptotic upper bound on utilities of firms in a tier, say k, by referencing to the asymptotic lower bounds on utilities of workers in tiers higher than k (Proposition A.7.2). As workers in higher tiers achieve high utilities, they are most likely to match with firms in high tiers, rather than firms in tier-k. Accordingly, the utilities of tier-k firms are asymptotically bounded above by the maximal utility that they can achieve by matching with workers in tiers near k.

As we finely partition the supports of the common-value distributions, the differences in common-values between adjacent tiers become small. Then, the asymptotic lower bound on utilities of tier-k firms will become close to the maximal utility achievable from workers in tier-k. In addition, the asymptotic upper bound also becomes close to the same level, since the maximal utility achievable from workers in tiers near k will also become close to the maximal utility achievable from workers in tier-k.

We divide the proof into three subsections. First, in Subsection A.7.2, we construct *two* tier structures from each profile of realized common-values. Then in Subsection A.7.3, we define three events related to the tier structures, and show that the all three events occur with a probability converging to 1 as the market becomes large. The real proof begins in subsection A.7.4. During the proof, we shall focus on the market instances where realized firms' or workers' common-values are all distinct. G^F and G^W are continuous, ensuring that realized common-values are all distinct with probability 1.

A.3.1 Tier-Grouping

We use the following notations.

1. ξ_q^F and ξ_q^W : q^{th} quantile of G^F and G^W.
2. $\hat{\xi}_{q;n}^F$ and $\hat{\xi}_{q;n}^W$: empirical q^{th} quantile of samples of size n from distributions G^F and G^W, respectively. We also use $\hat{\xi}_{q;n}^F$ and $\hat{\xi}_{q;n}^W$ to denote their realizations.

Since realized common-values $u_n^o = \langle u_w^o \rangle_{w \in W_n}$ and $v_n^o = \langle v_f^o \rangle_{f \in F_n}$ are all distinct with probability 1, we index firms and workers from $i = 1$ to n in the order of their common-values: i.e.

$$v^o_{f_i} > v^o_{f_j} \quad \text{and} \quad u^o_{w_i} > u^o_{w_j}, \quad \text{if} \quad i < j.$$

Then, $U^o_{w_i;n}$ and $V^o_{f_i;n}$ represent i^{th} highest values of n order statistics from G^W and G^F. Note that $U^o_{w_i;n} = \hat{\xi}^W_{(1-\frac{i-1}{n});n}$ by the relationship between order statistics and empirical quantiles (see Equation (A.1)).

We partition the support of G^W into

$$\begin{aligned}
I^W_1 &:= (\xi^W_{1-\frac{1}{K}}, \infty] \\
I^W_2 &:= (\xi^W_{1-\frac{2}{K}}, \xi^W_{1-\frac{1}{K}}] \\
&\vdots \\
I^W_k &:= (\xi^W_{1-\frac{k}{K}}, \xi^W_{1-\frac{k-1}{K}}] \\
&\vdots \\
I^W_K &:= [0, \xi^W_{\frac{1}{K}}].
\end{aligned}$$

We define **the set of workers in tier-k** (with respect to *workers'* common-values) as

$$W_k(u) := \{w \mid u^o_w \in I^W_k\} \quad \text{for} \quad k = 1, 2, \ldots, K,$$

and define **the set of firms in tier-k** (with respect to *workers'* common-values) as

$$F_k(u) := \{f_i \in F_n \mid w_i \in W_k(u)\}.$$

We will use the following notations.

1. $l_k(u) := |F_k(u)| = |W_k(u)|$: The size of tier-$k$ (with respect to *workers'* common-values).

2. $u^o_k := \xi^W_{1-\frac{k}{K}}$: The threshold level of tier-k and tier-$(k+1)$ workers' common-values. Note, $w \in W_k(u)$ if and only if $u^o_k < u^o_w \leq u^o_{k-1}$.

Remark 2. *The set of tier-k workers is defined with respect to workers' common-values, which is a random sample. Therefore, $W_k(U)$ is random, and so is $F_k(U)$. In particular,* *the size of tier-k, $l_k(U)$, is random; whereas, u_k^o is a constant.*

In parallel, we partition the support of G^F into

$$\begin{aligned} I_1^F &:= (\xi^F_{1-\frac{1}{K}}, \infty] \\ I_2^F &:= (\xi^F_{1-\frac{2}{K}}, \xi^F_{1-\frac{1}{K}}] \\ &\vdots \\ I_k^F &:= (\xi^F_{1-\frac{k}{K}}, \xi^F_{1-\frac{k-1}{K}}] \\ &\vdots \\ I_K^F &:= [0, \xi^F_{\frac{1}{K}}]. \end{aligned}$$

We define **the set of firms in tier-k** (with respect to *firms'* common-values) as

$$F_k(v) := \left\{ f \mid v_f^o \in I_k^F \right\} \quad \text{for} \quad k = 1, 2, \ldots, K,$$

and define **the set of workers in tier-k** (with respect to *firms'* common-values) as

$$W_k(v) := \{ w_i \in W_n \mid f_i \in F_k(v) \}.$$

Accordingly, we use the following notations.

1. $l_k(v) := |F_k(v)| = |W_k(v)|$: The size of tier-$k$ (with respect to *firms'* common-values).
2. $v_k^o := \xi^F_{1-\frac{k}{K}}$: The threshold level of tier-k and tier-$(k+1)$ firms' common-values. Note, $f \in F_k(u)$ if and only if $v_k^o < v_f^o \leq v_{k-1}^o$.

Remark 3. *Tiers with respect to workers' common-values are in general not the same as tiers with respect to firms' common-values. In particular, we are most likely to have $l_k(u) \neq l_k(v)$.*

Throughout the proof, we mainly use tiers defined with respect to workers' common-values. However, we need both tier structures in the last part of the proof. We simply write "tier-k" to denote tier-k with respect to workers' common-values, and use "(w.r.t firm) tier-k" to denote tier-k with respect to firms' common-values.

A.3.2 High-Probability Events

We introduce three events and show that the events occur with probabilities converging to 1 as the market becomes large. We provide proofs for completeness, but the main ideas are simply from the (weak) law of large numbers. In the next section, we will leave the probability that the following events do not occur as a remainder term converging to zero, and focus on the probabilities conditioned that the following events all occur.

A.3.2.1 No vanishing tiers

Event 1 ($\mathcal{E}_1$)**.** *Let $\bar{K} > K$. For all $k = 1, 2, \ldots, K$,*

$$\frac{l_k(U)}{n} > \frac{1}{\bar{K}}.$$

Proof. By definition,

$$\frac{l_k(U)}{n} := \frac{1}{n} \sum_{w \in W_n} \mathbf{1}\{U_w^o \in I_k^W\},$$

which converges to $\frac{1}{K}$ in probability by the (weak) law of large numbers. □

A.3.2.2 Distinct common-values of the firms in non-adjacent tiers.

Let $\tilde{\epsilon} > 0$ be such that for any $v, v' \in [0, \xi^F_{1-1/K}]$ and $|v - v'| \leq \tilde{\epsilon}$,

$$|G^F(v) - G^F(v')| < \frac{1}{3K}.$$

There exists such an $\tilde{\epsilon}$ since G^F is uniformly continuous on $[0, \xi^F_{1-1/K}]$.

Event 2 ($\mathcal{E}_2$)**.** *For every $k = 1, 2, \ldots, K-2$,*

$$\min_{\substack{f \in F_k(U) \\ f' \in F_{k+2}(U)}} |V_f^o - V_{f'}^o| > \tilde{\epsilon}.$$

Proof. Fix $k \in 1, 2, \ldots, K-2$ and realized u. For every $w_i \in W_k(u)$ and $w_j \in W_{k+2}(u)$,

$$u_{w_i}^o > u_k^o = \xi^W_{1-\frac{k}{K}}, \quad \text{and} \quad u_{w_j}^o \leq u_{k+1}^o = \xi^W_{1-\frac{k+1}{K}}. \tag{A.4}$$

For any $q \in (0,1)$, $\hat{\xi}^W_{q;n} \xrightarrow{p} \xi^W_q$ (Theorem A.1.4), from which the following inequalities hold with probability converging to 1 as $n \to \infty$.

$$\xi^W_{1-\frac{k}{K}} > \hat{\xi}^W_{1-\frac{k}{K}-\frac{1}{4K}} \quad \text{and} \quad \xi^W_{1-\frac{k+1}{K}} < \hat{\xi}^W_{1-\frac{k+1}{K}+\frac{1}{4K}}. \tag{A.5}$$

Considering (A.21) and the relation between order statistics and empirical quantiles (Equation (A.1)), if (A.22) holds, we have

$$1 - \frac{k}{K} - \frac{1}{4K} < \min_{w_i \in W_k(u)} \left(1 - \frac{i-1}{n}\right) = \min_{f_i \in F_k(u)} \left(1 - \frac{i-1}{n}\right)$$

and

$$1 - \frac{k+1}{K} + \frac{1}{4K} > \max_{w_j \in W_{k+2}(u)} \left(1 - \frac{j-1}{n}\right) = \max_{f_j \in F_{k+2}(u)} \left(1 - \frac{j-1}{n}\right).$$

Then for every $f_i \in F_k(u)$ and $f_j \in F_{k+2}(u)$,

$$v^o_{f_i} > \hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{4K}} \quad \text{and} \quad v^o_{f_j} < \hat{\xi}^F_{1-\frac{k+1}{K}+\frac{1}{4K}}.$$

Therefore,

$$\begin{aligned} P\Big(\inf_{\substack{f_i \in F_k(U) \\ f_j \in F_{k+2}(U)}} \left|V^o_{f_i} - V^o_{f_j}\right| \leq \tilde{\epsilon}\Big) &\leq P\left(\left|\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{4K}} - \hat{\xi}^F_{1-\frac{k+1}{K}+\frac{1}{4K}}\right| \leq \tilde{\epsilon}\right) + R_n \\ &\leq P\left(\left|G^F(\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{4K}}) - G^F(\hat{\xi}^F_{1-\frac{k+1}{K}+\frac{1}{4K}})\right| < \frac{1}{3K}\right) + R_n, \end{aligned} \tag{A.6}$$

where R_n corresponds to the probability that (A.22) is violated: i.e. $R_n \to 0$. The last inequality is by the definition of $\tilde{\epsilon}$.

Note that

$$G^F(\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{4K}}) - G^F(\hat{\xi}^F_{1-\frac{k+1}{K}+\frac{1}{4K}}) \xrightarrow{p} \frac{1}{2K}$$

by Theorem A.1.4 and continuity of G^F(Theorem A.1.3). As a result, the right hand side of (A.23) converges to 0.

□

A.3.2.3 Similarity between tiers w.r.t workers' common-values and tiers w.r.t firms' common-values

The following event is the case that all firms in tier-k with respect to workers' common-values are in a tier near k with respect to firms' common-values, and vice versa.

Event 3 ($\mathcal{E}_3$)**.** *For every* $k = 1, 2, 3, \ldots, K$*,*

$$F_k(U) \subset \bigcup_{k'=k-1}^{k+1} F_{k'}(V) \quad \textit{and} \quad W_k(V) \subset \bigcup_{k'=k-1}^{k+1} W_{k'}(U).^5$$

Proof. We prove the first part and omit the proof of the second part.

For each realized (u, v), we have

$$\{u^o_w | w \in W_k(u)\} \subset \left(u^o_k, u^o_{k-1}\right] = \left(\xi^W_{1-\frac{k}{K}}, \xi^W_{1-\frac{k-1}{K}}\right]. \tag{A.7}$$

Suppose

$$\left(\xi^W_{1-\frac{k}{K}}, \xi^W_{1-\frac{k-1}{K}}\right] \subset \left(\hat{\xi}^W_{1-\frac{k}{K}-\frac{1}{2K}}, \hat{\xi}^W_{1-\frac{k-1}{K}+\frac{1}{2K}}\right], \tag{A.8}$$

and

$$\left(\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{2K}}, \hat{\xi}^F_{1-\frac{k-1}{K}+\frac{1}{2K}}\right] \subset \left(\xi^F_{1-\frac{k+1}{K}}, \xi^F_{1-\frac{k-2}{K}}\right]. \tag{A.9}$$

If (A.25) hold, then (A.24) implies that for every tier-k worker w_i, we have

$$u^o_{w_i} \in \left(\hat{\xi}^W_{1-\frac{k}{K}-\frac{1}{2K}}, \hat{\xi}^W_{1-\frac{k-1}{K}+\frac{1}{2K}}\right],$$

and thus,

$$1 - \frac{i-1}{n} \in \left(1 - \frac{k}{K} - \frac{1}{2K}, 1 - \frac{k-1}{K} + \frac{1}{2K}\right].$$

Then for any tier-k firm f_i, we have

$$v^o_{f_i} \in \left(\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{2K}}, \hat{\xi}^F_{1-\frac{k-1}{K}+\frac{1}{2K}}\right],$$

[5] We simply assume that $F_0(V)$, $F_0(V)$, $W_{K+1}(U)$, and $W_{K+1}(U)$ are empty sets.

which implies that

$$\left\{v_f^o \mid f \in F_k(u)\right\} \subset \left(\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{2K}}, \hat{\xi}^F_{1-\frac{k-1}{K}+\frac{1}{2K}}\right].$$

Consequently if both (A.25) and (A.26) hold, then

$$\begin{aligned}\left\{v_f^o \mid f \in F_k(u)\right\} &\subset \left(\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{2K}}, \hat{\xi}^F_{1-\frac{k-1}{K}+\frac{1}{2K}}\right] \\ &\subset \left(\xi^F_{1-\frac{k+1}{K}}, \xi^F_{1-\frac{k-2}{K}}\right] \\ &= \bigcup_{k'=k-1}^{k+1} I^F_{k'}.\end{aligned}$$

In other words,

$$F_k(u) \subset \bigcup_{k'=k-1}^{k+1} F_{k'}(v).$$

(A.25) and (A.26) occur with probability converging to 1 (Theorem A.1.4), and thus the event $\mathcal{E}_3$ also occurs with probability converging to 1. □

A.3.3 Proof of the Theorem 1

We choose K large enough that

$$\max_{1\le k\le K-1}\left|u_k^o - u_{k+1}^o\right| \equiv \max_{1\le k\le K-1}\left|\xi^W_{1-\frac{k}{K}} - \xi^W_{1-\frac{k+1}{K}}\right| < \frac{\epsilon}{9}.^{6} \tag{A.10}$$

We divide the proof into two propositions. The first proposition finds an asymptotic lower bound on utilities of firms in each tier, using techniques from the theory of random bipartite graphs. Similarly, we have a proposition for an asymptotic lower bound on utilities of workers in each tier. The second proposition derives an asymptotic upper bound on utilities of firms in each tier, by referencing the lower bounds on utilities of workers in higher tiers. The Theorem 1 follows from the fact that the lower bound and the upper bound are close to each other.

[6] We can always satisfy the condition since G^W has a strictly positive density function.

Proposition A.3.1. *For each instance $\langle F, W, u, v\rangle$ and for each $\bar{k} = 1, 2, \ldots, K-2$, define*

$$\hat{B}^F_{\bar{k}}(\epsilon; u, v) := \left\{ f \in F_{\bar{k}}(u) : u_{\mu_W}(f) \leq u^o_{\bar{k}+2} + \bar{u} - \epsilon \right\}.^{7}$$

Then for any $\epsilon > 0$,

$$\frac{|\hat{B}^F_{\bar{k}}(\epsilon; U, V)|}{n} \xrightarrow{p} 0 \quad \textit{as} \quad n \to \infty.$$

Proof. For each instance $\langle F, W, u, v\rangle$ and for each $k = 1, 2, \ldots, K$, let $F_{\leq k}(u) := \bigcup_{k' \leq k} F_{k'}(u)$ and $F_{<k}(u) := \bigcup_{k'<k} F_{k'}(u)$. Similarly, we define $W_{\leq k}(u)$ and $W_{<k}(u)$.

Take any $\bar{k}$ from $\{1, 2, \ldots, K-2\}$. We construct a bipartite graph with $F_{\bar{k}}(u) \cup W_{\leq \bar{k}+2}(u)$ as a partitioned set of nodes. (see Section 3 for the related definitions.) Two vertices $f \in F_{\bar{k}}(u)$ and $w \in W_{\leq \bar{k}+2}(u)$ are joined by an edge if and only if

$$\zeta_{f,w} \leq \bar{u} - \epsilon \quad \text{or} \quad \eta_{f,w} \leq \bar{v} - \tilde{\epsilon},$$

where $\tilde{\epsilon}$ is the value taken before, while defining $\mathcal{E}_2$.

Let $\bar{W}_{\leq \bar{k}+2}(u, v)$ be the set of workers in tiers up to $\bar{k}+2$ who are not matched with firms in tiers up to $\bar{k}+1$ in μ_W. That is,

$$\bar{W}_{\leq \bar{k}+2}(u, v) := \left\{ w \in W_{\leq \bar{k}+2}(u) \,|\, \mu_W(w) \notin F_{\leq \bar{k}+1}(u) \right\}.$$

We now show that if $\mathcal{E}_2$ holds, then

$$\hat{B}^F_{\bar{k}}(\epsilon; u, v) \cup \bar{W}_{\leq \bar{k}+2}(u, v)$$

is a biclique.

Suppose, towards a contradiction, that a pair of $f \in \hat{B}^F_{\bar{k}}(\epsilon; u, v)$ and $w \in \bar{W}_{\leq \bar{k}+2}(u, v)$ is *not* joined by an edge: i.e.

$$\zeta_{f,w} > \bar{u} - \epsilon \quad \text{and} \quad \eta_{f,w} > \bar{v} - \tilde{\epsilon}.$$

[7] Note that $u^o_{\bar{k}+2} + \bar{u}$ is the maximal utility level a firm can achieve by matching with a worker in tier-$(\bar{k}+3)$.

Then, we first have

$$u_{f,w} = u^o_w + \zeta_{f,w} > u^o_{\bar{k}+2} + \zeta_{f,w} > u^o_{\bar{k}+2} + \bar{u} - \epsilon, \tag{A.11}$$

and also have

$$v_{f,w} = v^o_f + \eta_{f,w} \geq \min_{f' \in F_{\bar{k}}(u)} v^o_{f'} + \eta_{f,w} > \min_{f' \in F_{\bar{k}}(u)} v^o_{f'} + \bar{v} - \tilde{\epsilon}.^8$$

Conditioned on $\mathcal{E}_2$, we can proceed further and obtain

$$\begin{aligned} v_{f,w} &> \min_{f' \in F_{\bar{k}}(u)} v^o_{f'} + \bar{v} - \left(\min_{f' \in F_{\bar{k}}(u)} v^o_{f'} - \max_{f'' \in F_{\bar{k}+2}(u)} v^o_{f''} \right) \\ &= \max_{f'' \in F_{\bar{k}+2}(u)} v^o_{f''} + \bar{v}. \end{aligned} \tag{A.12}$$

On the other hand, $f \in \hat{B}^F_{\bar{k}}(\epsilon; u, v)$ implies that

$$u_{\mu_W}(f) \leq u^o_{\bar{k}+2} + \bar{u} - \epsilon,$$

and $w \in \bar{W}_{\leq \bar{k}+2}(u, v)$ implies that

$$v_{\mu_W}(w) \leq \max_{f'' \in F_{\bar{k}+2}(u)} v^o_{f''} + \bar{v},$$

since a worker can obtain utility higher than $\max_{f'' \in F_{\bar{k}+2}(u)} v^o_{f''} + \bar{v}$ only by matching with a firm in $F_{\leq \bar{k}+1}(u)$.

Then, (A.28) and (A.29) implies that (f, w) would have blocked μ_W, contradicting that μ_W is stable. Therefore,

$$\hat{B}^F_{\bar{k}}(\epsilon; u, v) \cup \bar{W}_{\leq \bar{k}+2}(u, v).$$

is a biclique, which is not necessarily balanced.

We now control the size of $\hat{B}^F_{\bar{k}}(\epsilon; U, V)$ by referencing Theorem 3. Let u^o and v^o be realized common-values such that events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. Then, the remaining randomness

[8] We should not replace $\min_{f' \in F_{\bar{k}}(u)} v^o_{f'}$ with $v^o_{\bar{k}}$. $F_{\bar{k}}(u)$ is defined with respect to workers' common-values, rather than firms' common-values.

of U and V is from ζ and η. Consider a random bipartite graph with $F_{\bar{k}}(U) \cup W_{\leq \bar{k}+2}(U)$ as a bi-partitioned set of nodes, where each pair of $f \in F_{\bar{k}}(U)$ and $w \in W_{\leq \bar{k}+2}(U)$ is joined by an edge if and only if

$$\zeta_{f,w} \leq \bar{u} - \epsilon \quad \text{or} \quad \eta_{f,w} \leq \bar{v} - \tilde{\epsilon}.$$

In other words, every pair is joined by an edge independently with probability

$$p(\epsilon) = 1 - \left(1 - \Gamma^W(\bar{u} - \epsilon)\right) \cdot \left(1 - \Gamma^F(\bar{v} - \tilde{\epsilon})\right).$$

We write $\beta(n) := 2 \cdot \log(l_{\leq \bar{k}+2}(U)) / \log \frac{1}{p(\epsilon)}$, and show that

$$P\left(|\hat{B}^F_{\bar{k}}(\epsilon; U, V)| \leq \beta(n)\right) \to 1 \quad \text{as} \quad n \to \infty.^9$$

First, observe that $\bar{W}_{\leq \bar{k}+2}(U, V)$ is the size of at least $l_{\bar{k}+2}(U)$. Among $l_{\leq \bar{k}+2}(U)$ workers in tiers up to $\bar{k}+2$ at most $l_{\leq \bar{k}+1}(U)$ are matched with firms in tiers up to $\bar{k}+1$. In addition, $l_{\bar{k}+2}(U) > \beta(n)$ with large n, since $\mathcal{E}_1$ holds. Therefore, with large n, we shall write

$$P\left(|\hat{B}^F_{\bar{k}}(\epsilon; U, V)| \leq \beta(n)\right) = P\left(\min\left\{|\hat{B}^F_{\bar{k}}(\epsilon; U, V)|, |\bar{W}_{\leq \bar{k}+2}(U, V)|\right\} \leq \beta(n)\right). \tag{A.13}$$

Let $\alpha(U, V) \times \alpha(U, V)$ be the size of a maximum balance biclique of the random graph

$$G\left(F_{\bar{k}}(U) \cup W_{\leq \bar{k}+2}(U)\ ,\ p(\epsilon)\right).$$

Since every realized $\hat{B}^F_{\bar{k}}(\epsilon; u, v) \cup \bar{W}_{\leq \bar{k}+2}(u, v)$ is a biclique, it contains a balanced biclique of the size equals to

$$\min\left\{|\hat{B}^F_{\bar{k}}(\epsilon; u, v)|\ ,\ |\bar{W}_{\leq \bar{k}+2}(u, v)|\right\}.$$

Therefore,

$$P\left(\min\left\{|\hat{B}^F_{\bar{k}}(\epsilon; U, V)|, |\bar{W}_{\leq \bar{k}+2}(U, V)|\right\} \leq \beta(n)\right) \geq P\left(\alpha(U, V) \leq \beta(n)\right). \tag{A.14}$$

[9] Note that we fixed common-values as a realization u^o and v^o such that the events $\mathcal{E}_1$ and $\mathcal{E}_2$ occur. Thus for now, the tier-structure is deterministic, and $\beta(n)$ is, in turn, a deterministic sequence.

Applying Theorem 3 to (A.14) and using (A.13),

$$P\left(|\hat{B}^F_{\bar{k}}(\epsilon; U, V)| \leq \beta(n)\right) \geq P\left(\alpha(U, V) \leq \beta(n)\right) \to 1. \tag{A.15}$$

Lastly, we consider random utilities U and V, in which common-values are yet realized. For every $\epsilon' > 0$,

$$\begin{aligned} P\left(\frac{|\hat{B}^F_{\bar{k}}(\epsilon; U, V)|}{n} > \epsilon'\right) &= P\left(|\hat{B}^F_{\bar{k}}(\epsilon; U, V)| > \epsilon' \cdot n\right) \\ &\leq P\left(|\hat{B}^F_{\bar{k}}(\epsilon; U, V)| > \beta(n) \mid \mathcal{E}_1, \mathcal{E}_2\right) + R_n, \quad \text{with large } n, \end{aligned}$$

where R_n is the probability that either $\mathcal{E}_1$ or $\mathcal{E}_2$ does not hold: i.e. $R_n \to 0$. The inequality is from the fact that $\epsilon' \cdot n > \beta(n)$ with large n. We complete the proof by applying (A.15). □

We also obtain the counterpart proposition of Proposition A.7.1 in terms of tiers defined with respect to firms' common-values.

Proposition A.7.1* *For each $\bar{k} = 1, 2, \ldots, K-2$, define*

$$\hat{B}^W_{\bar{k}}(\epsilon; u, v) := \left\{ w \in W_{\bar{k}}(v) | v_{\mu_F}(w) \leq v^o_{\bar{k}+2} + \bar{v} - \epsilon \right\}.$$

Then for any $\epsilon > 0$,

$$\frac{|\hat{B}^W_{\bar{k}}(\epsilon; U, V)|}{n} \xrightarrow{p} 0 \quad as \quad n \to \infty.$$

Proof. We omit the proof since it is analogous to the proof of Proposition A.7.1. □

For each instance $\langle F, W, u, v \rangle$ and for each $\bar{k} = 1, 2, \ldots, K$, we define

$$B^F_{\bar{k}}(\epsilon; u, v) := \{ f \in F_{\bar{k}}(u) | \Delta(f; u, v) \geq \epsilon \}.$$

Proposition A.3.2. *If $\bar{k} = 7, 8, \ldots, K-2$, then for any $\epsilon > 0$,*

$$\frac{|B^F_{\bar{k}}(\epsilon; U, V)|}{n} \xrightarrow{p} 0 \quad as \quad n \to \infty.$$

Proof. In Proposition A.7.1* with $k = 1, 2, \dots, K-3$, we replace ϵ with

$$\epsilon_k := v^o_{k+2} - v^o_{k+3},$$

and write

$$\hat{B}^W_k(\epsilon_k; u, v) = \left\{ w \in W_k(v) | v_{\mu_F}(w) \le v^o_{k+3} + \bar{v} \right\}.^{10}$$

Then,

$$\frac{|\hat{B}^W_k(\epsilon_k; U, V)|}{n} \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty. \tag{A.16}$$

Note that a worker receives utility higher than $v^o_{k+3} + \bar{v}$ only by matching with a firm in (w.r.t firm) tiers up to $k+3$.[11] Thus for $k = 5, 6, \dots, K$,

$$\{w \in W_{\le k-4}(V) : \mu(w) \in F_k(V)\} \subset \bigcup_{k'=1}^{k-4} \hat{B}^W_{k'}(\epsilon_{k'}; U, V). \tag{A.17}$$

If event $\mathcal{E}_3$ holds, we can translate (A.34) into an expression with tiers w.r.t workers' common-values. That is, for $k = 7, 8, \dots, K$,

$$\begin{aligned} \{w \in W_{\le k-6}(U) : \mu_F(w) \in F_k(U)\} &\subset \bigcup_{k'=k-1}^{k+1} \{w \in W_{\le k-6}(U) : \mu_F(w) \in F_{k'}(V)\} \\ &\subset \bigcup_{k'=k-1}^{k+1} \{w \in W_{\le k-5}(V) : \mu_F(w) \in F_{k'}(V)\} \\ &\subset \bigcup_{k'=k-1}^{k+1} \{w \in W_{\le k'-4}(V) : \mu_F(w) \in F_{k'}(V)\} \end{aligned}$$

where the first and second inequalities are from $\mathcal{E}_3$.

By applying (A.34), we obtain

$$\{w \in W_{\le k-6}(U) : \mu_F(w) \in F_k(U)\} \subset \bigcup_{k'=1}^{k-3} \hat{B}^W_{k'}(\epsilon_{k'}; U, V).$$

[10] Recall that v^o_k is a constant, defined as $v^o_k := \xi^F_{1-\frac{k}{K}}$.

[11] Recall that $f \in F_k(v)$ if and only if $v^o_k < v^o_f \le v^o_{k-1}$. Thus, if $f \in F_{>k+3}(v)$ then $v^o_f \le v^o_{k+3}$.

It follows that

$$\frac{|\{f \in F_k(U) : \mu_F(f) \in W_{\leq k-6}(U)\}|}{n} \xrightarrow{p} 0, \tag{A.18}$$

because for every $\epsilon > 0$,

$$P\left(\frac{|\{f \in F_k(U) : \mu_F(f) \in W_{\leq k-6}(U)\}|}{n} > \epsilon\right) \leq P\left(\sum_{k'=1}^{k-3} \frac{|\hat{B}_{k'}^W(\epsilon_{k'}; U, V)|}{n} > \epsilon\right) + R_n,$$

where R_n is the probability that $\mathcal{E}_3$ does not hold: i.e. $R_n \to 0$. The right hand side converges to 0 by (A.33).

We complete the proof of Proposition A.7.2 by proving the following claim. Proposition A.7.1 and (A.35) show that the normalized sizes of two sets on the right hand side of (A.36) converge to 0 in probability.

Claim A.3.1. *For $\bar{k} = 7, 8, \ldots, K-2$ and each instance $\langle F, W, u, v \rangle$,*

$$B_{\bar{k}}^F(\epsilon; u, v) \subset \hat{B}_{\bar{k}}^F(\epsilon/9; u, v) \cup \left\{f \in F_{\bar{k}}(u) | \mu_F(f) \in W_{\leq \bar{k}-6}(u)\right\}. \tag{A.19}$$

Proof of Claim A.7.1. If a firm $f \in F_{\bar{k}}(u)$ is *not* in $\hat{B}_{\bar{k}}^F(\epsilon/9; u, v)$, then

$$u_{\mu_W}(f) > u_{\bar{k}+2}^o + \bar{u} - \epsilon/9,$$

and if the firm f is *not* in $\left\{f \in F_{\bar{k}}(u) | \mu_F(f) \in W_{\leq \bar{k}-6}(u)\right\}$, then

$$u_{\mu_F}(f) \leq u_{\bar{k}-6}^o + \bar{u}.$$

Therefore, using (A.27) we obtain

$$u_{\mu_F}(f) - u_{\mu_W}(f) \leq u_{\bar{k}-6}^o - u_{\bar{k}+2}^o + \epsilon/9 < \epsilon,$$

and thus f is *not* in $B_{\bar{k}}^F(\epsilon; u, v)$.

□

□

Lastly, we complete the proof of Theorem 1 by the following inequalities.

$$
\begin{aligned}
P\left(\frac{|B^F(\epsilon;U,V)|}{n} > \frac{9}{K}\right) &= P\left(\sum_{1\le k\le K}\frac{|B_k^F(\epsilon;U,V)|}{n} > \frac{9}{K}\right)\\
&< P\left(\sum_{7\le k\le K-2}\frac{|B_k^F(\epsilon;U,V)|}{n} + \sum_{k=1,\ldots,6,K-1,K}\frac{l_k(U)}{n} > \frac{9}{K}\right).
\end{aligned}
$$

The last probability converges to 0. For each $k = 7,\ldots,K-2$, the proportion $\frac{|B_k^F(\epsilon;U,V)|}{n}$ converges to 0 in probability (Proposition A.7.2). For each $k = 1,\ldots,6,K-1,K$, the proportion $\frac{l_k(U)}{n}$ converges to $\frac{1}{K}$ in probability by the (weak) law of large numbers.

A.4 Proof of Theorem 4

For each $\epsilon > 0$, we first define

$$
B_E^F(\epsilon;u,v) := F\backslash A_E^F(\epsilon;u,v) = \{f \in F \mid \Delta_E(f;u,v) \ge \epsilon\},
$$

and show that

$$
E\left[\frac{\left|B_E^F(\epsilon;U,V)\right|}{n}\right] \to 0 \quad \text{as} \quad n\to\infty.
$$

For each n, let $f_n \in F_n$ and consider the resulting sequence $\langle f_n\rangle_{n=1}^{\infty}$. For any $\epsilon > 0$,

$$
\begin{aligned}
E\left[\frac{|B_E^F(\epsilon;U,V)|}{n}\right] &= E\left[\mathbf{1}\{f_n \in B_E^F(\epsilon;U,V)\}\right]\\
&= P\left(\Delta_E(f_n;U,V) \ge \epsilon\right).
\end{aligned}
$$

Thus if $\Delta_E(f_n;U,V) \xrightarrow{p} 0$, then for every ϵ, $\frac{|B_E^F(\epsilon,U,V)|}{n}$ converges to zero in mean, thereby completing the proof.

Claim A.4.1.

$$
\Delta_E(f_n;U,V) \xrightarrow{p} 0, \quad \textit{as} \quad n\to\infty.
$$

Proof. For every $\epsilon > 0$,

$$\begin{aligned} P\left(\Delta(f_n; U, V) \geq \epsilon\right) &= E\left[\mathbf{1}\{\Delta(f_n; U, V) \geq \epsilon\}\right] \\ &= E\left[\frac{|F \backslash A^F(\epsilon; U, V)|}{n}\right]. \end{aligned}$$

The last term converges to 0 by Theorem 1, and thus $\Delta(f_n; U, V) \xrightarrow{p} 0$.

Let $\bar{u}^o$ and $\bar{u}$ be upper bounds of common-value distribution and private-value distribution of workers, respectively. Then, $\Delta(f_n; U, V)$ is bounded above by $\lambda\,\bar{u}^o + (1-\lambda)\bar{u}$ with probability 1. We obtain by Theorem A.1.2 that

$$\lim_{n\to\infty} E[\Delta_E(f_n; U, V)] = \lim_{n\to\infty} E\left[E\left[\Delta(f_n; U, V)|\Pi_{f_n}\right]\right] = \lim_{n\to\infty} E\left[\Delta(f_n; U, V)\right] = 0.$$

The Claim A.4.1 follows by Theorem A.1.1. □

A.5 Additional Simulations on the Proportion of Unmatched Agents

The simulation results in Section 1.1.4 show that the short preference condition assumed in Roth and Peranson (1999), Immorlica and Mahdian (2005), and Kojima and Pathak (2009) may leave most agents in a large market unmatched in stable matchings. It is worth noting that random preferences in the previous simulations were generated by the setup of our model, rather than the previous studies' model. That is, the previous simulations do not directly represent features of previous models. In this section, we show the increasing proportions of unmatched agents with simulations based on the previous studies' model.

Let L be the maximum number of firms that each worker considers acceptable. We generate random preferences following the previous model, in particular Immorlica and Mahdian (2005). Immorlica and Mahdian studied one-to-one matching markets with generally distributed random preferences. For each market size n, a market is given two underlying distributions, one for firms and the other for workers, called *popularity distributions.*[12] A

[12] Immorlica and Mahdian (2005) construct random preferences only for workers: firms' preferences are arbitrarily given. In our simulation, we also generate firms' preferences randomly, rather than assuming

worker's preference list is constructed by sequentially sampling L firms from the popularity distribution without replacement. The firm chosen first is the most preferred, and the next chosen firm becomes the second most preferred. We similarly construct firms' preferences, except that firms' preferences are of length n: i.e., all workers are acceptable.

We use two classes of popularity distributions.

1. Normalized geometric distribution

   For each market size n, we define the normalized geometric distribution as:

   $$\text{PDF} : p_k = \frac{(1-q)^k}{\sum_{k'=1}^{n}(1-q)^{k'}}, \quad (0 \leq q < 1,\ k = 1, 2, \ldots, n).$$

   Consider a pair of firms, f_{k_1} and f_{k_2} ($k_1 < k_2 \leq n$). For each worker, the probability of choosing f_{k_1} before f_{k_2}, conditioned on at least one of the firms chosen, equals to

   $$\frac{(1-q)^{k_1}}{(1-q)^{k_1} + (1-q)^{k_2}} = \frac{1}{1 + (1-q)^{k_2 - k_1}}.$$

   which is independent of the market size n. If $q = 0$, we have the uniform popularity distribution over firms, so all firms have an equal chance of being chosen before another. As q becomes close to 1, more popular firms have higher chances of being chosen before other firms, which generates a commonality of preferences among workers.

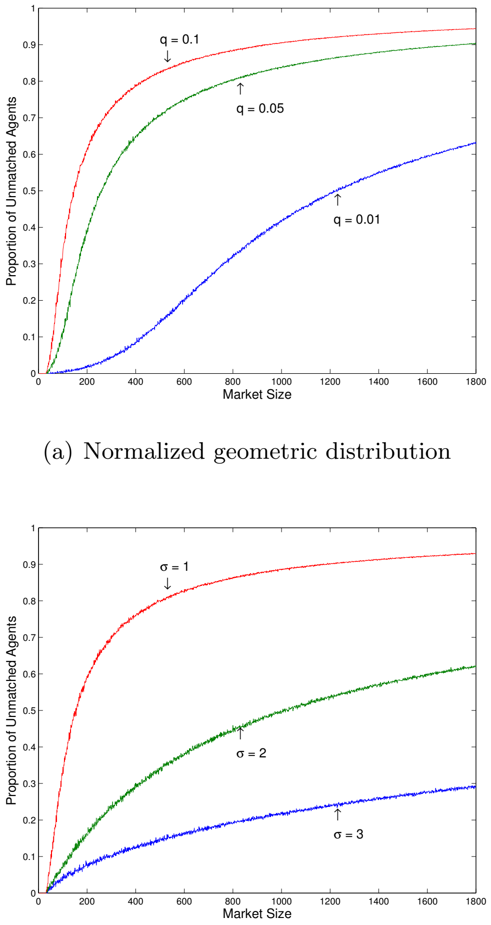
2. Normalized log-normal distribution

   Let $F(\,\cdot\,; \mu, \sigma)$ be the cumulative distribution function of a log-normal distribution. For each market size n, we define the normalized log-normal distribution as:

   $$\text{PDF} : p_k = \frac{F(\,k\,; \mu, \sigma) - F(\,k-1\,; \mu, \sigma)}{F(\,n\,; \mu, \sigma)}, \quad (\mu, \sigma \in \mathbb{R},\ k = 1, 2, \ldots, n).$$

   For each μ, as σ increases, firms have similar probabilities to be chosen. This generates a weaker commonality of preferences among workers.

particular preferences. Accordingly, we can measure the general likely proportions of unmatched agents.



(a) Normalized geometric distribution

(b) Normalized log-normal distribution

Figure A.1: Proportions of unmatched agents in stable matchings.

Figure A.1 shows that the proportion of unmatched agents in stable matchings increases as a market becomes large. Each graph represents the proportion of unmatched agents, when workers consider 30 most preferred firms acceptable. The proportions are averaged over 10 repetitions.

A.6 An Extended Model

We extend the model to allow that (i) the number of firms may differ from the number workers and (ii) some workers (or firms) may not be acceptable to some firms (or workers). Accordingly, firms (or workers) may remain unmatched in a stable matching. Moreover, we allow that (iii) private values for each pair of a firm and a worker are possibly correlated. Theorem 1 in the main paper holds in this extended model as well.

Let F be the set of n firms and W be the set of m workers. Utilities are represented by $n \times m$ random matrices $U = [U_{f,w}]$ and $V = [V_{f,w}]$. When a firm f and a worker w match with one another, the firm f receives utility $U_{f,w}$ and the worker w receives utility $V_{f,w}$. For each pair (f, w), utilities are defined as

$$U_{f,w} = \lambda\, U^o_w + (1-\lambda)\, \zeta_{f,w} \qquad \text{and}$$
$$V_{f,w} = \lambda\, V^o_f + (1-\lambda)\, \eta_{f,w} \qquad (0 < \lambda \le 1).$$

We call U^o_w and V^o_f *common-values*, and $\zeta_{f,w}$ and $\eta_{f,w}$ *private-values.*[13]

Common-values are defined as random vectors

$$U^o := \langle U^o_w \rangle_{w \in W} \quad \text{and} \quad V^o := \langle V^o_f \rangle_{f \in F}.$$

$\langle U^o_w \rangle_{w \in W}$ is an i.i.d sample of size m from a distribution with a positive density function on a bounded support in $\mathbb{R}$. $\langle V^o_f \rangle_{f \in F}$ is defined similarly.

Independent private-values are defined as two $n \times m$ random matrices

$$\zeta := [\zeta_{f,w}] \quad \text{and} \quad \eta := [\eta_{f,w}].$$

Each pair $(\zeta_{f,w}, \eta_{f,w})$ is randomly drawn from a joint distribution on a bounded support in $\mathbb{R}^2$. We normalize utilities such that firms and workers remaining unmatched receive 0 utility.

A random market is defined as a tuple $\langle F, W, U, V \rangle$. We denote realized matrices of U and V by u and v. A market instance is then denoted by $\langle F, W, u, v \rangle$. With probability 1, the market has all distinct utilities, none of which equals to 0. As such, for each $\langle F, W, u, v \rangle$, we can derive a strict preference list $\succ_f$ as

$$\succ_f = w, w', \ldots, f, \ldots, w''$$

[13] Note that we exclude the pure private value case ($\lambda = 0$).

if and only if

$$u_{f,w} > u_{f,w'} > \cdots > 0 > \cdots > u_{f,w''}.$$

Take any $\alpha \in (0, \infty)$, and consider a sequence m_n such that $\frac{m_n}{n}$ converges to α. We study properties of stable matchings in the sequence of random markets $\langle F_n, W_{m_n}, U_{n\times m_n}, V_{n\times m_n}\rangle_{n=1}^{\infty}$. We often omit the indexes n and m_n, or simply write n and m.

Given a market instance $\langle F, W, u, v\rangle$ and a matching μ, we let $u_\mu(\cdot)$ and $v_\mu(\cdot)$ denote utilities from the matching: i.e. $u_\mu(f) := u_{f,\mu(f)}$ and $v_\mu(w) := v_{\mu(w),w}$. For each $f \in F$, we define $\Delta(f; u, v)$ as the difference between utilities from firm-optimal and worker-optimal stable matchings: i.e.

$$\Delta(f; u, v) := u_{\mu_F}(f) - u_{\mu_W}(f).$$

For every $\epsilon > 0$, we have the set of firms whose utilities are within ϵ of one another for all stable matchings, which is denoted by

$$A^F(\epsilon; u, v) := \{f \in F \mid \Delta(f; u, v) < \epsilon\}.$$

Theorem A.6.1. *For every $\epsilon > 0$,*

$$E\left[\frac{|A^F(\epsilon; U, V)|}{n}\right] \to 1, \quad \textit{as} \quad n \to \infty.$$

We have similar notations and a theorem for workers, which are omitted here.

The intuition of Theorem A.6.1 is from the fact that the set of unmatched firms and workers is the same for all stable matchings (McVitie and Wilson (1970)). Firms and workers who remain unmatched have no difference in utilities from all stable matchings. Firms and workers who are matched in stable matchings have small differences in utilities by Theorem 1 in the main paper.

A.7 Proof of Theorem A.6.1

We prove the theorem when $0 < \lambda < 1$. If $\lambda = 1$, assortative matching forms a unique stable matching, and Theorem 1 follows immediately.

We first simplify the notations by compressing λ and $1-\lambda$ and considering utilities defined as

$$U_{f,w} = U^o_w + \zeta_{f,w} \quad \text{and} \quad V_{f,w} = V^o_f + \eta_{f,w}.$$

We do not lose generality since we can regard common-values and private-values as the ones already multiplied by λ and $1-\lambda$, respectively.

Let $U^o := \langle U^o_w \rangle_{w \in W}$ be an i.i.d sample of size m from a distribution G^W, and $V^o := \langle V^o_f \rangle_{f \in F}$ be an i.i.d sample of size n from a distribution G^F. G^W and G^F have strictly positive density functions on $\mathbb{R}$. Each pair $(\zeta_{f,w}, \eta_{f,w})$ is randomly drawn from a joint distribution Γ with a support bounded above by $(\bar{u}, \bar{v})$.

We define

$$B^F(\epsilon; u, v) := F \backslash A^F(\epsilon; u, v) = \{ f \in F \mid \Delta(f; u, v) \geq \epsilon \}$$

and prove that $\frac{|B^F(\epsilon; U, V)|}{n}$ converges to 0 in probability, which is equivalent to proving convergence to 0 in the mean (Theorem A.1.2). That is, we fix $\epsilon > 0$ and $K \in \mathbb{N}$, and prove that

$$P\left(\frac{|B^F(\epsilon; U, V)|}{n} > \frac{14}{K} \right) \to 0, \quad \text{as} \quad n \to \infty.$$

A.7.1 Preliminary Notations

1. ξ^F_q (or ξ^W_q) : q^{th} quantile of G^F (or G^W).

2. $\hat{\xi}^F_{q;n}$: empirical q^{th} quantile of a sample of size n from G^F. We also use $\hat{\xi}^F_{q;n}$ to denote its realization.

3. $\hat{\xi}^W_{q;m}$: empirical q^{th} quantile of a sample of size m from G^W. We also use $\hat{\xi}^W_{q;m}$ to denote its realization.

Since common-values are all distinct with probability 1, we index firms and workers in the order of their common-values: i.e.

$$v^o_{f_i} > v^o_{f_j} \quad \text{and} \quad u^o_{w_i} > u^o_{w_j}, \quad \text{if} \quad i < j.$$

Then, $U^o_{w_i;m}$ (or $V^o_{f_i;n}$) represents the i^{th} highest value of m (or n) order statistics from G^W

(or G^F). Note that $U^o_{w_i;m} = \hat{\xi}^W_{(1-\frac{i-1}{m});m}$ and $V^o_{f_i;n} = \hat{\xi}^F_{(1-\frac{i-1}{n});n}$ by the relationship between order statistics and empirical quantiles (see Appendix A.1).

Some firms may remain unmatched in stable matchings due to unequal populations of firms and workers, or because some firms (or workers) are not acceptable to some workers (or firms). Especially if a firm has a common value less than $\bar{u}$, all workers consider the firm unacceptable. Roughly, $G^W(-\bar{u})$ is the proportion of workers who are not acceptable to any firm, and $G^F(-\bar{v})$ is the proportion of firms which are not acceptable to any worker. Accordingly, we denote an asymptotic upper bound of the proportion of firms matched in stable matchings by

$$\beta := \min\{\alpha(1 - G^W(-\bar{u})), \quad 1 - G^F(-\bar{v})\}.$$

A.7.2 Tier-Grouping

We partition $\mathbb{R}$ into

$$\begin{aligned}
&I^W_1 := (\xi^W_{1-\frac{1}{\alpha K}}, \infty) \\
&I^W_2 := (\xi^W_{1-\frac{2}{\alpha K}}, \xi^W_{1-\frac{1}{\alpha K}}] \\
&\cdots \\
&I^W_k := (\xi^W_{1-\frac{k}{\alpha K}}, \xi^W_{1-\frac{k-1}{\alpha K}}] \\
&\cdots \\
&I^W_{K'} := (\xi^W_{1-\frac{K'}{\alpha K}}, \xi^W_{1-\frac{K'-1}{\alpha K}}] \\
&I^W_{K'+1} := (-\infty, \xi^W_{1-\frac{K'}{\alpha K}}],
\end{aligned}$$

where $K' = \lceil \beta K \rceil$.[14]

For each $\langle F, W, u, v \rangle$, we define **the set of workers in tier-k** (with respect to *workers'*

[14] K' is the smallest integer which is greater than or equal to βK. If $1 - \frac{K'}{\alpha K} \leq 0$, we let $\xi^W_{1-\frac{K'}{\alpha K}}$ equals to the infimum of the support of G^W.

common-values) as

$$W_k(u) := \left\{ w \mid u_w^o \in I_k^W \right\} \quad \text{for} \quad k = 1, 2, \dots, K' + 1$$

and define **the set of firms in tier-k** (with respect to *workers'* common-values) as

$$\begin{aligned} &F_k(u) := \{ f_i \in F \mid w_i \in W_k(u) \} \quad \text{for} \quad k = 1, 2, \dots, K', \quad \text{and} \\ &F_{K'+1}(u) := F \backslash \bigcup_{k=1}^{K'} F_k(u). \end{aligned}$$

Note that $F_{K'+1}(u)$ may include firms with indexes larger than the number of workers.

Similarly, we partition $\mathbb{R}$ into

$$\begin{aligned} &I_1^F := (\xi^F_{1-\frac{1}{K}}, \infty) \\ &I_2^F := (\xi^F_{1-\frac{2}{K}}, \xi^F_{1-\frac{1}{K}}] \\ &\dots \\ &I_k^F := (\xi^F_{1-\frac{k}{K}}, \xi^F_{1-\frac{k-1}{K}}] \\ &\dots \\ &I_{K'}^F := (\xi^F_{1-\frac{K'}{K}}, \xi^F_{1-\frac{K'-1}{K}}] \\ &I_{K'+1}^F := (-\infty, \xi^F_{1-\frac{K'}{K}}]. \end{aligned}$$

where $K' = \lceil \beta K \rceil$.

We define **the set of firms in tier-k** (with respect to *firms'* common-values) as

$$F_k(v) := \left\{ f \mid v_f^o \in I_k^F \right\} \quad \text{for} \quad k = 1, 2, \dots, K', K' + 1$$

and define **the set of workers in tier-k** (with respect to *firms'* common-values) as

$$\begin{aligned} &W_k(v) := \{ w_i \in W \mid f_i \in F_k(v) \} \quad \text{for} \quad k = 1, 2, \dots, K', \quad \text{and} \\ &W_{K'+1}(v) := W \backslash \bigcup_{k=1}^{K'} W_k(v). \end{aligned}$$

Note that $W_{K'+1}(v)$ may include workers with indexes larger than the number of firms.

We use the following notations.

1. $u_k^o := \xi^W_{1-\frac{k}{\alpha K}}$ for $k = 1, 2, \ldots, K'$: The threshold level of tier-k and tier-$(k+1)$ workers' common-values. That is, $w \in W_k(u)$ if and only if $u_k^o < u_w^o \leq u_{k-1}^o$.

2. $v_k^o := \xi^F_{1-\frac{k}{K}}$ for $k = 1, 2, \ldots, K'$: The threshold level of tier-k and tier-$(k+1)$ firms' common-values. That is, $f \in F_k(v)$ if and only if $v_k^o < v_f^o \leq v_{k-1}^o$.

Remark 4. *1. The set of tier-k workers (with respect to workers' common-values) is defined with a random sample. Therefore, $W_k(U)$ is random, and so is $F_k(U)$; whereas, u_k^o is a constant. Similarly, $F_k(V)$ and $W_k(V)$ are random; whereas, v_k^o is a constant.*

2. Tiers with respect to workers' common-values are in general not the same as tiers with respect to firms' common-values. In particular, we are most likely to have $|F_k(U)| \neq |F_k(V)|$.

Throughout the proof, we mainly use tiers defined with respect to workers' common-values. However, we need both tier structures in the last part of the proof. We simply write "tier-k" to denote tier-k with respect to workers' common-values, and use "(w.r.t firm) tier-k" to denote tier-k with respect to firms' common-values.

A.7.3 High-Probability Events

We introduce three events and show that the events occur with probabilities converging to 1 as the market becomes large. We provide proofs for completeness, but the main ideas are simply from the (weak) law of large numbers. In the next section, we will leave the probability that the following events do not occur as a remainder term converging to zero, and focus on the cases where the following events all occur.

A.7.3.1 No vanishing tier and an equal number of firms and workers in each tier.

Event 4 ($\mathcal{E}_1$)**.** *1. For $k = 1, 2, \ldots, K'$, the sets $F_k(U)$, $W_k(U)$, $F_k(V)$, and $W_k(V)$ are all non empty.*

2. For $k = 1, 2, \dots, K' - 1$,

$$|F_k(U)| = |W_k(U)| \quad \textit{and} \quad |F_k(V)| = |W_k(V)|.$$

Proof. The second part immediately follows from the first part. For instance, $F_{K'}(U) \neq \emptyset$ implies that the total number of firms is larger than the number of workers in tier up to $K' - 1$. By definition of tiers with respect to workers' common-values, we have $|F_k(U)| = |W_k(U)|$ for all $k = 1, 2, \dots, K' - 1$.

We only prove that $F_{K'}(U)$ and $W_{K'}(U)$ are non empty with probability converging to one as the market becomes large. Proofs for $k = 1, 2, \dots, K' - 1$ are almost analogous, and we omit here.

Note that

$$1 - \frac{K' - 1}{\alpha K} > 1 - \frac{\beta K}{\alpha K} \geq 0,$$

which implies that for each $w \in W$,

$$P\left(u_w^o \in I_{K'}^W\right) = G^W\left(\xi^W_{1 - \frac{K'-1}{\alpha K}}\right) - G^W\left(\xi^W_{1 - \frac{K'}{\alpha K}}\right) > 0.$$

As such, $W_{K'}(U) = \emptyset$ occurs with probability converging to 0 as the market becomes large.

When $W_{K'}(U)$ is not empty, $F_{K'}(U)$ remains empty only if the total number of firms is no more than the number of workers in tiers up to $K' - 1$. That is,

$$1 \leq \frac{1}{n} \sum_{k=1}^{K'-1} |W_k(U)|. \tag{A.20}$$

Note that

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^{K'-1} |W_k(U)| &= \frac{m}{n} \cdot \frac{1}{m} \sum_{k=1}^{m} \mathbf{1}\{U_w^o \geq u_{K'-1}^o\} \\ &\xrightarrow{p} \alpha \cdot \frac{K' - 1}{\alpha K} = \frac{\lceil \beta K \rceil}{K} - \frac{1}{K} \leq 1 - \frac{1}{K}. \end{aligned}$$

The convergence in probability is by the (weak) law of large numbers and Theorem A.1.3. Therefore, the inequality (A.20) holds with probability converging to zero, and thus $F_{K'}(U)$

is not empty with probability converging to 1. □

A.7.3.2 Distinct common-values of the agents in non-adjacent tiers.

Let $\tilde{\epsilon} > 0$ be such that for any $v, v' \in \mathbb{R}$,

$$|v - v'| \leq \tilde{\epsilon} \implies |G^F(v) - G^F(v')| < \frac{1}{3K},$$

and for any $u, u' \in \mathbb{R}$,

$$|u - u'| \leq \tilde{\epsilon} \implies |G^W(u) - G^W(u')| < \frac{1}{3\alpha K}.$$

There exists such an $\tilde{\epsilon}$ since G^F and G^W are continuous on their bounded supports, so uniformly continuous.

Event 5 ($\mathcal{E}_2$)**.** *For every $k = 1, 2, \dots, K' - 2$,*

$$\min_{\substack{f \in F_k(U) \\ f' \in F_{k+2}(U)}} |V^o_f - V^o_{f'}| > \tilde{\epsilon} \quad \textit{and} \quad \min_{\substack{w \in W_k(V) \\ w' \in W_{k+2}(V)}} |U^o_w - U^o_{w'}| > \tilde{\epsilon}.$$

Proof. We prove only the first part. Fix a realized matrix u such that $\mathcal{E}_1$ holds. For any $k \in 1, 2, \dots, K' - 2$ and for any $w_i \in W_k(u)$ and $w_j \in W_{k+2}(u)$,

$$u^o_{w_i} > u^o_k = \xi^W_{1-\frac{k}{\alpha K}} \quad \text{and} \quad u^o_{w_j} \leq u^o_{k+1} = \xi^W_{1-\frac{k+1}{\alpha K}}. \tag{A.21}$$

For any $q \in (0,1)$, $\hat{\xi}^W_{q;m} \xrightarrow{p} \xi^W_q$ (Theorem A.1.4), from which the following inequalities hold with probability converging to 1.

$$\xi^W_{1-\frac{k}{\alpha K}} > \hat{\xi}^W_{1-\frac{k}{\alpha K}-\frac{1}{8\alpha K}} \quad \text{and} \quad \xi^W_{1-\frac{k+1}{\alpha K}} < \hat{\xi}^W_{1-\frac{k+1}{\alpha K}+\frac{1}{8\alpha K}}. \tag{A.22}$$

Considering (A.21) and the relation between order statistics and empirical quantiles (see Appendix A.1), if (A.22) holds, we have

$$1 - \frac{k}{\alpha K} - \frac{1}{8\alpha K} < \min_{w_i \in W_k(u)} \left(1 - \frac{i-1}{m}\right) = \min_{f_i \in F_k(u)} \left(1 - \frac{i-1}{m}\right),$$

which implies that

$$1-\frac{k}{K}-\frac{1}{8K}<\min_{f_i\in F_k(u)}\left(1-\frac{\alpha(i-1)}{m}\right)<\min_{f_i\in F_k(u)}\left(1-\frac{i-1}{n}+\frac{1}{8K}\right)\quad\text{with large } n.$$

In addition, we have

$$1-\frac{k+1}{K}+\frac{1}{8K}>\max_{w_j\in W_{k+2}(u)}\left(1-\frac{\alpha(j-1)}{m}\right)=\max_{f_j\in F_{k+2}(u)}\left(1-\frac{\alpha(j-1)}{m}\right),$$

which implies that

$$1-\frac{k+1}{K}+\frac{1}{4K}>\max_{f_j\in F_{k+2}(u)}\left(1-\frac{j-1}{n}\right)\quad\text{with large } n.$$

As such for every $f_i\in F_k(u)$ and $f_j\in F_{k+2}(u)$,

$$v^o_{f_i}>\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{4K}}\quad\text{and}\quad v^o_{f_j}<\hat{\xi}^F_{1-\frac{k+1}{K}+\frac{1}{4K}}.$$

Therefore,

$$\begin{aligned}P\Big(\inf_{\substack{f_i\in F_k(U)\\ f_j\in F_{k+2}(U)}}\left|V^o_{f_i}-V^o_{f_j}\right|\le\tilde{\epsilon}\Big) &\le P\left(\left|\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{4K}}-\hat{\xi}^F_{1-\frac{k+1}{K}+\frac{1}{4K}}\right|\le\tilde{\epsilon}\right)+R_n\\ &\le P\left(\left|G^F(\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{4K}})-G^F(\hat{\xi}^F_{1-\frac{k+1}{K}+\frac{1}{4K}})\right|<\frac{1}{3K}\right)+R_n \qquad (A.23)\end{aligned}$$

where R_n corresponds to the probability that either $\mathcal{E}_1$ does not hold or (A.22) is violated: i.e. $R_n\to 0$. The last inequality is by the definition of $\tilde{\epsilon}$.

Note that

$$G^F(\hat{\xi}^F_{1-\frac{k}{K}-\frac{1}{4K}})-G^F(\hat{\xi}^F_{1-\frac{k+1}{K}+\frac{1}{4K}})\xrightarrow{p}\frac{1}{2K}$$

by Theorem A.1.4 and continuity of G^F(Theorem A.1.3). As a result, the right hand side of (A.23) converges to 0. □

A.7.3.3 Similarity between tiers w.r.t workers' common-values and tiers w.r.t firms' common-values

Event 6 ($\mathcal{E}_3$)**.** *For every* $k = 1, 2, 3, \ldots, K' + 1$,

$$F_k(U) \subset \bigcup_{k'=k-1}^{k+1} F_{k'}(V) \quad \textit{and} \quad W_k(V) \subset \bigcup_{k'=k-1}^{k+1} W_{k'}(U).^{15}$$

Proof. We prove the first part for $k = 1, \ldots, K'$ under the condition that $\mathcal{E}_1$ holds.[16]

For each realized (u, v), we have

$$\{u^o_w | w \in W_k(u)\} \subset \left(u^o_k, u^o_{k-1}\right] = \left(\xi^W_{1-\frac{k}{\alpha K}}, \xi^W_{1-\frac{k-1}{\alpha K}}\right]. \tag{A.24}$$

Suppose

$$\left(\xi^W_{1-\frac{k}{\alpha K}}, \xi^W_{1-\frac{k-1}{\alpha K}}\right] \subset \left(\hat{\xi}^W_{1-\frac{k}{\alpha K}-\frac{1}{3\alpha K}}, \hat{\xi}^W_{1-\frac{k-1}{\alpha K}+\frac{1}{3\alpha K}}\right], \tag{A.25}$$

and

$$\left(\hat{\xi}^F_{1-\frac{k}{K}-\frac{2}{3K}}, \hat{\xi}^F_{1-\frac{k-1}{K}+\frac{2}{3K}}\right] \subset \left(\xi^F_{1-\frac{k+1}{K}}, \xi^F_{1-\frac{k-2}{K}}\right]. \tag{A.26}$$

If (A.25) hold, then (A.24) implies that for every tier-k worker w_i, we have

$$u^o_{w_i} \in \left(\hat{\xi}^W_{1-\frac{k}{\alpha K}-\frac{1}{3\alpha K}}, \hat{\xi}^W_{1-\frac{k-1}{\alpha K}+\frac{1}{3\alpha K}}\right],$$

and thus,

$$1 - \frac{i-1}{m} \in \left(1 - \frac{k}{\alpha K} - \frac{1}{3\alpha K}, 1 - \frac{k-1}{\alpha K} + \frac{1}{3\alpha K}\right],$$

which implies that

$$1 - \frac{i-1}{n} \in \left(1 - \frac{k}{K} - \frac{2}{3K}, 1 - \frac{k-1}{K} + \frac{2}{3K}\right] \quad \text{with large } n.$$

[15] We define $F_0(V)$, $W_0(V)$, $W_{K'+2}(U)$, and $W_{K'+2}(U)$ as empty sets.

[16] For $k = 1, 2$, we need to modify the proof by replacing the intervals such as $(\xi^W_{1-\frac{k}{\alpha K}}, \xi^W_{1-\frac{k-1}{\alpha K}}]$ with $(\xi^W_{1-\frac{k}{\alpha K}}, \infty)$ and $(\xi^F_{1-\frac{k+1}{K}}, \xi^F_{1-\frac{k-2}{K}}]$ with $(\xi^F_{1-\frac{k+1}{K}}, \infty)$. We omit the modifications since they are trivial and tedious.

Then for any tier-k firm f_i, we have

$$v^o_{f_i} \in \left(\hat{\xi}^F_{1-\frac{k}{K}-\frac{2}{3K}}, \hat{\xi}^F_{1-\frac{k-1}{K}+\frac{2}{3K}}\right],$$

which implies that

$$\left\{v^o_f \mid f \in F_k(u)\right\} \subset \left(\hat{\xi}^F_{1-\frac{k}{K}-\frac{2}{3K}}, \hat{\xi}^F_{1-\frac{k-1}{K}+\frac{2}{3K}}\right].$$

Consequently if both (A.25) and (A.26) hold, then

$$\begin{aligned}\left\{v^o_f \mid f \in F_k(u)\right\} &\subset \left(\hat{\xi}^F_{1-\frac{k}{K}-\frac{2}{3K}}, \hat{\xi}^F_{1-\frac{k-1}{K}+\frac{2}{3K}}\right] \\ &\subset \left(\xi^F_{1-\frac{k+1}{K}}, \xi^F_{1-\frac{k-2}{K}}\right] \\ &= \bigcup_{k'=k-1}^{k+1} I^F_{k'}.\end{aligned}$$

In other words,

$$F_k(u) \subset \bigcup_{k'=k-1}^{k+1} F_{k'}(v).$$

Inequalities (A.25) and (A.26), and $\mathcal{E}_1$ occur with probability converging to 1 (Theorem A.1.4), and thus the event $\mathcal{E}_3$ for $k = 1, 2, \dots, K'$ also occurs with probability converging to 1.

Lastly for $k = K' + 1$,

$$F_{K'+1}(U) \subset F_{K'}(V) \cup F_{K'+1}(V)$$

occurs with probability converging to 1, since the event occurs whenever

$$F_k(V) \subset \bigcup_{k'=k-1}^{k+1} F_{k'}(U) \quad \text{for all } k = 1, 2, \dots, K' - 1$$

holds.

□

A.7.4 Proof of Theorem 1

We choose K large enough that

$$\max_{1\leq k\leq K'-2}\left|u_k^o - u_{k+1}^o\right| \equiv \max_{1\leq k\leq K'-2}\left|\xi^W_{1-\frac{k}{\alpha K}} - \xi^W_{1-\frac{k+1}{\alpha K}}\right| < \frac{\epsilon}{9}.^{17} \tag{A.27}$$

The proof of Theorem 1 is completed by the following inequalities.

$$\begin{aligned} P\left(\frac{|B^F(\epsilon;U,V)|}{n} > \frac{14}{K}\right) &= P\left(\sum_{1\leq k\leq K'+1}\frac{|B^F_k(\epsilon;U,V)|}{n} > \frac{14}{K}\right)\\ &< P\left(\sum_{7\leq k\leq K'-3}\frac{|B^F_k(\epsilon;U,V)|}{n} + \sum_{\substack{k=1,\dots,6,\\ K'-2,K'-1,K'}}\frac{F_k(U)}{n} + \frac{|B^F_{K'+1}(\epsilon;U,V)|}{n} > \frac{14}{K}\right). \end{aligned}$$

We show that the last term converges to 0. We first prove that for each $k = 7,\dots,K'-3$, the proportion $\frac{|B^F_k(\epsilon;U,V)|}{n}$ converges to 0 in probability (Proposition A.7.2). The proof identifies asymptotic upper and lower bounds of utilities from all stable matchings and shows that the two bounds are close to each other. We then prove that $\frac{|B^F_{K'+1}(\epsilon;U,V)|}{n}$ is asymptotically bounded above by $\frac{4}{K}$ (Proposition A.7.3). The proof shows that most tier-$K'+1$ firms remain unmatched in stable matchings, and thus have no difference in utilities. Lastly, for each $k = 1,\dots,6, K'-2, K'-1, K'$, the proportion $\frac{F_k(U)}{n}$ converges to at most $\frac{1}{K}$ in probability by the (weak) law of large numbers.

A.7.4.1 For $k = 7,\dots,K'-3$, $\frac{|B^F_k(\epsilon;U,V)|}{n} \xrightarrow{p} 0$.

We first identify an asymptotic lower bound on utilities of firms in each tier, using techniques from the theory of random bipartite graphs (Proposition A.7.1). Similarly, we find an asymptotic lower bound on utilities of workers in each tier (Proposition A.7.1*). The asymptotic lower bound on utilities of workers induces an asymptotic upper bound on utilities of firms in each tier. Lastly, we complete the proof by showing that the asymptotic lower and upper bounds are close to each other (Proposition A.7.2).

[17] We can always satisfy the condition since G^W has a strictly positive density function on a bounded support.

Proposition A.7.1. *For each instance $\langle F, W, u, v\rangle$ and for each $\bar{k} = 1, 2, \ldots, K'-3$, define*

$$\hat{B}^F_{\bar{k}}(\epsilon; u, v) := \left\{ f \in F_{\bar{k}}(u) : u_{\mu_W}(f) \leq u^o_{\bar{k}+2} + \bar{u} - \epsilon \right\}.$$[18]

Then for any $\epsilon > 0$,

$$\frac{|\hat{B}^F_{\bar{k}}(\epsilon; U, V)|}{n} \xrightarrow{p} 0 \quad as \quad n \to \infty.$$

Proof. For each instance $\langle F, W, u, v\rangle$ and for each $k = 1, 2, \ldots, K'+1$, let $F_{\leq k}(u) := \bigcup_{k' \leq k} F_{k'}(u)$ and $F_{<k}(u) := \bigcup_{k'<k} F_{k'}(u)$. We similarly define $W_{\leq k}(u)$ and $W_{<k}(u)$.

Take any $\bar{k}$ from $\{1, 2, \ldots, K'-3\}$. We construct a bipartite graph with $F_{\bar{k}}(u) \cup W_{\leq \bar{k}+2}(u)$ as a bi-partitioned set of nodes. Two vertices $f \in F_{\bar{k}}(u)$ and $w \in W_{\leq \bar{k}+2}(u)$ are joined by an edge if and only if

$$\zeta_{f,w} \leq \bar{u} - \epsilon \quad \text{or} \quad \eta_{f,w} \leq \bar{v} - \tilde{\epsilon},$$

where $\tilde{\epsilon}$ is the value taken before, while defining $\mathcal{E}_2$.

Let $\bar{W}_{\leq \bar{k}+2}(u, v)$ be the set of workers in tiers up to $\bar{k}+2$ who are *not* matched with firms in tiers up to $\bar{k}+1$ in μ_W. That is,

$$\bar{W}_{\leq \bar{k}+2}(u, v) := \left\{ w \in W_{\leq \bar{k}+2}(u) \,|\, \mu_W(w) \notin F_{\leq \bar{k}+1}(u) \right\}.$$

We now show that if $\mathcal{E}_2$ holds, then

$$\hat{B}^F_{\bar{k}}(\epsilon; u, v) \cup \bar{W}_{\leq \bar{k}+2}(u, v)$$

is a biclique.

Suppose, towards a contradiction, that a pair of $f \in \hat{B}^F_{\bar{k}}(\epsilon; u, v)$ and $w \in \bar{W}_{\leq \bar{k}+2}(u, v)$ is *not* joined by an edge: i.e.

$$\zeta_{f,w} > \bar{u} - \epsilon \quad \text{and} \quad \eta_{f,w} > \bar{v} - \tilde{\epsilon}.$$

[18] Note that $u^o_{\bar{k}+2} + \bar{u}$ is the maximal utility that a firm can achieve by being matched with a worker in tier-$(\bar{k}+3)$.

Then, we have

$$u_{f,w} = u^o_w + \zeta_{f,w} > u^o_{\bar{k}+2} + \zeta_{f,w} > u^o_{\bar{k}+2} + \bar{u} - \epsilon, \tag{A.28}$$

and

$$v_{f,w} = v^o_f + \eta_{f,w} \geq \min_{f' \in F_{\bar{k}}(u)} v^o_{f'} + \eta_{f,w} > \min_{f' \in F_{\bar{k}}(u)} v^o_{f'} + \bar{v} - \tilde{\epsilon}.\text{[19]}$$

Conditioned on $\mathcal{E}_2$, we can proceed further and obtain

$$\begin{aligned} v_{f,w} \quad &> \quad \min_{f' \in F_{\bar{k}}(u)} v^o_{f'} + \bar{v} - \left(\min_{f' \in F_{\bar{k}}(u)} v^o_{f'} - \max_{f'' \in F_{\bar{k}+2}(u)} v^o_{f''} \right) \\ &= \quad \max_{f'' \in F_{\bar{k}+2}(u)} v^o_{f''} + \bar{v}. \end{aligned} \tag{A.29}$$

On the other hand, $f \in \hat{B}^F_{\bar{k}}(\epsilon; u, v)$ implies that

$$u_{\mu_W}(f) \leq u^o_{\bar{k}+2} + \bar{u} - \epsilon,$$

and $w \in \bar{W}_{\leq \bar{k}+2}(u, v)$ implies that

$$v_{\mu_W}(w) \leq \max_{f'' \in F_{\bar{k}+2}(u)} v^o_{f''} + \bar{v},$$

since a worker can obtain utility higher than $\max_{f'' \in F_{\bar{k}+2}(u)} v^o_{f''} + \bar{v}$ only by matching with a firm in $F_{\leq \bar{k}+1}(u)$.

Equations (A.28) and (A.29) imply that (f, w) would have blocked μ_W, contradicting that μ_W is stable. Therefore,

$$\hat{B}^F_{\bar{k}}(\epsilon; u, v) \cup \bar{W}_{\leq \bar{k}+2}(u, v).$$

is a biclique, which is not necessarily balanced.

We now control the size of $\hat{B}^F_{\bar{k}}(\epsilon; U, V)$ by referencing Theorem 3. Let u^o and v^o be realized common-values such that events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. Then, the remaining randomness

[19] We should not replace $\min_{f' \in F_{\bar{k}}(u)} v^o_{f'}$ with $v^o_{\bar{k}}$. $F_{\bar{k}}(u)$ is defined with respect to workers' common-values, rather than firms' common-values.

of U and V is from ζ and η. Consider a random bipartite graph with $F_{\bar{k}}(U) \cup W_{\leq \bar{k}+2}(U)$ as a bi-partitioned set of nodes, where each pair of $f \in F_{\bar{k}}(U)$ and $w \in W_{\leq \bar{k}+2}(U)$ is joined by an edge if and only if

$$\zeta_{f,w} \leq \bar{u} - \epsilon \quad \text{or} \quad \eta_{f,w} \leq \bar{v} - \tilde{\epsilon}.$$

In other words, every pair is joined by an edge independently with probability $p(\epsilon) = 1 - \Gamma(\bar{u} - \epsilon, \bar{v} - \tilde{\epsilon})$.

We write $\beta(n) := 2 \cdot \log(|W_{\leq \bar{k}+2}(U)|)/\log \frac{1}{p(\epsilon)}$ and show that

$$P\left(|\hat{B}^F_{\bar{k}}(\epsilon; U, V)| \leq \beta(n)\right) \to 1 \quad \text{as} \quad n \to \infty. \tag{A.30}$$

Consider that

$$P\left(|\hat{B}^F_{\bar{k}}(\epsilon; U, V)| \leq \beta(n)\right) \geq P\left(\min\{|\hat{B}^F_{\bar{k}}(\epsilon; U, V)|, |\bar{W}_{\leq \bar{k}+2}(U, V)|\} \leq \beta(n)\right) - P\left(|\bar{W}_{\leq \bar{k}+2}(U, V)| \leq \beta(n)\right).$$

We show that the two terms on the right hand side converge respectively to 1 and 0 in probability.

Let $\alpha(U, V) \times \alpha(U, V)$ be the size of a maximum balance biclique of the random graph

$$G\left(F_{\bar{k}}(U) \cup W_{\leq \bar{k}+2}(U)\ ,\ p(\epsilon)\right).$$

Since every realized $\hat{B}^F_{\bar{k}}(\epsilon; u, v) \cup \bar{W}_{\leq \bar{k}+2}(u, v)$ is a biclique, it contains a balanced biclique of the size equals to

$$\min\left\{|\hat{B}^F_{\bar{k}}(\epsilon; u, v)|\ ,\ |\bar{W}_{\leq \bar{k}+2}(u, v)|\right\}.$$

Therefore,

$$P\left(\min\left\{|\hat{B}^F_{\bar{k}}(\epsilon; U, V)|, |\bar{W}_{\leq \bar{k}+2}(U, V)|\right\} \leq \beta(n)\right) \geq P\left(\alpha(U, V) \leq \beta(n)\right) \to 1, \tag{A.31}$$

where the convergence is from Theorem 3.

On the other hand, observe that $\bar{W}_{\leq \bar{k}+2}(U, V)$ is the size of at least $|W_{\bar{k}+2}(U)|$. Among workers in tiers up to $\bar{k} + 2$ at most $|W_{\leq \bar{k}+1}(U)|$ are matched with firms in tiers up to $\bar{k} + 1$.

In addition, $\frac{|W_{\bar{k}+2}(U)|}{n}$ converges to $\frac{1}{K}$ by the (weak) law of large numbers. Therefore,

$$P\left(|\bar{W}_{\leq\bar{k}+2}(U,V)| \leq \beta(n)\right) \to 0. \tag{A.32}$$

Equations (A.31) and (A.32) imply that (A.30) holds.

Lastly, we consider random utilities U and V, in which common-values are yet realized. For every $\epsilon' > 0$,

$$\begin{aligned} P\left(\frac{|\hat{B}^F_{\bar{k}}(\epsilon;U,V)|}{n} > \epsilon'\right) &= P\left(|\hat{B}^F_{\bar{k}}(\epsilon;U,V)| > \epsilon' \cdot n\right) \\ &\leq P\left(|\hat{B}^F_{\bar{k}}(\epsilon;U,V)| > \beta(n) \mid \mathcal{E}_1, \mathcal{E}_2, \beta(n) \leq \epsilon' n\right) + R_n, \quad \text{with large } n, \end{aligned}$$

where R_n is the probability that either $\mathcal{E}_1$ or $\mathcal{E}_2$ does not hold, or $\beta(n) \leq \epsilon' n$ is violated: i.e. $R_n \to 0$. We complete the proof by applying (A.30). □

We also obtain the counterpart proposition of Proposition A.7.1 in terms of tiers defined with respect to firms' common-values.

Proposition A.7.1* *For each $\bar{k} = 1, 2, \ldots, K'-3$, define*

$$\hat{B}^W_{\bar{k}}(\epsilon;u,v) := \left\{w \in W_{\bar{k}}(v) | v_{\mu_F}(w) \leq v^o_{\bar{k}+2} + \bar{v} - \epsilon\right\}.$$

Then for any $\epsilon > 0$,

$$\frac{|\hat{B}^W_{\bar{k}}(\epsilon;U,V)|}{n} \xrightarrow{p} 0 \quad as \quad n \to \infty.$$

Proof. We omit the proof since it is analogous to the proof of Proposition A.7.1. □

For each instance $\langle F, W, u, v\rangle$, we define

$$B^F_{\bar{k}}(\epsilon;u,v) := \{f \in F_{\bar{k}}(u) | \Delta(f;u,v) \geq \epsilon\} \quad \text{for} \quad \bar{k} = 1, 2, \ldots, K'+1.$$

Proposition A.7.2. *If $\bar{k} = 7, 8, \ldots, K' - 3$, then for any $\epsilon > 0$,*

$$\frac{|B_{\bar{k}}^F(\epsilon; U, V)|}{n} \xrightarrow{p} 0 \quad \textit{as} \quad n \to \infty.$$

Proof. In Proposition A.7.1*, for $k = 1, 2, \ldots, K' - 3$, let

$$\epsilon_k := v_{k+2}^o - v_{k+3}^o,$$

and write

$$\hat{B}_k^W(\epsilon_k; u, v) = \left\{ w \in W_k(v) | v_{\mu_F}(w) \leq v_{k+3}^o + \bar{v} \right\}.^{20}$$

By Proposition A.7.1*,

$$\frac{|\hat{B}_k^W(\epsilon_k; U, V)|}{n} \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty. \tag{A.33}$$

Note that a worker receives utility higher than $v_{k+3}^o + \bar{v}$ only by matching with a firm in (w.r.t firm) tiers up to $k + 3$.[21] Thus for $k = 5, 6, \ldots, K' + 1$,

$$\{w \in W_{\leq k-4}(V) : \mu(w) \in F_k(V)\} \subset \bigcup_{k'=1}^{k-4} \hat{B}_{k'}^W(\epsilon_{k'}; U, V). \tag{A.34}$$

If event $\mathcal{E}_3$ holds, we can translate (A.34) into an expression with tiers w.r.t workers' common-values. That is, for $k = 7, 8, \ldots, K' + 1$,

$$\begin{aligned} \{w \in W_{\leq k-6}(U) : \mu_F(w) \in F_k(U)\} &\subset \bigcup_{k'=k-1}^{k+1} \{w \in W_{\leq k-6}(U) : \mu_F(w) \in F_{k'}(V)\} \\ &\subset \bigcup_{k'=k-1}^{k+1} \{w \in W_{\leq k-5}(V) : \mu_F(w) \in F_{k'}(V)\} \\ &\subset \bigcup_{k'=k-1}^{k+1} \{w \in W_{\leq k'-4}(V) : \mu_F(w) \in F_{k'}(V)\} \end{aligned}$$

where the first and second inequalities are from $\mathcal{E}_3$.

[20] Recall that v_k^o is a constant, defined as $v_k^o := \xi_{1-\frac{k}{K}}^F$.

[21] Recall that $f \in F_k(v)$ if and only if $v_k^o < v_f^o \leq v_{k-1}^o$. Thus, if $f \in F_{>k+3}(v)$ then $v_f^o \leq v_{k+3}^o$.

By applying (A.34), we obtain

$$\{w \in W_{\leq k-6}(U) : \mu_F(w) \in F_k(U)\} \subset \bigcup_{k'=1}^{k-3} \hat{B}^W_{k'}(\epsilon_{k'}; U, V).$$

It follows that

$$\frac{|\{f \in F_k(U) : \mu_F(f) \in W_{\leq k-6}(U)\}|}{n} \xrightarrow{p} 0, \tag{A.35}$$

because for every $\epsilon > 0$,

$$P\left(\frac{|\{f \in F_k(U) : \mu_F(f) \in W_{\leq k-6}(U)\}|}{n} > \epsilon\right) \leq P\left(\sum_{k'=1}^{k-3} \frac{|\hat{B}^W_{k'}(\epsilon_{k'}; U, V)|}{n} > \epsilon\right) + R_n,$$

where R_n is the probability that $\mathcal{E}_3$ does not hold: i.e. $R_n \to 0$. The right hand side converges to 0 by (A.33).

We complete the proof of Proposition A.7.2 by proving the following claim. Proposition A.7.1 and (A.35) show that the normalized sizes of two sets on the right hand side of (A.36) converge to 0 in probability.

Claim A.7.1. *For $\bar{k} = 7, 8, \ldots, K' - 3$ and each instance $\langle F, W, u, v\rangle$,*

$$B^F_{\bar{k}}(\epsilon; u, v) \subset \hat{B}^F_{\bar{k}}(\epsilon/9; u, v) \cup \left\{f \in F_{\bar{k}}(u) | \mu_F(f) \in W_{\leq \bar{k}-6}(u)\right\}. \tag{A.36}$$

Proof of Claim A.7.1. If a firm $f \in F_{\bar{k}}(u)$ is *not* in $\hat{B}^F_{\bar{k}}(\epsilon/9; u, v)$, then

$$u_{\mu_W}(f) > u^o_{\bar{k}+2} + \bar{u} - \epsilon/9,$$

and if the firm f is *not* in $\left\{f \in F_{\bar{k}}(u) | \mu_F(f) \in W_{\leq \bar{k}-6}(u)\right\}$, then

$$u_{\mu_F}(f) \leq u^o_{\bar{k}-6} + \bar{u}.$$

Therefore, using (A.27) we obtain

$$u_{\mu_F}(f) - u_{\mu_W}(f) \leq u^o_{\bar{k}-6} - u^o_{\bar{k}+2} + \epsilon/9 < \epsilon,$$

and thus f is *not* in $B^F_k(\epsilon; u, v)$. □

□

A.7.4.2 Firms in tier $K' + 1$

We show that most firms in tier-$(K'+1)$ remain unmatched in stable matchings. Unmatched firms' utilities from μ_F and μ_W are clearly less than ϵ difference from each other.

Proposition A.7.3.

$$P\left(\frac{|B^F_{K'+1}(\epsilon; U, V)|}{n} > \frac{4}{K}\right) \to 0 \quad as \quad n \to \infty.$$

Proof. We divide the proof into two cases.

Case 1. $\beta = 1 - G^F(-\bar{v})$: only a small proportion of firms in tier $K' + 1$ are acceptable to workers.

For each $\langle F, W, u, v\rangle$, if $\mathcal{E}_3$ holds,

$$F_{K'+1}(u) \subset F_{K'}(v) \cup F_{K'+1}(v).$$

If $f \in F_{K'+1}(v)$,

$$v^o_f \le \xi^F_{1-\frac{K'}{K}} = \xi^F_{1-\frac{\lceil \beta K \rceil}{K}} \le \xi^F_{1-\beta} = -\bar{v}.^{22}$$

That is, if there is a firm in tier-$K' + 1$, the firm is unacceptable to all workers regardless of the firm's private values to the workers. The firm remains unmatched in all stable matchings and have no difference in utilities from stable matchings. Therefore, conditioned on $\mathcal{E}_3$,

$$\frac{|B^F_{K'+1}(\epsilon; U, V)|}{n} \le \frac{|F_{K'}(V)|}{n}.$$

Proposition A.7.3 holds from the following convergence result.

$$\frac{|F_{K'}(V)|}{n} \xrightarrow{p} \frac{\beta K - (\lceil \beta K \rceil - 1)}{K} \quad \text{as} \quad n \to \infty.$$

[22] Note that $f \in F_{K'+1}(v)$ implies $1 - \frac{K'}{K} > 0$ and $G^F(-\bar{v}) > 0$, which we used to derive the inequalities.

Case 2. $\beta = \alpha(1-G^W(-\bar{u}))$: firms in tier-$(K'+1)$ see only a small proportion of acceptable workers available.

For each market $\langle F, W, u, v\rangle$, if $w \in W_{K'+1}(u)$,

$$u^o_w \le \xi^W_{1-\frac{\lceil \beta K \rceil}{\alpha K}} \le \xi^W_{1-\frac{\beta}{\alpha}} = -\bar{u}.^{23}$$

That is, workers in $W_{K'+1}(u)$ are unacceptable to all firms. Therefore, the total number of matched workers in stable matchings is no more than the total number of workers in tiers up to K': i.e.

$$|\{w \in W | \mu_W(w) \in F\}| \le \sum_{k=1}^{K'} |W_k(U)|,$$

which implies that

$$|\{f \in F | \mu_W(f) \in W\}| \le \sum_{k=1}^{K'} |W_k(U)|.$$

As such, we have

$$\begin{aligned}|\{f \in F_{K'+1}(U) | \mu_W(f) \in W\}| &= |\{f \in F | \mu_W(f) \in W\}| - \sum_{k=1}^{K'} |\{f \in F_k(U) | \mu_W(f) \in W\}| \\ &\le \sum_{k=1}^{K'} |W_k(U)| - \sum_{k=1}^{K'} |\{f \in F_k(U) | \mu_W(f) \in W\}|.\end{aligned}$$

Conditioned on $\mathcal{E}_1$,

$$\begin{aligned}|\{f \in F_{K'+1}(U) | \mu_W(f) \in W\}| &\le \sum_{k=1}^{K'-3} (|F_k(U)| - |\{f \in F_k(U) | \mu_W(f) \in W\}|) + \sum_{k=K'-2}^{K'} |W_k(U)| \\ &= \sum_{k=1}^{K'-3} |\{f \in F_k(U) | \mu_W(f) \notin W\}| + \sum_{k=K'-2}^{K'} |W_k(U)|.\end{aligned}$$

[23] Note that $w \in W_{K'+1}(u)$ implies $1 - \frac{K'}{K} > 0$ and $G^W(-\bar{u}) > 0$, which we used to derive the inequalities.

With a small $\epsilon' > 0$,

$$\begin{aligned}\frac{|B^F_{K'+1}(\epsilon; U, V)|}{n} &\leq \frac{|\{f \in F_{K'+1}(U) | \mu_W(f) \in W\}|}{n} \\ &\leq \sum_{k=1}^{K'-3} \frac{|\hat{B}^F_k(\epsilon'; U, V)|}{n} + \sum_{k=K'-2}^{K'} \frac{|W_k(U)|}{n} \\ &\xrightarrow{p} 0 + \frac{3 + (\beta K - \lceil \beta K \rceil)}{K},\end{aligned}$$

where the convergence in probability is from Proposition A.7.1 and the (weak) law of large numbers.

□

Appendix B

Appendix to Chapter 2: Plea Bargaining

B.1 Existence of a Symmetric Voting Equilibrium.

Let $S := \{c, a\} \times \{c, a\}$ be the set of pure strategies; 'c' represents voting for conviction and 'a' for acquittal. A generic strategy $s \in S$ is a pair (s_g, s_i) consisting of voting decisions with signal g and i. Let $\Sigma := \Delta(\{c, a\}) \times \Delta(\{c, a\})$. A generic mixed strategy $\sigma = (\sigma_g, \sigma_i) \in \Sigma$ consists of probabilities of conviction voting with signal g and i. Define continuous functions $u_g(\sigma'_g, \sigma)$ or $u_i(\sigma'_i, \sigma)$ as a juror's expected utility when she receives signal g or i respectively and uses strategy σ', while all other jurors use strategy σ. Clearly, u_g and u_i are continuous in σ' and σ in our model.

We proceed similarly to the existence proof of Nash equilibrium in Nash (1951). For each pure strategy $s \in S$, define a continuous function h as

$$h^s(\sigma) = (h^s_1(\sigma),\, h^s_2(\sigma)) := \big(\max\{\, 0\, ,\, u_g(s_g, \sigma) - u_g(\sigma_g, \sigma)\}\ ,\ \max\{\, 0\, ,\, u_i(s_i, \sigma) - u_i(\sigma_i, \sigma)\}\big).$$

For each $s \in S$, define a continuous function as

$$y^s(\sigma) := \left(\frac{\sigma_{g:s_g} + h^s_1(\sigma)}{1 + \sum_{t \in \{c,a\}} h^t_1(\sigma)}\ ,\ \frac{\sigma_{g:s_i} + h^s_2(\sigma)}{1 + \sum_{t \in \{c,a\}} h^t_2(\sigma)} \right)$$

where $\sigma_{g:s_g}$ and $\sigma_{g:s_i}$ are the probabilities that the mixed strategy $\sigma = (\sigma_g, \sigma_i)$ assigns to each pure strategy s_g and s_i.

The set of functions $y^s(\cdot)$ for all $s \in S$ defines a mapping $y(\cdot)$ from the set of mixed strategy to itself. Similar to the existence proof of Nash equilibrium, a fixed point of $y(\cdot)$ is a symmetric Bayesian Nash Equilibrium (a symmetric equilibrium voting behavior). Since the set of mixed strategies is compact and convex, $y(\cdot)$ has a fixed point by the Brouwer fixed point theorem.

B.2 Proof of Proposition 2.3.1

For each level of belief π, we first find all symmetric equilibrium voting behaviors. Then we compare the jurors' expected payoffs and take the most efficient symmetric voting behavior.

B.2.1 Finding All Symmetric Equilibrium Voting Behaviors.

Non-responsive equilibrium voting behavior $(\sigma_g = 1, \sigma_i = 1)$ is an equilibrium voting behavior for any $1 \leq \hat{k} < n$. given that other jurors always vote for conviction, a juror is never pivotal. (Her vote never changes the judicial decisions.) In such a case, no juror has an incentive to change her voting strategy from $(\sigma_g = 1, \sigma_i = 1)$. Similarly, $(\sigma_g = 0, \sigma_i = 0)$ is an equilibrium voting behavior when $1 < \hat{k} \leq n$.

$(\sigma_g = 1, \sigma_i = 1)$ is not an equilibrium when $\hat{k} = n$. Given that other jurors always vote for conviction, being pivotal does not give any additional information. Each juror then fully relies on her own private signal. If a juror receives an innocent signal, then she votes for conviction (or acquittal) if and only if

$$\frac{1-p}{p}\,\frac{\pi}{1-\pi} \quad \geq (\text{or} \leq) \quad \frac{q}{1-q}.$$

Note that the evidence innately supports innocent defendants ($\frac{1-p}{p} < 1$ and $\frac{\pi}{1-\pi} \leq 1$), and reasonable doubt is in favor of acquittal ($\frac{q}{1-q} \geq 1$). A juror receiving an innocent signal does not have enough evidence to vote for conviction; $\sigma_i = 1$ is not a best response to $(\sigma_g = 1, \sigma_i = 1)$.

In a similar fashion, when $\hat{k} = 1$, $(\sigma_g = 0, \sigma_i = 0)$ is an equilibrium voting behavior only if $\pi \leq \bar{\pi}(1)$. Being pivotal does not provide any additional evidence, and a juror compares her private signal (g or i), belief (π), and reasonable doubt (q). If the belief π is low, even a guilty signal gives insufficient evidence for conviction voting.

Responsive equilibrium voting behavior A responsive voting behavior has $0 < \sigma_g$ and $\sigma_i < 1$; otherwise, $\sigma_g = \sigma_i$, and it is not responsive. We define r_G and r_I as conviction probabilities of guilty and innocent defendants, computed as

$$r_G = p\sigma_g + (1-p)\sigma_i, \quad r_I = (1-p)\sigma_g + p\sigma_i$$

When the jury follows responsive voting behavior, it does not always convict nor acquit defendants ($0 < r_G, r_I < 1$). In such a case, voting criteria (2.5) and (2.6), are well defined.

We consider each strategy case and find necessary levels of belief π consistent with the strategy as an equilibrium voting behavior. We explicitly compute the equilibria to use later for selecting the most efficient one.

Case 1 : $(0 < \sigma_g < 1,\ \sigma_i = 0)$

Conviction and acquittal must be indifferent to a juror receiving signal g. That is

$$\frac{r_G^{\hat{k}-1}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}-1}(1-r_I)^{n-\hat{k}}}\,\frac{p}{1-p}\,\frac{\pi}{1-\pi} = \frac{q}{1-q}.$$

Substituting in $r_G = p\,\sigma_g$ and $r_I = (1-p)\,\sigma_g$, we obtain

$$\left(\frac{1-p\sigma_g}{1-(1-p)\sigma_g}\right)^{n-\hat{k}}\left(\frac{p}{1-p}\right)^{\hat{k}}\frac{\pi}{1-\pi} = \frac{q}{1-q}. \tag{B.1}$$

Under the unanimity rule ($\hat{k} = n$), the first term in LHS is equal to 1, and the equality holds when $\pi = \bar{\pi}(\hat{k})$. Then, any $\sigma_g \in (0,1)$ with $\sigma_i = 0$ is an equilibrium voting behavior.

Consider a general super-majority rule $\hat{k}$ ($1 \leq \hat{k} < n$). Since $\frac{1-p\sigma_g}{1-(1-p)\sigma_g}$ is strictly decreasing in σ_g, by plugging $\sigma_g = 0$ and $\sigma_g = 1$ in (B.1), we can verify that $\bar{\pi}(\hat{k}) < \pi < \bar{\pi}(2\hat{k}-n)$ is necessary for $(0 < \sigma_g < 1, \sigma_i = 0)$ to be an equilibrium voting behavior. Moreover, at most one value of σ_g satisfies the equality. By algebraic manipulation of (B.1), we find $(\sigma_g, \sigma_i = 0)$ is an equilibrium voting strategy with

$$\sigma_g(\pi) = \frac{\psi_1 - 1}{(1-p)\psi_1 - p} \quad \text{where} \quad \psi_1 = \left(\frac{1-p}{p}\right)^{\frac{\hat{k}}{n-\hat{k}}} \left(\frac{q}{1-q}\,\frac{1-\pi}{\pi}\right)^{\frac{1}{n-\hat{k}}} \tag{B.2}$$

Case 2 : $(\sigma_g = 1,\, \sigma_i = 0)$

A juror receiving signal g prefers conviction, whereas a juror receiving signal i prefers acquittal. Substituting in $r_G = p$ and $r_I = 1 - p$ to voting criteria (2.5) and (2.6), we obtain

$$\left(\frac{p}{1-p}\right)^{2(\hat{k}-1)-n} \le \frac{q}{1-q}\,\frac{1-\pi}{\pi} \le \left(\frac{p}{1-p}\right)^{2\hat{k}-n} \tag{B.3}$$

The first inequality is from the criterion with signal i, and the second inequality is from the criterion with signal g. The above inequality is equivalent to $\bar{\pi}(2\hat{k}-n) \le \pi \le \bar{\pi}(2(\hat{k}-1)-n)$. When π is between $\bar{\pi}(2\hat{k}-n)$ and $\bar{\pi}(2(\hat{k}-1)-n)$, $(\sigma_g = 1, \sigma_i = 0)$ is an equilibrium voting behavior; every juror follows her own signal.

Case 3 : $(\sigma_g = 1,\, 0 < \sigma_i < 1)$

Jurors receiving signal i treat conviction and acquittal equally. That is

$$\frac{r_G^{\hat{k}-1}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}-1}(1-r_I)^{n-\hat{k}}}\,\frac{1-p}{p}\,\frac{\pi}{1-\pi} = \frac{q}{1-q}$$

Substituting in $r_G = p + (1-p)\sigma_i$ and $r_I = (1-p) + p\sigma_i$, we get

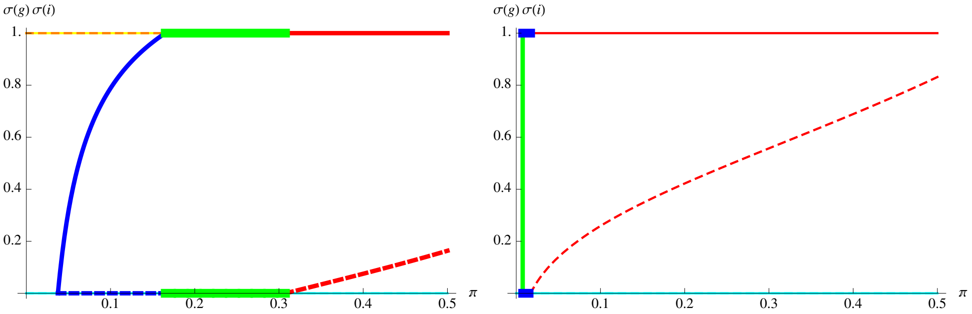
$$\left(\frac{p + (1-p)\sigma_i}{(1-p) + p\sigma_i}\right)^{\hat{k}-1} \left(\frac{1-p}{p}\right)^{n-\hat{k}+1} \frac{1-\pi}{\pi} = \frac{q}{1-q} \tag{B.4}$$

Note that $\frac{p+(1-p)\sigma_i}{(1-p)+p\sigma_i}$ is strictly decreasing in σ_i. By plugging in $\sigma_i = 0$ and $\sigma_i = 1$, we can verify that $\bar{\pi}(2(\hat{k}-1)-n) < \pi \le .5$ is necessary if $\sigma_g = 1$ and $0 < \sigma_i < 1$ is an equilibrium voting behavior.

For each level of belief π such that $\bar{\pi}(2(\hat{k}-1)-n) < \pi < .5$, at most one σ_i satisfies

General super-majority rules ($1 \leq \hat{k} < n$)		The unanimity rule ($\hat{k} = n$)	
Non-responsive voting			
$\forall\ \pi \in [0, .5]$	$(\sigma_g = \sigma_i = 1)$	$\forall\ \pi \in [0, .5]$	$(\sigma_g = \sigma_i = 0)$
$\pi \in [0, .5](\hat{k} > 1)$, $\pi \in [0, \bar{\pi}(1)](\hat{k} = 1)$	$(\sigma_g = \sigma_i = 0)$		
Responsive voting			
$\bar{\pi}(\hat{k}) < \pi < \bar{\pi}(2\hat{k} - n)$	$(0 < \sigma_g < 1, \sigma_i = 0)$	$\pi = \bar{\pi}(n)$	$(0 < \sigma_g < 1, \sigma_i = 0)$
$\bar{\pi}(2\hat{k} - n) \leq \pi \leq \bar{\pi}(2(\hat{k} - 1) - n)$	$(\sigma_g = 1, \sigma_i = 0)$	$\bar{\pi}(n) \leq \pi \leq \bar{\pi}(n - 2)$	$(\sigma_g = 1, \sigma_i = 0)$
$\bar{\pi}(2(\hat{k} - 1) - n) < \pi \leq .5$	$(\sigma_g = 1, 0 < \sigma_i < 1)$	$\bar{\pi}(2n - 2) < \pi \leq .5$	$(\sigma_g = 1, 0 < \sigma_i < 1)$

Table B.1: Symmetric voting equilibrium behavior in jury trial.



(a) A super-majority rule ($\hat{k} = 8$).

(b) The unanimity rule ($\hat{k} = 12$).

Figure B.1: Symmetric equilibrium voting behavior with $n = 12$, $p = \frac{6}{10}$, and $q = \frac{6}{10}$

the equality. This σ_i combined with $\sigma_g = 1$ forms a symmetric equilibrium voting behavior, and σ_i is determined as

$$\sigma_i(\pi) = \frac{p - \psi_2(1-p)}{p\,\psi_2 - (1-p)} \quad \text{where} \quad \psi_2 = \left(\frac{p}{1-p}\right)^{\frac{n-\hat{k}+1}{\hat{k}-1}} \left(\frac{q}{1-q}\,\frac{1-\pi}{\pi}\right)^{\frac{1}{\hat{k}-1}} \tag{B.5}$$

Table B.1 summarizes all symmetric equilibrium voting behavior. Figure B.1 illustrates equilibrium voting behaviors with $n = 12$, $p = \frac{6}{10}$, and $q = \frac{6}{10}$, when voting rules are $\hat{k} = 8$ and $\hat{k} = 12$. We used solid lines for σ_g and dashed lines for σ_i. For each π, the pair of σ_g and σ_i forming a strategy profile (σ_g, σ_i) share the same thickness. In this example, we observe all three equilibrium cases, but we may not observe some cases under other parameter values. For instance, $\bar{\pi}(2(\hat{k} - 1) - n)$, one of the threshold levels of belief, may not be defined or may be larger than .5. In such a case, $(\sigma_g = 1, \sigma_i = 0)$ is not an equilibrium voting behavior for any $\pi \in [0, .5]$.

B.2.2 Finding an Efficient Equilibrium Voting Behavior.

For each belief π, there may be several symmetric equilibrium voting behaviors. If a responsive equilibrium voting behavior exists, intuitively it must be more efficient than non-responsive equilibrium voting behavior, because jurors essentially *use* private signals to form judgements. We confirm this intuition by comparing responsive equilibrium voting outcomes with non-responsive equilibrium voting outcomes. If there is no responsive equilibrium voting behavior for a belief π, then one of the non-responsive equilibria, $(\sigma_g = 1, \sigma_i = 1)$ or $(\sigma_g = 0, \sigma_i = 0)$, is an efficient equilibrium voting behavior.

Given a belief π, conviction probabilities, (P_G, P_I), change the jurors' expected payoff by

$$-q \cdot (1-\pi) \cdot P_I - (1-q) \cdot \pi \cdot (1 - P_G).$$

The first term corresponds to mistakenly convicting innocent defendants, and the second term corresponds to mistakenly acquitting guilty defendants.

Between two non-responsive equilibrium voting behaviors, $(\sigma_g = \sigma_i = 0)$ and $(\sigma_g = \sigma_i = 1)$, the former gives a higher jurors' expected utility than the latter, because $q\,(1-\pi)$ is larger than $(1-q)\,\pi$.

When $\pi > \bar{\pi}(\hat{k})$, there is a responsive equilibrium voting behavior, and responsive voting is more efficient than $(\sigma_g = \sigma_i = 0)$ if and only if the conviction probabilities (P_G, P_I) of responsive voting satisfy

$$-q\ (1-\pi)\ P_I - (1-q)\ \pi\ (1-P_G) > -(1-q)\ \pi$$

which we can rewrite as

$$\frac{P_G}{P_I} = \frac{\sum_{j=\hat{k}}^{n} \binom{n}{j} r_G^j (1-r_G)^{n-j}}{\sum_{j=\hat{k}}^{n} \binom{n}{j} r_I^j (1-r_I)^{n-j}} > \frac{q}{1-q}\frac{1-\pi}{\pi}. \tag{B.6}$$

If the above inequality holds as an equality, then responsive voting behavior and $(\sigma_g = 0, \sigma_i = 0)$ are both equally efficient.

We proceed separately with general super-majority rules and the unanimity rule.

General super-majority rules ($\hat{k} < n$) In order to verify (B.6), first note that $k' > k$ and $r_G > r_I > 0$ implies

$$\frac{r_G^{k'}(1-r_G)^{n-k'}}{r_I^{k'}(1-r_I)^{n-k'}} > \frac{r_G^{k}(1-r_G)^{n-k}}{r_I^{k}(1-r_I)^{n-k}}. \tag{B.7}$$

Also note that

$$\text{if } x, x' > 0 \text{ and } y, y' > 0, \quad \frac{x'}{y'} > \frac{x}{y} \quad \text{implies} \quad \frac{x+x'}{y+y'} > \frac{x}{y}. \tag{B.8}$$

Sequentially applying (B.7) and using (B.8), we obtain

$$\frac{\sum_{k=\hat{k}}^{n}\binom{n}{k} r_G^{k}(1-r_G)^{n-k}}{\sum_{k=\hat{k}}^{n}\binom{n}{k} r_I^{k}(1-r_I)^{n-k}} > \frac{r_G^{\hat{k}}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}}(1-r_I)^{n-\hat{k}}}.$$

Therefore, to prove (B.6), it is enough to show

$$\frac{r_G^{\hat{k}}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}}(1-r_I)^{n-\hat{k}}} \geq \frac{q}{1-q}\,\frac{1-\pi}{\pi}. \tag{B.9}$$

We proceed with each case of responsive equilibrium voting behavior.

Case 1 : $(0 < \sigma_g < 1, \sigma_i = 0)$, where $\bar{\pi}(\hat{k}) < \pi < \bar{\pi}(2\hat{k} - n)$.

By substituting in $r_G = p\sigma_g$ and $r_I = (1-p)\sigma_g$, the LHS of (B.9) becomes

$$\frac{r_G^{\hat{k}}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}}(1-r_I)^{n-\hat{k}}} = \left(\frac{1-p\sigma_g}{1-(1-p)\sigma_g}\right)^{n-\hat{k}}\left(\frac{p}{1-p}\right)^{\hat{k}}.$$

The equilibrium restriction (B.1) implies that the RHS of the above expression is equal to the RHS of (B.9). Thus (B.9) holds under equality.

Case 2 : $(\sigma_g = 1, \sigma_i = 0)$, where $\bar{\pi}(2\hat{k} - n) \leq \pi \leq \bar{\pi}(2(\hat{k} - 1) - n)$.

Since $r_G = p$ and $r_I = 1 - p$, the LHS of (B.9) is

$$\frac{r_G^{\hat{k}}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}}(1-r_I)^{n-\hat{k}}} = \left(\frac{p}{1-p}\right)^{2\hat{k}-n}.$$

From (B.3), equation (B.9) must be true.

Case 3 : $(\sigma_g = 1, 0 < \sigma_i < 1)$, where $\bar{\pi}(2(\hat{k}-1)-n) < \pi \leq .5$.

Note that (B.4) is a necessary equilibrium restriction. Since $\pi \leq .5$ and $p > .5$,

$$\left(\frac{p+(1-p)\sigma_i}{(1-p)+p\sigma_i}\right)^{\hat{k}-1}\left(\frac{1-p}{p}\right)^{n-\hat{k}+1} = \frac{q}{1-q}\,\frac{1-\pi}{\pi}$$

By substituting in $r_G = p + (1-p)\,\sigma_i$, $r_I = (1-p) + p\,\sigma_i$, we obtain

$$\frac{r_G^{\hat{k}}(1-r_G)^{n-\hat{k}}}{r_I^{\hat{k}}(1-r_I)^{n-\hat{k}}} = \left(\frac{p+(1-p)\sigma_i}{(1-p)+p\sigma_i}\right)^{\hat{k}}\left(\frac{1-p}{p}\right)^{n-\hat{k}} \geq \left(\frac{p+(1-p)\sigma_i}{(1-p)+p\sigma_i}\right)^{\hat{k}-1}\left(\frac{1-p}{p}\right)^{n-\hat{k}+1}$$

Inequality (B.9) is derived from the above two inequalities.

The unanimity rule ($\hat{k} = n$) If the voting rule follows the unanimity rule, then (B.6) becomes

$$\frac{P_G}{P_I} = \left(\frac{r_G}{r_I}\right)^n > \frac{q}{1-q}\,\frac{1-\pi}{\pi}. \tag{B.10}$$

If the above inequality holds, responsive voting is more efficient than $(\sigma_g = 0, \sigma_i = 0)$; if LHS and RHS are equal, both responsive equilibrium voting and $(\sigma_g = 0, \sigma_i = 0)$ are equally efficient.

Case 1: $(0 < \sigma_g < 1, \sigma_i = 0)$, where $\pi = \bar{\pi}(n)$.

By substituting in $r_G = p\sigma_g$ and $r_I = (1-p)\sigma_g$, the LHS of (B.10) becomes

$$\left(\frac{r_G}{r_I}\right)^n = \left(\frac{p}{1-p}\right)^n.$$

By definition of $\bar{\pi}(\cdot)$ and $\pi = \bar{\pi}(n)$, (B.10) holds as an equality. Thus, both $(0 < \sigma_g < 1, \sigma_i = 0)$ and $(\sigma_g = 0, \sigma_i = 0)$ are equally efficient.

Case 2: $(\sigma_g = 1, \sigma_i = 0)$, where $\bar{\pi}(2\hat{k}-n) \leq \pi \leq \bar{\pi}(2(\hat{k}-1)-n)$.

Since $r_G = p$ and $r_I = 1-p$, the LHS of (B.10) is

$$\left(\frac{r_G}{r_I}\right)^n = \left(\frac{p}{1-p}\right)^n.$$

By definition of $\bar{\pi}(\cdot)$, (B.10) holds as an equality when $\pi = \bar{\pi}(2\hat{k}-n) = \bar{\pi}(n)$; otherwise if $\bar{\pi}(n) < \pi \leq \bar{\pi}(2(\hat{k}-1)-n)$ then (B.10) holds with a strict inequality. Thus, when $\pi = \bar{\pi}(n)$, both $(\sigma_g = 1, \sigma_i = 0)$ and $(\sigma_g = 0, \sigma_i = 0)$ are equally efficient; when $\bar{\pi}(n) < \pi \leq \bar{\pi}(2(\hat{k}-1)-n)$, responsive equilibrium voting $(\sigma_g = 1, \sigma_i = 0)$ is more efficient than $(\sigma_g = \sigma_i = 0)$.

Case 3: $(\sigma_g = 1, 0 < \sigma_i < 1)$, where $\bar{\pi}(2(\hat{k}-1)-n) < \pi \leq .5$.

By substituting in $r_G = p + (1-p)\,\sigma_i$, $r_I = (1-p) + p\,\sigma_i$, we obtain

$$\left(\frac{r_G}{r_I}\right)^n = \left(\frac{p+(1-p)\sigma_i}{(1-p)+p\sigma_i}\right)^n > \left(\frac{p+(1-p)\sigma_i}{(1-p)+p\sigma_i}\right)^{n-1}\frac{p}{1-p} = \frac{q}{1-q}\frac{1-\pi}{\pi}$$

where the last equality is from the voting criterion (B.4). Responsive equilibrium voting is the most efficient equilibrium voting behavior.

B.3 Other Notions of Equilibrium Refinements.

We use the most efficient equilibrium as an equilibrium refinement, but it is a theoretically interesting question whether other previously studied refinement concepts are also applicable. It turns out that equilibrium refinement using *trembling hand perfection* by Austen-Smith and Feddersen (2005) or *weakly un-dominated strategies* by Gerardi and Yariv (2007) does not generate equilibrium voting behavior satisfying natural properties in Proposition 2.3.2. We prove this by showing that, when the voting rule is a super-majority and π is small, both $\sigma_g = \sigma_i = 0$ and $\sigma_g = \sigma_i = 1$ are weakly undominated strategies, and none of them passes trembling hand perfection.

First, we show that both $\sigma_g = \sigma_i = 0$ and $\sigma_g = \sigma_i = 1$ are weakly undominated strategies. Assume that $1 \leq \hat{k} < n$ and $\pi = \bar{\pi}(\hat{k}) - \epsilon$. We showed in the proof of Proposition 2.3.1 that only $\sigma_g = \sigma_i = 1$ and $\sigma_g = \sigma_i = 0$ are symmetric equilibria. The level of belief is low enough that $\hat{k}$ number of guilty signals give a single dictating juror insufficient evidence to convict the defendant. However, with slightly more evidence, the juror will have enough incentive to convict the defendant.

We first consider $\sigma_g = \sigma_i = 0$. Suppose that all other jurors except juror j play (σ_g', σ_i')

in which $\sigma'_g = 1$ and $\frac{1}{2} < \sigma'_i < 1$. Being pivotal implies that $\hat{k}-1$ other jurors vote for conviction. Such an event combined with juror j's guilty signal provides less incentive to vote for conviction than the event that juror j herself observes $\hat{k}$ number of guilty signals, because some other jurors' conviction votes may come from i signals. The best response for juror j with signal g is to vote for acquittal. Clearly, the best response when the signal is i is also to vote for acquittal. Therefore, $\sigma_g = \sigma_i = 0$ is not a weakly dominated strategy.

We next consider $\sigma_g = \sigma_i = 1$. Suppose that all other jurors except juror j play (σ''_g, σ''_i) in which $0 < \sigma''_g < \frac{1}{2}$ and $\sigma''_i = 0$. Being pivotal implies that $\hat{k}-1$ other jurors vote for conviction. Such an event gives more incentive to vote for conviction than the event that juror j herself observes $\hat{k}$ number of guilty signals, because some other jurors' acquittal votes may come from g signals. The best response for juror j is to vote for conviction regardless of her own signal. Since $\sigma_g = \sigma_i = 1$ is the best response, it is not a weakly dominated strategy.

On the other hand, neither $\sigma_g = \sigma_i = 0$ nor $\sigma_g = \sigma_i = 1$ passes trembling hand perfection. Trembling hand perfection modified to our Bayesian game requires us to construct a sequence of perturbed games. In each perturbation, players assign strictly positive probabilities to both pure strategies: $(\sigma^n_g = \epsilon^n_1, \sigma^n_i = \epsilon^n_2)$ and $(\sigma^n_g = 1 - \epsilon^n_3, \sigma^n_i = \epsilon^n_4)$. Trembling hand perfection requires that the strategy must constitute a Bayesian Nash equilibrium of a corresponding sequence of perturbed games, and the sequence of equilibria must converge to the Bayesian Nash equilibrium of the original game, $(\sigma_g = \sigma_i = 0)$ and $(\sigma_g = \sigma_i = 1)$, respectively. However, since guilty signal g gives a strictly higher incentive to vote for conviction than a signal i, such a sequence of perturbed games does not exist. In no case is a juror indifferent between voting for conviction and voting for acquittal with both signals, g and i. Therefore, neither $\sigma_g = \sigma_i = 0$ nor $\sigma_g = \sigma_i = 1$ passes trembling hand perfection.

B.4 Proof of Proposition 2.3.2.

The conviction probabilities of guilty defendants and innocent defendants, $\{(P_G, P_I)|\pi\}$, are determined by

$$P_G = \sum_{k=\hat{k}}^{n} \binom{n}{k} r_G^k (1-r_G)^{n-k}$$
$$P_I = \sum_{k=\hat{k}}^{n} \binom{n}{k} r_I^k (1-r_I)^{n-k}$$

where $r_G = p\sigma_g + (1-p)\sigma_i$ and $r_I = (1-p)\sigma_g + p\sigma_i$, where (σ_g, σ_i) is the efficient equilibrium voting behavior.

When the efficient equilibrium voting behavior is $(\sigma_g = 0, \sigma_i = 0)$, $P_G \geq P_I$ clearly holds, because the conviction probabilities are all equal to zero. If the efficient equilibrium voting behavior is responsive, we showed that (B.6) holds and $\frac{q}{1-q}\frac{1-\pi}{\pi} \geq 1$. Thus, $P_G \geq P_I$ (Item 1).

From the closed form solutions of responsive equilibrium voting behavior, we observed that σ_g and σ_i are constant on $[0, \bar{\pi}(\hat{k})]$ and $[\bar{\pi}(2\hat{k}-n), \bar{\pi}(2(\hat{k}-1)-n)]$, and non-decreasing in π on both intervals $(\bar{\pi}(\hat{k}), \bar{\pi}(2\hat{k}-n))$ and $(\bar{\pi}(2\hat{k}-n), .5]$. By comparing across intervals, we can check that σ_g and σ_i are non-decreasing in π over $[0, .5]$. From the closed form solutions of efficient equilibrium voting behavior, it is also easy to see that σ_g and σ_i are increasing in $\hat{k}$ (Item 2).

Lastly, $f_G(\pi)$ and $f_I(\pi)$ are non-decreasing in π, because the conviction probabilities are strictly increasing in σ_g and σ_i, and σ_g and σ_i are non-decreasing in π (Item 3).

B.5 Proof of Proposition 2.4.1

We first prove the following lemma.[1]

Lemma B.5.1. *Conviction probability of guilty defendants $f_G(\pi)$ is an upper hemicontinuous correspondence in π with non-empty convex values.*

Proof. Note that the efficient equilibrium voting behavior σ_g and σ_i are unique for every π, except when $\pi = \bar{\pi}(n)$ and the rule is unanimous, in which efficient equilibrium voting behavior is any pair of $(\sigma_i = 0, 0 \leq \sigma_g \leq 1)$. Since $\sum_{k'=\hat{k}}^{n} \binom{n}{k'} r_G^{k'} (1-r_G)^{n-k'}$ is a continuous

[1] The lemma holds also for $f_I(\pi)$, but we do not need this observation in proving Proposition 2.4.1.

function of σ_g and σ_i, $f_G(\pi)$ is convex valued for all π (Intermediate Value Theorem). In addition, closed form solutions of efficient equilibrium voting behavior (σ_g and σ_i) are upper hemicontinuous in π. Since f_G is continuous in σ_g and σ_i, $f_G(\pi)$ inherits upper hemicontinuity in π. □

Now, suppose $\theta \leq P_G$. It is necessary that $\theta \in [0, \bar{\theta}]$ where $\bar{\theta} := \sup f_G(.5)$. There exists a π such that $\theta \in f_G(\pi)$, because $f_G(\pi)$ is upper hemicontinuous in π with non-empty convex values (Intermediate Value Theorem). Suppose by contradiction that $\theta < P_G$. Every guilty defendant pleads guilty, and only innocent defendants may or may not go to trial. In such a case, jurors reasonably believe that all defendants in trials are innocent ($\pi = 0$), which consequently leads conviction probability to equal zero. This contradicts $\theta < P_G$. $\theta = P_G$ must be true (Item 1).

Otherwise, we have $\theta > P_G$ as a part of an equilibrium outcome. No defendant pleads guilty, and the jurors' reasonable beliefs π will be equal to .5. The conviction probabilities (P_G, P_I) must be in $\{(P'_G, P'_I)|.5\}$ (Item 2).

B.6 Proof of Proposition 2.4.2

B.6.1 Simplifying the prosecutor's problem

The prosecutor's problem is described below.

$$\max_{\theta \in [0,1]} -\frac{1}{2}q'\Big(\phi_I\theta + (1-\phi_I)P_I\Big) - \frac{1}{2}(1-q')\Big(\phi_G(1-\theta) + (1-\phi_G)(1-P_G)\Big) \tag{B.11}$$

$$\text{such that} \quad \begin{array}{ll} (a.1) & \phi_G \in \arg\min_{\phi' \in [0,1]} \phi'\theta + (1-\phi')P_G \\ (a.2) & \phi_I \in \arg\min_{\phi' \in [0,1]} \phi'\theta + (1-\phi')P_I \\ (b) & \pi = \begin{cases} 0 & \text{if } \phi_G = \phi_I = 1 \\ \frac{1-\phi_G}{(1-\phi_G)+(1-\phi_I)} & \text{otherwise.} \end{cases} \\ (c) & (P_G, P_I) \in \{(P'_C, P'_I)|\pi\}. \end{array}$$

Using Proposition 2.4.1, we simplify the above expressions. To begin with, we can restrict without loss of generality that a prosecutor can offer $\theta \in [0, \bar{\theta}]$, because he can obtain any utility level from offering $\theta > \bar{\theta}$ by offering $\theta = \bar{\theta}$; all players perceive the same ex-ante punishments in both cases. In the former case (offering $\theta > \bar{\theta}$), all defendants plead not guilty and receive $(P_G, P_I) \in \{(P'_G, P'_I)|.5\}$ conviction probabilities. In the latter case, some guilty defendants may plead guilty, but the punishment for a guilty plea is equal to the conviction probability: i.e. the expected punishment from a jury trial. As far as the ex-ante punishments are the same, the prosecutor and the defendant are indifferent between pleading guilty and pleading not guilty.

Once the prosecutor offers $\theta \in [0, \bar{\theta}]$, Proposition 2.4.1 ensures that $\theta = P_G \geq P_I$. Pleading decisions of guilty defendants are straightforward; guilty defendants are indifferent toward pleading guilty or pleading not guilty, thus any $\phi_G \in [0, 1]$ is a best response. Pleading decisions of innocent defendants depend on whether $\theta = P_I$ or $\theta > p_I$. $P_G = P_I$ holds only when $\theta = P_G = P_I = 0$; otherwise, $\theta = P_G > P_I$. In the former case, any pleading decision behavior incurs the same expected prosecutor's utility, $-\frac{1}{2}(1 - q')$ including when $\phi_I = 1$ (no punishment). In the latter case, $\phi_I = 1$ must be true, since only pleading not guilty is the best response. In all, when the prosecutor offers $\theta \in [0, \bar{\theta}]$, it is innocuous for the prosecutor to assume that $\phi_I = 1$. By applying these observations, we simplify the prosecutor's decision as

$$\max_{\theta \in [0, \bar{\theta}]} -\frac{1}{2} q' P_I - \frac{1}{2}(1 - q')(1 - \theta)$$

$$\text{such that} \quad \begin{array}{ll} (a) & \phi_G \in [0, 1] \\ (b) & \pi = \begin{cases} 0 & \text{if } \phi_G = 1 \\ \frac{1-\phi_G}{2-\phi_G} & \text{otherwise.} \end{cases} \\ (c) & (\theta, P_I) \in \{(P'_G, P'_I)|\pi\}. \end{array}$$

It is convenient to define a function $\tilde{P_I} : [0, \bar{\theta}] \to [0, 1]$ as follows.

$$\tilde{P_I}(\theta) = p_I, \quad \text{where} \quad \exists\, \pi, \quad (\theta, p_I) \in \{(P'_G, P'_I)|\pi\}.$$

Referencing the proof of Proposition 2.3.1, we can verify that the function $\tilde{P_I}$ is well-defined; For every $\theta \in [0, \bar{\theta}]$, the value of $\tilde{P_I}(\theta)$ exists and is unique. There are four cases: (1) $\theta = 0$, (2) $\theta \in (0, \hat{p_G})$, (3) $\theta = \hat{p_G}$, or (4) $\theta \in (\hat{p_G}, \bar{\theta}]$, in which $\hat{p_G}$ is the conviction probability of guilty defendants when jurors vote by following their own signals ($\sigma_g = 1, \sigma_i = 0$).

If $\theta = 0$, p_I must be 0. If $\theta = \hat{p_G}$, p_I is unique and the value is derived from the voting strategy ($\sigma_g = 1, \sigma_i = 0$). For other cases, recall that the conviction probabilities are defined as

$$P_G = \sum_{k=\hat{k}}^{n} \binom{n}{k} r_G^k (1 - r_G)^{n-k}, \quad P_I = \sum_{k=\hat{k}}^{n} \binom{n}{k} r_I^k (1 - r_I)^{n-k}$$

where $r_G = p\sigma_g + (1-p)\sigma_i$ and $r_I = (1-p)\sigma_g + p\sigma_i$. When $\theta \in (0, \hat{p_G})$, $\sigma_i = 0$ and both P_G and P_I are strictly increasing in σ_g. Since P_G is continuous in r_G which is also continuous in σ_g, for any $\theta \in (0, \hat{p_G})$, there exists a unique σ_g inducing $P_G = \theta$. Such a σ_g combined with $\sigma_i = 0$ gives a unique p_I such that $(\theta, p_I) \in \{(P_G', P_I')|\pi\}$. A similar procedure applies when $\theta \in (\hat{p_G}, \bar{\theta}]$.

Through the above argument, the function $\tilde{P_I}$ is not only well-defined, but strictly increasing and continuous on $[0, \bar{\theta}]$, and differentiable on $(0, \hat{p_G})$ and $(\hat{p_G}, \bar{\theta})$. Using $\tilde{P_I}$, the prosecutor's problem becomes

$$\max_{\theta \in [0, \bar{\theta}]} U(\theta) := -\frac{1}{2} q' \tilde{P}_I(\theta) - \frac{1}{2}(1 - q')(1 - \theta). \tag{B.12}$$

We show that the objective function above is strictly concave. Thus, the First Order Condition (FOC) will be the necessary and sufficient condition of the maximizer θ^*. We later use the FOC to prove Proposition 2.4.2.

B.6.2 $U(\theta)$ is strictly concave in θ.

Since $\tilde{P_I}$ is continuous in θ, the objective function is, too. Moreover, $\tilde{P_I}$ is differentiable on $(0, \hat{p_G})$ and $(\hat{p_G}, \bar{\theta})$, and $U(\theta)$ is a linear combination of θ and $\tilde{P_I}$. Thus, $U(\theta)$ is also differentiable with respect to θ on $(0, \hat{p_G})$ and $(\hat{p_G}, \bar{\theta})$. If we show that the derivative of $\tilde{P_I}$ is decreasing on $(0, \hat{p_G})$ and $(\hat{p_G}, \bar{\theta})$, and the left derivate is greater than the right at $\hat{p_G}$, then the concavity of $\tilde{P_I}$ follows. Since $U(\theta)$ is a linear combination of θ and $\tilde{P_I}$, concavity of the objective function directly follows from the concavity of $\tilde{P_I}$.

When $\theta \in (0, \hat{p_G})$, P_G and P_I are differentiable with respect to σ_g. The derivative of P_G is

$$\begin{aligned}
\frac{\partial P_G}{\partial \sigma_g} &= \frac{\partial}{\partial \sigma_g} \sum_{k=\hat{k}}^{n} \binom{n}{k} (r_G)^k (1-r_G)^{n-k} \\
&= \sum_{k=\hat{k}}^{n-1} \Bigg(\frac{n!}{k!(n-k)!} k r_G^{k-1} (1-r_G)^{n-k} r'_G \\
&\quad - \frac{n!}{k!(n-k-1)!} r_G^{k} (n-k)(1-r_G)^{n-k-1} r'_G \Bigg) + n r_G^{n-1} r'_G \\
&= n\, r'_G \binom{n-1}{\hat{k}-1} r_G^{\hat{k}-1} (1-r_G)^{n-\hat{k}} \qquad \text{(B.13)}
\end{aligned}$$

Using a similar operation, we obtain

$$\frac{\partial P_I}{\partial \sigma_g} = n\, r'_I \binom{n-1}{\hat{k}-1} r_I^{\hat{k}-1} (1-r_I)^{n-\hat{k}} \qquad \text{(B.14)}$$

Therefore,

$$\frac{\partial \tilde{P_I}(\theta)}{\partial \theta} = \frac{\partial P_I / \partial \sigma_g}{\partial P_G / \partial \sigma_g} = \frac{r'_I\, r_I^{\hat{k}-1} (1-r_I)^{n-\hat{k}}}{r'_G\, r_G^{\hat{k}-1} (1-r_G)^{n-\hat{k}}}. \qquad \text{(B.15)}$$

Since $r_G = p\sigma_g$ and $r_I = (1-p)\sigma_g$, (B.15) becomes

$$\left(\frac{1-p}{p} \right)^{\hat{k}} \left(\frac{1-(1-p)\sigma_g}{1-p\sigma_g} \right)^{n-\hat{k}}. \qquad \text{(B.16)}$$

As θ increases in $(0, \hat{p_G})$, the corresponding σ_g increases, and the above derivative strictly decreases. Therefore, $\frac{\partial \tilde{P_I}(\theta)}{\partial \theta}$ is decreasing in $\theta \in (0, \hat{p_G})$.

When $\theta \in (\hat{p_G}, \bar{\theta})$, σ_g is fixed equal to 1 and only σ_i varies. Similar to (B.13) and (B.14), we obtain

$$\frac{\partial \tilde{P_I}(\theta)}{\partial \theta} = \frac{\partial P_I / \partial \sigma_i}{\partial P_G / \partial \sigma_i} = \frac{r'_I\, r_I^{\hat{k}-1} (1-r_I)^{n-\hat{k}}}{r'_G\, r_G^{\hat{k}-1} (1-r_G)^{n-\hat{k}}}. \qquad \text{(B.17)}$$

By substituting in $r_G = p + (1-p)\sigma_i$ and $r_I = (1-p) + p\sigma_i$, we obtain

$$\left(\frac{(1-p)+p\sigma_i}{p+(1-p)\sigma_i}\right)^{\hat{k}-1}\left(\frac{p}{1-p}\right)^{n-\hat{k}+1}. \tag{B.18}$$

Again, as θ increases in $(\hat{p_G}, \bar{\theta})$, the corresponding σ_i increases, and the above derivative decreases. Therefore, $\frac{\partial \tilde{P_I}(\theta)}{\partial \theta}$ is decreasing in $\theta \in (\hat{p_G}, \bar{\theta})$

Lastly, at $\theta = \hat{p_G}$, the left derivative is greater than the right derivative, because the limit of (B.16) as σ_g goes to 1 is greater than the limit of (B.18) as σ_i goes to 0. This concludes that $\tilde{P_I}$ is strictly concave in θ, and thus the objective function in (B.12) is also strictly concave in θ.

B.6.3 First Order Condition

Since the prosecutor's objective function is strictly concave in θ, the FOC gives the necessary and sufficient condition of optimizer θ^*. Instead of finding the closed form solution, we use the FOC and prove Proposition 2.4.2. We proceed for each case of the optimizer θ^*.

Interior Solutions

$(0 < \theta^* < \hat{p_G})$: Using (B.16), FOC of (B.12) becomes

$$\left(\frac{p}{1-p}\right)^{\hat{k}}\left(\frac{1-p\sigma_g}{1-(1-p)\sigma_g}\right)^{n-\hat{k}} = \frac{q'}{1-q'}.$$

Recall that a juror receiving a guilty signal uses a mixed strategy at this level of conviction probability for guilty defendants. (Equation (B.2) holds.) We obtain

$$\frac{q}{1-q}\frac{1-\pi}{\pi} = \frac{q'}{1-q'}$$

$(\hat{p_G} < \theta^* < \bar{\theta})$: Using (B.18), FOC of (B.12) becomes

$$\left(\frac{p+(1-p)\sigma_i}{(1-p)+p\sigma_i}\right)^{\hat{k}-1}\left(\frac{1-p}{p}\right)^{n-\hat{k}+1} = \frac{q'}{1-q'}.$$

Recall that a juror receiving an innocent signal uses a mixed strategy at this level of conviction probability for guilty defendants. (Equation (B.5) holds.) We obtain

$$\frac{q}{1-q}\frac{1-\pi}{\pi} = \frac{q'}{1-q'}$$

Boundary Solutions

$(\theta^* = \hat{p_G})$: The prosecutor offers this punishment for a guilty plea, when

$$\lim_{\theta \downarrow \hat{p_G}} \frac{\partial U(\theta)}{\partial \theta} \leq 0 \leq \lim_{\theta \uparrow \hat{p_G}} \frac{\partial U(\theta)}{\partial \theta}$$

Replacing (B.16) and (B.18) for $\frac{\partial \tilde{P_I}(\theta)}{\partial \theta}$, we can rewrite the above inequalities as

$$\left(\frac{(1-p)+p\sigma_i}{p+(1-p)\sigma_i}\right)^{\hat{k}-1}\left(\frac{p}{1-p}\right)^{n-\hat{k}+1} \leq \frac{1-q'}{q'} \leq \left(\frac{1-p}{p}\right)^{\hat{k}}\left(\frac{1-(1-p)\sigma_g}{1-p\sigma_g}\right)^{n-\hat{k}},$$

or

$$\left(\frac{p}{1-p}\right)^{2(\hat{k}-1)-n} \leq \frac{q'}{1-q'} \leq \left(\frac{p}{1-p}\right)^{2\hat{k}-n}$$

Compared with (B.3), when the prosecutor chooses $\theta^* = \hat{p_G}$, the jurors' voting behavior with π and q is exactly the same as the voting behavior when jurors' belief is equal to .5 and reasonable doubt is equal to q'.

$(\theta^* = 0)$: The right derivative at $\theta = 0$ must be less than or equal to 0. By applying (B.16) to the derivative of the objective function in (B.12) and taking $\sigma_g \to 0$, we obtain

$$\left(\frac{p}{1-p}\right)^{\hat{k}} \leq \frac{q'}{1-q'}.$$

Note that θ^* induces the equilibrium voting behavior $\sigma_g = \sigma_i = 0$. This strategy profile becomes an efficient equilibrium voting behavior when the RHS of (B.1) is greater than or equal to the LHS, which implies

$$\left(\frac{p}{1-p}\right)^{\hat{k}} \leq \frac{q}{1-q}\frac{1-\pi}{\pi}.$$

By comparing the above two inequalities, we observe that the equilibrium voting behavior is the same as the voting behavior when jurors' beliefs are equal to .5 and reasonable doubt is equal to q'.

$(\theta^* = \bar{\theta})$: The left derivative at $\theta = \bar{\theta}$ must be non-negative. Applying (B.18) to the derivative of $U(\theta)$, we must obtain

$$\lim_{\theta \uparrow \bar{\theta}} \frac{\partial U(\theta)}{\partial \theta} \geq 0$$

or

$$\left(\frac{p + (1-p)\bar{\sigma_i}}{(1-p) + p\bar{\sigma_i}} \right)^{\hat{k}-1} \left(\frac{1-p}{p} \right)^{n-\hat{k}+1} \geq \frac{q'}{1-q'}$$

where $\bar{\sigma_i}$ with $\sigma_g = 1$ is an equilibrium voting behavior with the belief $\pi = .5$.

Note that in this situation, a juror receiving an innocent signal is indifferent between conviction and acquittal. Thus (B.4) becomes

$$\left(\frac{p + (1-p)\bar{\sigma_i}}{(1-p) + p\bar{\sigma_i}} \right)^{\hat{k}-1} \left(\frac{1-p}{p} \right)^{n-\hat{k}+1} = \frac{q}{1-q}.$$

Thus, $\frac{q}{1-q} \geq \frac{q'}{1-q'}$, or $q \geq q'$.

When $q \geq q'$, the prosecutor offers $\theta^* = \bar{\theta}$, and all defendants plead not guilty ($\pi = .5$). Jurors vote with threshold $\frac{q}{1-q}$, which is the same as the threshold in the jury model without plea bargaining. Although we have restricted the prosecutor's strategy space to $[0, \bar{\theta}]$, any θ^* higher than $\bar{\theta}$ induces the same prosecutor's equilibrium expected utility as $\theta^* = \bar{\theta}$.

Proposition 2.4.2 summarizes these results of FOC.

B.6.4 Proof of Corollary 5

First, note that efficient equilibrium voting behavior is responsive if $\pi > \bar{\pi}(\hat{k})$. Since $\bar{\pi}(l)$ is strictly decreasing in l, the efficient equilibrium voting behaviors are responsive for all $\pi > 0$ as $n \to \infty$.

Given π, p, and a voting rule ($\hat{k} = n$), efficient equilibrium voting leads the conviction probabilities to converge to $1 - \left(\frac{(1-q)(1-p)\pi}{qp(1-\pi)}\right)^{\frac{1-p}{2p-1}}$ for guilty defendants, and to $\left(\frac{(1-q)(1-p)\pi}{qp(1-\pi)}\right)^{\frac{p}{2p-1}}$ for innocent defendants. These convergence results directly follow Proposition 2 in Feddersen and Pesendorfer (1998). (Our parameter values satisfy all conditions assumed in their Propositions.)

For general super-majority rules, regardless of the jury size n, we have $\frac{\pi}{1-\pi} = 1$ (if $q > q'$) or $\frac{1-q}{q}\frac{\pi}{1-\pi} = \frac{1-q'}{q'}$ (if $q \leq q'$). As we replace $\frac{1-q}{q}\frac{\pi}{1-\pi} = \frac{1-\tilde{q}}{\tilde{q}}$ where $\tilde{q} = \max\{q, q'\}$, the conviction probabilities for guilty defendants and innocent defendants directly follow Proposition 3 in Feddersen and Pesendorfer (1998); the conviction probability for guilty defendants converges to 1 and for innocent defendants converges to 0.

Lastly from Proposition 2.4.1 in this paper, we can relate the ex-ante punishments, one for guilty defendants and another for innocent defendants, to the conviction probabilities in jury trials.

Appendix C

Appendix to Chapter 3: Zero-sum Games

We prove the main theorem.

The necessity of (PC), (PE), and (INT) Suppose a joint choice correspondence f is Nash-rationalizable by a zero-sum game $(A, \succeq, \preceq)$. The necessity of (PC) and (PE) is obvious from the definition of Nash equilibrium. To show the necessity of (INT), let $B = B_1 \times B_2 \in \mathcal{A}$ and $b = (b_1, b_2), b' = (b'_1, b'_2) \in f(B)$. Note that $b_1, b'_1 \in B_1$ and $b_2, b'_2 \in B_2$, which implies that (b_1, b'_2) and (b'_1, b_2) are also in B.

Since (b_1, b_2) is a Nash equilibrium of the game $(B, \succeq, \preceq)$,

i) player 1 prefers (b_1, b_2) to (b'_1, b_2): i.e. $(b_1, b_2) \succeq (b'_1, b_2)$, and

ii) player 2 prefers (b_1, b_2) to (b_1, b'_2): i.e. $(b_1, b_2) \preceq (b_1, b'_2)$, or equivalently $(b_1, b'_2) \succeq (b_1, b_2)$.

In addition, since (b'_1, b'_2) is a Nash equilibrium of the game $(B, \succeq, \preceq)$,

iii) player 1 prefers (b'_1, b'_2) to (b_1, b'_2): i.e. $(b'_1, b'_2) \succeq (b_1, b'_2)$, and

iv) player 2 prefers (b'_1, b'_2) to (b'_1, b_2): i.e. $(b'_1, b'_2) \preceq (b'_1, b_2)$, or equivalently $(b'_1, b_2) \succeq (b'_1, b'_2)$.

By transitivity of $\succeq$, from (i) and (iv) we obtain $(b_1, b_2) \succeq (b'_1, b_2) \succeq (b'_1, b'_2)$, and from (ii) and (iii) we obtain $(b'_1, b'_2) \succeq (b_1, b'_2) \succeq (b_1, b_2)$. Therefore, (b_1, b_2), (b'_1, b_2), (b_1, b'_2), and (b'_1, b'_2) are all indifferent for player 1 and player 2.

In this situation, (b'_1, b_2) is a Nash equilibrium of the game $(B, \succeq, \preceq)$: for any $b''_1 \in B_1$, since (b_1, b_2) is a Nash equilibrium, we have $(b_1, b_2) \succeq (b''_1, b_2)$, and thus $(b'_1, b_2) \succeq (b''_1, b_2)$; from player 2's viewpoint, for any $b''_2 \in B_2$, since (b'_1, b'_2) is a Nash equilibrium, we have $(b'_1, b'_2) \preceq (b'_1, b''_2)$, and thus $(b'_1, b_2) \preceq (b'_1, b''_2)$. Similarly, (b_1, b'_2) is also a Nash equilibrium of the game $(B, \succeq, \preceq)$. In all, $\{b\} \vee \{b'\}$ is a subset of the set of Nash equilibria of the game $(B, \succeq, \preceq)$, and therefore a subset of $f(B)$.

The sufficiency of (PC), (PE), and (INT) To prove sufficiency, we construct a preference $\succeq$ over A, with which for all $B \in \mathcal{A}$, $f(B)$ coincides with the set of all Nash equilibria of $(B, \succeq, \preceq)$.

In individual choice theory, given a finite alternative set X and a choice correspondence g, Sen (1971) defines *base relation* R^* as

$$xR^*y \text{ if and only if } x \in g(\{x, y\}).$$

Similarly, we define two relations $\succeq^*$ and $\succeq^{**}$ as follows: for any $a = (a_1, a_2), b = (b_1, b_2) \in A$,

$$\begin{aligned} a \succeq^* b \quad & \text{if and only if} \quad a_2 = b_2 \text{ and } a \in f(\{a_1, b_1\} \times \{a_2\}), \\ a \succeq^{**} b \quad & \text{if and only if} \quad a_1 = b_1 \text{ and } b \in f(\{a_1\} \times \{a_2, b_2\}) \end{aligned}$$

Note that $\succeq^*$ and $\succeq^{**}$ are disjoint, and $\succeq^{**}$ is defined "inversely" from the convention of individual choice theory. Finally, let $\succeq$ be the union of $\succeq^*$ and $\succeq^{**}$. We arrange player 1's conceivable actions in a column and player 2's actions in a row, thereby constructing a table of joint actions. Then, in each line (PC) is equivalent to Sen's α and β, and $\succeq^*$ and $\succeq^{**}$ are defined as analogous with the base relation. $\succeq^*$ represents the base relation in each column, and $\succeq^{**}$ represents the base relation in each row, except $\succeq^{**}$ is defined inversely. In such case, Sen (1971) shows that $\succeq^*$ is a weak order in each column, and $\succeq^{**}$ is an inverse relation of a weak order in each row; therefore, the union $\succeq$ is a weak order in both columns and rows. Note that $\succeq$ is not yet defined on any pair of joint actions across the lines. In order to construct a complete relation over A, we need some preliminary definitions.

Definition 6 (Consistency)**.** *Let R be a relation over $X = \{x^1, x^2, \dots, x^l, \dots\}$ and P be the strict counterpart of R. A sequence $x^1Rx^2R\cdots Rx^lPx^1$ is called a PR-cycle (or a cycle). If a relation does not have any cycle, we say that it is **consistent**.*

Definition 7 (Extension)**.** *Given any arbitrary binary relation R over X, if a binary relation R' over X is such that*

$$xRy \text{ implies } xR^{'}y$$
$$xPy \text{ implies } xP^{'}y$$

*then $R^{'}$ is called an **extension** of R.*

In the following proof, we show using interchangeability that $\succeq$ is consistent (Section C.1). Then, we show using (PE) and (PC) that any weak order extension of $\succeq$ Nash-rationalizes the joint choice correspondence by a zero-sum game (Section C.2).

C.1 $\succeq$ is consistent.

By means of contradiction, suppose that there exists $\{a^1, \cdots, a^N\} \subset A$ such that $a^1 \succeq a^2 \succeq \cdots \succeq a^N \succ a^1$. Since $\succeq$ is the union of two disjoint sets, $\succeq^*$ and $\succeq^{**}$, $\succeq$ is either $\succeq^*$ or $\succeq^{**}$ depending on whether $\{a^i, a^j\}$ is in a column or a row.

Hereafter, we restrict our attention to cycles of an even length of at least 4 where the links in the cycle alternate between $\succeq^{**}$ and $\succeq^*$. This restriction does not lead to a loss of generality. First, we only need to consider cycles that alternate because any cycle containing consecutive $\succeq^*$ or $\succeq^{**}$ can be reduced by means of transitivity to a shorter cycle without consecutive $\succeq^*$ or $\succeq^{**}$. In addition, there is no cycle with a length of 2 such as $a^1 \succeq^* a^2 \succ^{**} a^1$. By definition of $\succeq^*$, $a^1_2 = a^2_2$, and by definition of $\succeq^{**}$, $a^1_1 = a^2_1$, which together imply that $a^1 = a^2$. Then, we have $a^1 \succ^{**} a^1$. We can also rule out cycles of odd lengths, since we can shorten any cycle of a odd length by transitivity to a cycle of an even length. For instance, the cycle $a \succeq^{**} b \succeq^* c \succeq^{**} d \succeq^* e \succ^{**} a$ of length 5 can be reduced to the cycle $b \succeq^* c \succeq^{**} d \succeq^* e \succ^{**} b$ of length 4. We also restrict attention to the case where the cycle begins with $\succeq^{**}$. The case where the cycle begins with $\succeq^{*}$ is omitted, but can be proved in a similar way.

First, we prove that there is no cycle of length 4. Suppose $a \succeq^{**} b \succeq^{*} c \succeq^{**} d \succ^{*} a$. By definition, we have $a_1 = b_1$, $b_2 = c_2$, $c_1 = d_1$ and $d_2 = a_2$. Then $\{a, b, c, d\}$ makes feasible sets as depicted in Figure C.1. In part (i) of the figure, each dashed arrow corresponds to either $\succeq^{*}$ or $\succeq^{**}$ and the solid arrow corresponds to $d \succ^{*} a$. The tail of each arrow is the element from the left hand side of the preference relation.

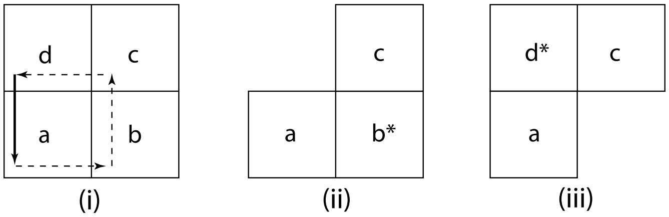


Figure C.1: A cycle of length 4

Parts (ii) and (iii) of Figure C.1 illustrate the choice correspondence generating $\succeq^{*}$ and $\succeq^{**}$ for each feasible set. Note that $b \in f(\{a, b\}) \cap f(\{b, c\})$, and $d \in f(\{a, d\}) \cap f(\{c, d\})$.[1] Then (PE) implies that $b \in f(\{a, b, c, d\})$ and $d \in f(\{a, b, c, d\})$. Since f is interchangeable, and since $a_1 = b_1$ and $a_2 = d_2$, $a = (b_1, d_2)$ must also be chosen; i.e., $a \in f(\{a, b, c, d\})$. Likewise, $c = (d_1, b_2)$ implies $c \in f(\{a, b, c, d\})$. Finally, (PC) implies that $a \in f(\{a, d\})$, which contradicts $d \succ^{*} a$. So, there cannot be any cycle with length 4.

Now, let us make the induction hypothesis that there is no cycle of length $2(n-1)$ where $n \geq 3$. Given this hypothesis, we prove that there is no cycle of length $2n$.

By reordering the list of individual actions for player 1 and for player 2 from a cycle $a^1 \succeq a^2 \succeq \cdots a^{2n} \succ a^1$, we can generate the table of joint actions in Figure C.2. Here, the dashed arrows and the solid arrow represent the links in the cycle as in Figure C.1.

The proof by induction argument requires the following steps. Step 1 to 3 gives preferences shown in Figure C.7, and Step 4 shows other preferences as reflected in Figure C.9-(ii). Step 5 shows the contradiction of these preferences identified in Step 1 to 3 and Step 4.

[1] Again for player 2, $\succeq^{**}$ is defined inversely from the convention of base relation. Accordingly, arrows in the figures inversely represent player 2's revealed preference.

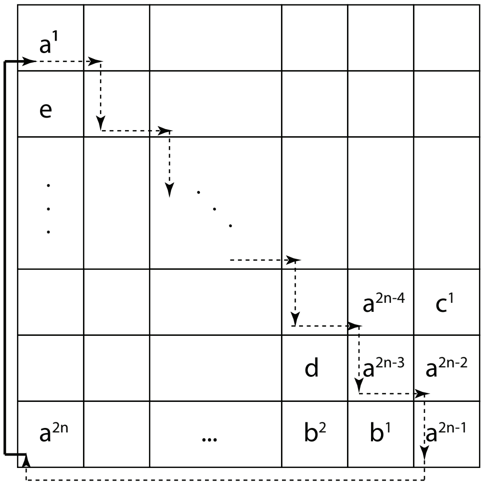


Figure C.2: A cycle of length 2n ($n \geq 3$)

Step 1: Consider the feasible set $\{a^{2n-3}, a^{2n-2}, a^{2n-1}, b^1\}$. In addition to the known preferences from the cycle, we can verify $f(\{a^{2n-3}, b^1\})$ and $f(\{b^1, a^{2n-1}\})$. The four cases in Figure C.3 below contain all possible cases of $f(\{a^{2n-3}, b^1\})$ and $f(\{b^1, a^{2n-1}\})$. In these two feasible sets, it must not be the case that either $a^{2n-3} \in f(\{a^{2n-3}, b^1\})$ and $a^{2n-1} \in f(\{b^1, a^{2n-1}\})$ (fig (i)), or $b^1 \in f(\{a^{2n-3}, b^1\})$ and $b^1 \in f(\{b^1, a^{2n-1}\})$ (fig (ii)).

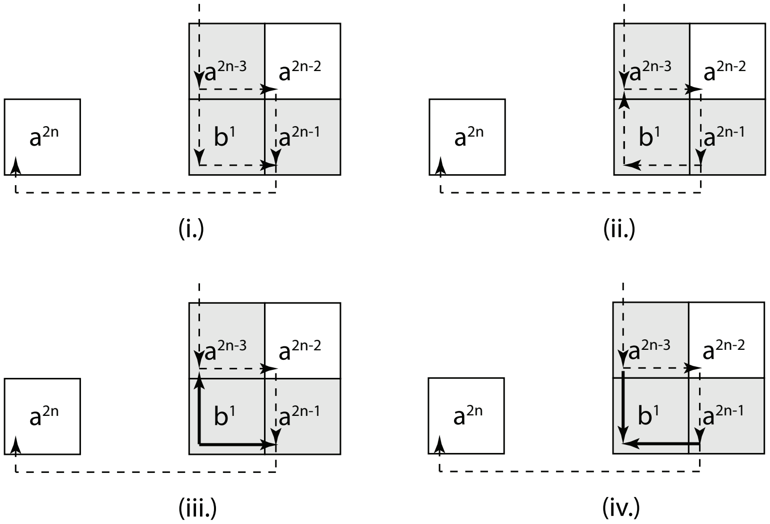


Figure C.3: A part of the cycle with length $2n$

In case (i), $a^{2n-4} \succeq^* b^1$ by transitivity of $\succeq^*$ in the left column, and $b^1 \succeq^{**} a^{2n}$ by transitivity of $\succeq^{**}$ in the bottom row. These two preferences induce the cycle $a^1 \succeq$

$\cdots \succeq a^{2n-4} \succeq b^1 \succeq a^{2n} \succ a^1$ which has length $2(n-1)$, a contradiction. In case (ii), $b^1 \in f(\{a^{2n-3}, b^1\}) \cap f(\{b^1, a^{2n-1}\})$ and $a^{2n-2} \in f(\{a^{2n-3}, a^{2n-2}\}) \cap f(\{a^{2n-2}, a^{2n-1}\})$. (PE) induces that a^{2n-2} and b^1 are in $f(\{a^{2n-3}, a^{2n-2}, a^{2n-1}, b^1\})$; interchangeability of f implies that all four joint actions are in $f(\{a^{2n-3}, a^{2n-2}, a^{2n-1}, b^1\})$. Therefore, we have an indifference relation $\sim$ in $\{a^{2n-3}, b^1\}$ and $\{b^1, a^{2n-1}\}$, which gives a special case of (i).

Excluding case (i) and (ii), either (iii) or (iv) must be true. We will prove that the induction step is true in case (iii). The proof in case (iv) is omitted here as it can be shown with exactly the same approach as that taken in case (iii).

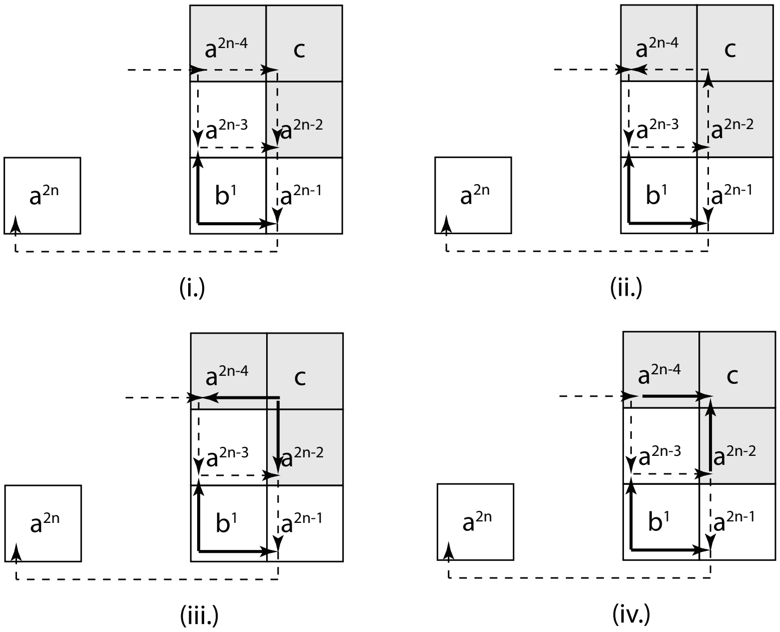


Figure C.4: Verifying more preferences

Step 2: Figure C.4 contains every possible case of $f(\{a^{2n-4}, c^1\})$ and $f(\{c^1, a^{2n-2}\})$. Using the same argument used for the case (i) and (ii) of $f(\{a^{2n-3}, b^1\})$ and $f(\{b^1, a^{2n-1}\})$ in Step 1, we can rule out the cases of (i) and (ii) in Figure C.4. In addition, case (iii), $\{a^{2n-4}\} = f(\{a^{2n-4}, c^1\})$ and $\{c^1\} = f(\{c^1, a^{2n-2}\})$, is not possible either. This can be shown first by observing $b^1 \succ^* a^{2n-4}$. If it is not the case, completeness of $\succeq^*$ in the left column gives $a^{2n-4} \succeq^* b^1$ which, combined with $b^1 \succeq^{**} a^{2n}$ by transitivity of $\succeq^{**}$ in the bottom row, induces the cycle $a^1 \succeq^{**} \cdots \succeq^{**} a^{2n-4} \succeq^* b^1 \succeq^{**} a^{2n} \succ^* a^1$ whose length is

$2(n-1)$.

Once (iii) and $b^1 \succ^* a^{2n-4}$ are obtained (see Figure C.5), we consider the set of joint actions $\{a^{2n-4}, c^1, b^1, a^{2n-1}\}$. Any choice from this feasible set violates the (PC) in one feasible subset of $\{a^{2n-4}, c^1, b^1, a^{2n-1}\}$. Suppose $c^1 \in f(\{a^{2n-4}, c^1, b^1, a^{2n-1}\})$, then $c \notin f(\{a^{2n-4}, c^1\})$ violates (PC). Likewise any joint action in $\{a^{2n-4}, c^1, b^1, a^{2n-1}\}$ is not a choice. Thus case (iv), $\{c^1\} = f(\{a^{2n-4}, c^1\})$ and $\{a^{2n-2}\} = f(\{c^1, a^{2n-2}\})$, must be true.

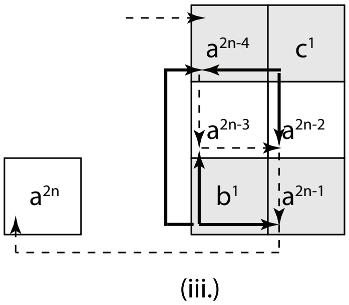


Figure C.5: Ruling out the case (iii)

Step 3: Considering $f(\{a^{2n-5}, d\})$ and $f(\{d, a^{2n-3}\})$, we can rule out the cases of either $a^{2n-5} \in f(\{a^{2n-5}, d\})$ and $a^{2n-3} \in f(\{d, a^{2n-3}\})$, or $d \in f(\{a^{2n-5}, d\})$ and $d \in f(\{d, a^{2n-3}\})$ by the same argument used for $f(\{a^{2n-3}, b^1\})$ & $f(\{b^1, a^{2n-1}\})$ and $f(\{a^{2n-4}, c\})$ & $f(\{c, a^{2n-2}\})$ in the previous steps. Accordingly, we only have cases of either $\{a^{2n-5}\} = f(\{a^{2n-5}, d\})$ and $\{d\} = f(\{d, a^{2n-3}\})$, or $\{d\} = f(\{a^{2n-5}, d\})$ and $\{a^{2n-3}\} = f(\{d, a^{2n-3}\})$; case (i) or case (ii) in Figure C.6, respectively. Case (i) is ruled out because once we have $a^{2n-5} \succ^* d$, it must be that $a^{2n-2} \succ^{**} d$. If this is not true, then $d \succeq^{**} a^{2n-2}$, which induces one of the following cases.

1. If the cycle has length 6 (a^{2n-5} is a^1 and there is no # in fig(i)), b^2 is equal to a^{2n}. Thus we have $a^{2n-1} \succeq^{**} b^2$ and $b^2 \succ^* d$ by transitivity of $\succeq^*$. As a result, $d \succeq^{**} a^{2n-2}$ makes a cycle with length 4, $d \succeq^{**} a^{2n-2} \succeq^* a^{2n-1} \succeq^{**} b^2 \succ^* d$, which contradicts the induction hypothesis.

2. If the cycle has length 8 or more (there is a^{2n-6} , '#' in the fig (i), which is not a^1), $a^{2n-6} \succeq^* d \succeq^{**} a^{2n-2}$ by transitivity of $\succeq^*$ and $\succeq^{**}$ in the left column and the middle

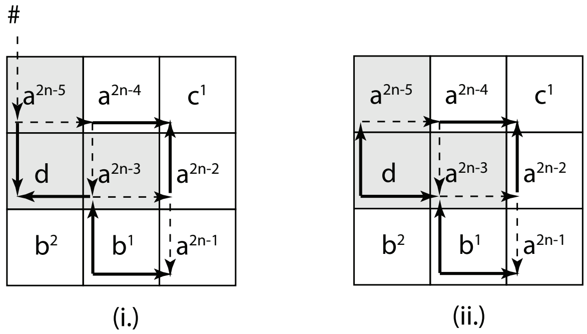


Figure C.6: Verifying more preferences.

row. These preferences shorten the cycle, which contradicts the induction hypothesis.

Therefore, $a^{2n-2} \succ^{**} d$ must be true in case (i). Regardless of what is in $f(\{a^{2n-5}, d, c^1, a^{2n-2}\})$, it violates (PC). For instance, if $d \in f(\{a^{2n-5}, d, c^1, a^{2n-2}\})$ then it must be $d \in f(\{a^{2n-5}, d\})$, which violates $a^{2n-5} \succ^{*} d$. Consequently, case (ii) in Figure C.6 must be the option.

By applying Step 2 and 3 sequentially, we can verify more preferences. Figure C.7 summarizes the result of this process. In the following proof, Step 4 is necessary only for a cycle whose length is at least 8. For a cycle with length 6, we already know all the preferences that we will verify in Step 4.

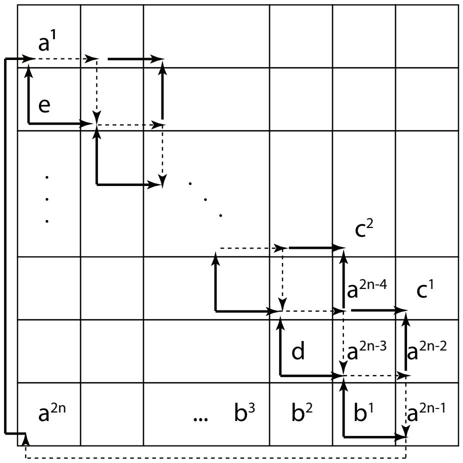


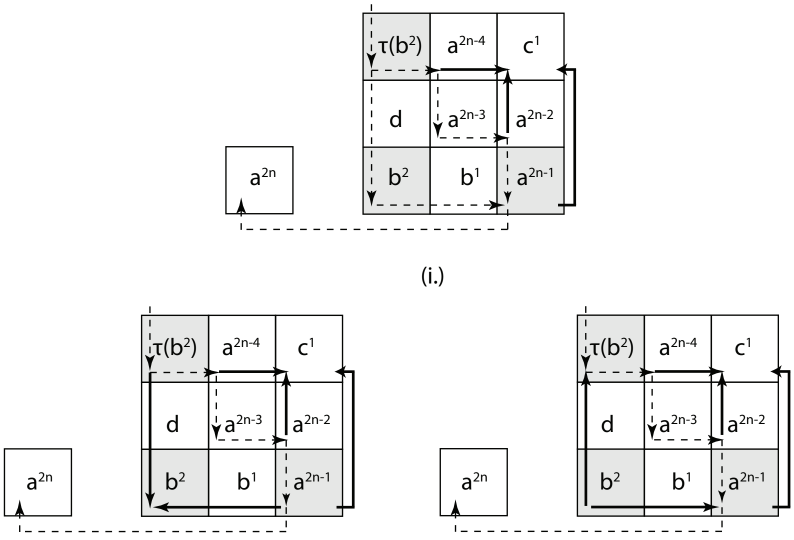
Figure C.7: Preferences verified in Step 2 and 3

Step 4: Denote the joint action $(a_1^{2n-1}, a^{2(n-k)-1})$ as b^m and the joint action $(a_1^{2(n-m-1)}, a_2^{2(n-m)})$ as c^m, where $k = 1, 2, \ldots, n-2$. Figure C.7 shows where b^m and c^m $(1 \leq m \leq n-2)$ are located. Let τ be a function from $\{b^1, b^2, \ldots, b^{n-2}\}$ to A such that $\tau(b^m) = (a_1^{2n-(2m+1)}, b_2^m)$. Figures C.8, C.9, and C.10 show how the function values are located in the feasible set table. ($\tau(b^m)$ takes its place on the stairway of which b^m is at the bottom.) We prove the following claim.

Claim C.1.1. *For any b^m $(1 \leq m \leq n-2)$, $b^m \succ \tau(b^m)$ and $b^m \succ a^{2n-1}$*

Proof. We prove by induction. Note that we already proved in Step 2 that this claim holds for b^1.

<u>Induction 1</u>: The claim holds for b^2. That is, $b^2 \succ^* \tau(b^2)$ (or a^{2n-5}) and $b^2 \succ^{**} a^{2n-1}$.



(i.)

(ii.)

(iii.)

Figure C.8: Verifying more preferences involving b^2

Proof. Considering feasible sets, $\{\tau(b^2), b^2\}$ and $\{b^2, a^{2n-1}\}$ (see Figure C.8), it is not the case that $\tau(b^2) \in f(\{\tau(b^2), b^2\})$ and $a^{2n-1} \in f(\{b^2, a^{2n-5}\})$ (case (i)). Otherwise, it shortens the cycle with $a^{2n-5} = \tau(b^2) \succeq^* b^2 \succeq^{**} a^{2n}$. (We used transitivity in the bottom row.) Therefore, by completeness in each line, we should have either $a^{2n-1} \succ^{**} b^2$ or $b^2 \succ^* \tau(b^2)$. In the former case, in order not to have a cycle of length 6, which includes $\{\tau(b^2), a^{2n-4}, a^{2n-3}, a^{2n-2}, a^{2n-1}, b^2\}$, f must give $\tau(b^2) \succ^* b^2$ (fig (ii)). In the latter case, in order not to have a cycle of length 6, f must give $b^2 \succ^{**} a^{2n-1}$ (fig (iii)). However, case (ii) is ruled out by considering the feasible set, $\{\tau(b^2), c^1, b^2, a^{2n-1}\}$. To demonstrate this, note that $a^{2n-1} \succ^* c^1$. Otherwise, $\tau(b^2) \succeq^{**} c^1 \succeq^* a^{2n-1}$ shortens the cycle. If case (ii) is true, then any choice from $\{\tau(b^2), c^1, b^2, a^{2n-1}\}$ violates (PC). For example, if $\tau(b^2) \in f(\{\tau(b^2), c^1, b^2, a^{2n-1}\})$, then it must be true that $\tau(b^2) \in f(\{\tau(b^2), c^1\})$. This contradicts $\tau(b^2) \succ^{**} c^1$. (Note again that $\succeq^{**}$ is defined inversely.) Therefore, (iii) must be the case in Figure C.8. □

Induction 2: If the claim holds for b^{m-2}, it also holds for b^m $(3 \leq m \leq n-2)$.

Proof. With the same approach as Induction 1, f should not give $\tau(b^m) \succeq^* b^m$ and $b^m \succeq^{**} a^{2n-1}$; otherwise, we have a shorter cycle including $\tau(b^m) \succeq^* b^m \succeq^{**} a^{2n}$. Thus, it must be either $a^{2n-1} \succ^{**} b^m$ or $b^m \succ^* \tau(b^m)$. In the former case, not to have a cycle, $b^m \succeq^* \tau(b^m) \succeq^{**} \cdots \succeq^* a^{2n-1} \succ^{**} b^m$ which has length $2m+2 \leq 2(n-1)$, it must be true that $\tau(b^m) \succ^* b^m$ (case (i) in Figure C.9).[2] In the latter case, not to have a cycle, $\tau(b^m) \succeq^{**} \cdots \succeq^* a^{2n-1} \succeq^* b^m \succ^* \tau(b^m)$ which has length $2m+2 \leq 2(n-1)$, it must be true that $b^m \succ^{**} a^{2n-1}$. (case (ii) in Figure C.9.) However, case (i) is ruled out. First, observe that $b^{m-2} \succ^* c^{m-1}$ must be true; otherwise $\tau(b^m) \succeq^{**} c^{m-1} \succeq^* b^{m-2} \succeq^{**} a^{2n}$ leads to a shorter cycle. In addition, transitivity of $\succeq^{**}$ in the bottom row gives $b^{m-2} \succ^{**} b^m$. Then, in the feasible set, $\{\tau(b^m), b^m, b^{m-2}, c^{m-1}\}$, any choice violates (PC). Therefore, (ii) must be the case in $f(\{\tau(b^m), b^m\})$ and $f(\{b^m, a^{2n-1}\})$.

□

By induction, $b^m \succ \tau(b^m)$ and $b^m \succ a^{2n-1}$ for $m = 1, \ldots, n-2$. Claim C.1.1 holds.

□

Step 5: Results from Steps 2 and 3, and results from Step 4 contradict each other.

[2] Although we explicitly write the proof only for the case of cycle beginning with $\succeq^{**}$, every single step so far could have been reproduced for cases where cycles begin with $\succ^*$. Here, we used the induction hypothesis, "there is no cycle with a length of $2(n-1)$," from the counterpart proof of cycles begining with $\succeq^*$.

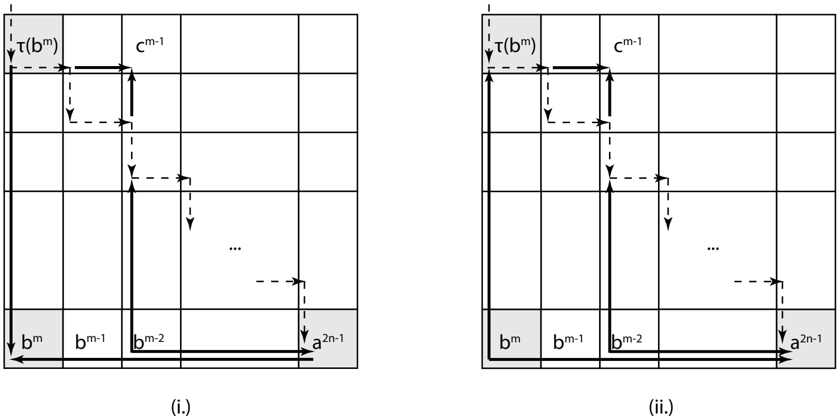


Figure C.9: Verifying preferences involving b^m

Proof. If we denote the joint action $(\tau(b^{n-2})_1, a_2^1)$ as e (see Figure C.10), then Step 2 and 3 gives $e \succ^* a^1$ and $e \succ^{**} \tau(b^{n-2})$. We showed in Step 4 that $b^{n-2} \succ^* \tau(b^{n-2})$ and $b^{n-2} \succ^{**} a^{2n-1}$. Moreover, it must be true that $e \succ^* a^{2n}$, since otherwise, $a^{2n} \succeq^* e \succ^{**} a^4$ shortens the cycle. On the other hand, $b^{n-2} \succ^{**} a^{2n}$ by transitivity of $\succeq^{**}$ in the bottom row. We can observe that any choice from the feasible set, $\{e, \tau(b^{n-2}), a^{2n}, b^{n-2}\}$, violates (PC). This contradiction completes the proof of Step 5, thereby completing the proof of consistency of $\succeq$. □

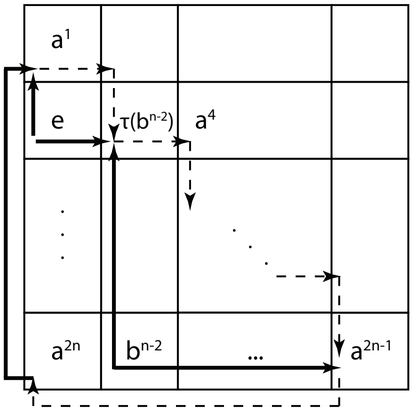


Figure C.10: A contradiction

C.2 Characterizing a rationalizing preference relation.

Claim C.2.1. *For all $B \in \mathcal{A}$, $f(B)$ coincides with the set of all Nash equilibria of the game $(B, \succeq, \preceq)$.*

Proof. Take any $B = B_1 \times B_2 \in \mathcal{A}$, and let $\mathrm{NE}(B)$ be the set of all Nash equilibria of the game $(B, \succeq, \preceq)$. First, to show $f(B) \subset \mathrm{NE}(B)$, we take any $b^* = (b_1^*, b_2^*) \in f(B)$. Since f satisfies (PC), $b^* \in f(B^{'})$ for all $B^{'} \in \mathcal{A}$ and $B^{'} \subset B$. Therefore, for any $\{b^*, (b_1, b_2^*)\} \subset B$, $b^* \in f(\{b^*, (b_1, b_2^*)\})$. By the definition of $\succeq^*$, we have $b^* \succeq^* (b_1, b_2^*)$, which is equal to $b^* \succeq (b_1, b_2^*)$. Similarly, for any $\{b^*, (b_1^*, b_2)\} \subset B$, $b^* \in f(\{b^*, (b_1^*, b_2)\})$. The definition of $\succeq^{**}$ gives $(b_1^*, b_2) \succeq^{**} b^*$, which is equal to $(b_1^*, b_2) \succeq b^*$, or $b^* \preceq (b_1^*, b_2)$. Since $b^* \succeq (b_1, b_2^*)$ and $b^* \preceq (b_1^*, b_2)$, for all $(b_1, b_2^*) \in B$ and $(b_1^*, b_2) \in B$, b^* is a Nash equilibrium of the game $(B, \succeq, \preceq)$.

Conversely, if $b^* \in \mathrm{NE}(B)$, for any $(b_1, b_2^*) \in B$, $b^* \succeq (b_1, b_2^*)$. Since, only $\succeq^*$, and not $\succeq^{**}$, is defined in columns, we have $b^* \succeq^* (b_1, b_2^*)$. The definition of $\succeq^*$ gives $b^* \in f(\{b^*, (b_1, b_2^*)\})$, and (PE) implies $b^* \in f(B_1 \times \{b_2^*\})$ (#). $b^* \in \mathrm{NE}(B)$ implies $b^* \preceq (b_1^*, b_2)$ for all $(b_1^*, b_2) \in B$ (or $(b_1^*, b_2) \succeq b^*$). Because we defined only $\succeq^{**}$, and not $\succeq^*$, in rows, we have $(b_1^*, b_2) \succeq^{**} b^*$. The definition of $\succeq^{**}$ gives $b^* \in f(\{b^*, (b_1^*, b_2)\})$ and (PE) induces $b^* \in f(\{b_1^*\} \times B_2)$ (##). Lastly, (#), (##), and (PE) imply that $b^* \in f(B)$. □

We have shown that $\succeq$ is consistent and $f(B)$ coincides with $\mathrm{NE}(B)$ for all $B \in \mathcal{A}$. Suzumura (1976) shows that a consistent relation has a weak order extension. Since the extension generates additional preferences only between two joint choices which are not in a line, this extension does not affect the result of Claim C.2.1. Therefore, Claim C.2.1 is still valid with the weak order extension of $\succeq$. This completes the proof of the main theorem.

Bibliography

Abdulkadiroglu, A., P. Pathak, and A. Roth (2009): "Strategy-proofness versus Efficiency in Matching with Indifferences: Redesigning the NYC High School Match," *The American Economic Review*, 99(5), 1954–1978.

Abdulkadiroglu, A., P. Pathak, A. Roth, and T. Sönmez (2006): "Changing the Boston school choice mechanism," .

Abdulkadiroglu, A., and T. Sönmez (2003): "School choice: A mechanism design approach," *The American Economic Review*, 93(3), 729–747.

Alcalde, J., and S. Barberà (1994): "Top dominance and the possibility of strategy-proof stable solutions to matching problems," *Economic theory*, 4(3), 417–435.

Ashlagi, I., M. Braverman, and A. Hassidim (2011): "Stability in Large Matching Markets with Complementarities," Discussion paper, Mimeo.

Austen-Smith, D., and J. S. Banks (1994): *Positive Political Theory I : Collective Preference*. Univ. of Michigan Press, Ann Arbor.

——— (1996): "Information Aggregation, Rationality, and the Condorcet Jury Theorem," *The American Political Science Review*, 90(1), 34–45.

Austen-Smith, D., and T. Feddersen (2005): "Deliberation and voting rules," *Social Choice and Strategic Decisions*, pp. 269–316.

Austen-Smith, D., and T. Feddersen (2006): "Deliberation, preference uncertainty, and voting rules," *American Political Science Review*, 100(02), 209–217.

AZEVEDO, E. (2010): “Imperfect Competition in Two-Sided Matching Markets,” Discussion paper, working paper, Harvard University.

AZEVEDO, E., AND J. LESHNO (2011): “The College Admissions Problem With a Continuum of Students,” Discussion paper, Mimeo.

BACCARA, M., A. COLLARD-WEXLER, L. FELLI, AND L. YARIV (2010): “Child adoption matching: preferences for gender and race,” .

BERNAL, R., L. HU, C. MORIGUCHI, AND E. NAGYPAL (2007): “Child Adoption in the United States: Historical Trends and the Determinants of Adoption Demand and Supply, 1951-2002,” .

BIBAS, S. (2004): “Plea Bargaining outside the Shadow of Trial,” *Harvard Law Review*, 117(8), 2463–2547.

BOGOMOLNAIA, A., AND H. MOULIN (2001): “A new solution to the random assignment problem,” *Journal of Economic Theory*, 100(2), 295–328.

BULOW, J., AND J. LEVIN (2006): “Matching and Price Competition,” *American Economic Review*, 96(3), 652–668.

CARVAJAL, A., R. DEB, J. FENSKE, AND J. QUAH (2010): “Revealed Preference Tests of the Cournot Model,” *Working Paper No. 14998, University of Oxford.*

CHAMBERS, C., AND F. ECHENIQUE (2011): “Testable implications of Bargaining Theories,” *Social Science Working Paper No. 1348, Caltech.*

CHE, Y., AND F. KOJIMA (2010): “Asymptotic equivalence of probabilistic serial and random priority mechanisms,” *Econometrica*, 78(5), 1625–1672.

CHERCHYE, L., T. DEMUYNCK, AND B. DE ROCK (2011): “The Empirical Content of Cournot Competition,” *Working Paper, Erasmus Research Institute of Management.*

CHO, I.-K., AND D. M. KREPS (1987): “Signaling Games and Stable Equilibria,” *The Quarterly Journal of Economics*, 102(2), 179–221.

Choo, E., and A. Siow (2006): "Who marries whom and why," *Journal of Political Economy*, 114(1), 175.

Clark, S. (2006): "The uniqueness of stable matchings," *The BE Journal of Theoretical Economics*, 6(1), 8.

Coles, P., J. Cawley, P. Levine, M. Niederle, A. Roth, and J. Siegfried (2010): "The Job Market for New Economists: A Market Design Perspective," *Journal of Economic Perspectives*, 24(4), 187–206.

Condorcet, M. (1785): "Essai sur lapplication de lanalyse à la probabilité des decisions rendues a la pluralité des voix," *Paris: Limprimerie royale.*

Cooter, R., and D. Rubinfeld (1989): "Economic analysis of legal disputes and their resolution," *Journal of Economic Literature*, 27(3), 1067–1097.

Coughlan, P. (2000): "In defense of unanimous jury verdicts: Mistrials, communication, and strategic voting," *The American Political Science Review*, 94(2), 375–393.

Crawford, V. (2008): "The flexible-salary match: a proposal to increase the salary flexibility of the national resident matching program," *Journal of Economic Behavior & Organization*, 66(2), 149–160.

Dawande, M., P. Keskinocak, J. Swaminathan, and S. Tayur (2001): "On bipartite and multipartite clique problems," *Journal of Algorithms*, 41(2), 388–403.

Demange, G., D. Gale, and M. Sotomayor (1987): "A further note on the stable matching problem," *Discrete Applied Mathematics*, 16, 217–222.

Demuynck, T., and L. Lauwers (2009): "Nash rationalization of collective choice over lotteries," *Mathematical Social Sciences*, 57(1), 1–15.

Dubins, L., and D. Freedman (1981): "Machiavelli and the Gale-Shapley algorithm," *The American Mathematical Monthly*, 88(7), 485–494.

Duersch, P., J. Oechssler, and B. Schipper (2011): "Unbeatable Imitation," *Working Paper, University of California, Davis.*

——— (forthcoming): "Pure Strategy Equilibria in Symmetric Two-Player Zero-Sum Games," *International Journal of Game Theory.*

Echenique, F., and L. Ivanov (2011): "Implications of Pareto efficiency for two-agent (household) choice," *Journal of Mathematical Economics*, 47(2), 129–136.

Eeckhout, J. (2000): "On the uniqueness of stable marriage matchings," *Economics Letters*, 69(1), 1–8.

Feddersen, T., and W. Pesendorfer (1998): "Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting," *The American Political Science Review*, 92(1), 23–35.

Feddersen, T. J., and W. Pesendorfer (1996): "The Swing Voter's Curse," *The American Economic Review*, 86(3), 408–424.

Forges, F., and E. Minelli (2009): "Afriat's theorem for general budget sets," *Journal of Economic Theory*, 144(1), 135–145.

Galambos, A. (2009): "The complexity of nash rationalizability," *Working Paper, Lawrence University.*

Gale, D., and L. Shapley (1962): "College admissions and the stability of marriage," *American Mathematical Monthly*, 69(1), 9–15.

Gerardi, D., and L. Yariv (2007): "Deliberative voting," *Journal of Economic Theory*, 134(1), 317–338.

Goeree, J., and L. Yariv (Forthcoming): "An experimental study of collective deliberation," *Econometrica.*

Grossman, G., and M. Katz (1983): "Plea bargaining and social welfare," *The American Economic Review*, 73(4), 749–757.

Guarnaschelli, S., R. D. McKelvey, and T. R. Palfrey (2000): "An Experimental Study of Jury Decision Rules," *The American Political Science Review*, 94(2), 407–423.

Halaburda, H. (2010): "Unravelling in two-sided matching markets and similarity of preferences," *Games and Economic Behavior*, 69(2), 365–393.

Hitsch, G., A. Hortaçsu, and D. Ariely (2010): "What makes you click?Mate preferences in online dating," *Quantitative Marketing and Economics*, 8(4), 393–427.

Immorlica, N., and M. Mahdian (2005): "Marriage, honesty, and stability," in *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 53–62. Society for Industrial and Applied Mathematics.

Knuth, D. (1976): *Mariages stables.* Les Presse De L'Universite De Montreal.

Kojima, F., and M. Manea (2010): "Incentives in the probabilistic serial mechanism," *Journal of Economic Theory*, 145(1), 106–123.

Kojima, F., and P. Pathak (2009): "Incentives and stability in large two-sided matching markets," *The American Economic Review*, 99(3), 608–627.

Kojima, F., P. Pathak, and A. Roth (2010): "Matching with Couples: Stability and Incentives in Large Markets," *NBER Working Papers.*

Liu, Q., and M. Pycia (2011): "Ordinal Efficiency, Fairness, and Incentives in Large Markets," .

Luce, R., and H. Raiffa (1989): *Games and decisions: Introduction and critical survey.* Dover Pubns.

McKinney, C., M. Niederle, and A. Roth (2003): "The collapse of a medical clearinghouse (and why such failures are rare)," .

McVitie, D., and L. Wilson (1970): "Stable marriage assignment for unequal sets," *BIT Numerical Mathematics*, 10(3), 295–309.

Mnookin, R. H., and L. Kornhauser (1979): "Bargaining in the Shadow of the Law: The Case of Divorce," *The Yale Law Journal*, 88(5), 950–997.

Moulin, H. (1985): “Choice functions over a finite set: A summary,” *Social Choice and Welfare*, 2(2), 147–160.

Nash, J. (1951): “Non-cooperative games,” *Annals of mathematics*, 54(2), 286–295.

Niederle, M., and A. Roth (2005): “The gastroenterology fellowship market: Should there be a match?,” *The American economic review*, 95(2), 372–375.

Pathak, P., and T. Sönmez (2008): “Leveling the playing field: Sincere and sophisticated players in the Boston mechanism,” *The American Economic Review*, 98(4), 1636–1652.

Peleg, B., and S. Tijs (1996): “The consistency principle for games in strategic form,” *International Journal of Game Theory*, 25(1), 13–34.

Pittel, B. (1989): “The Average Number of Stable Matchings,” *SIAM Journal on Discrete Mathematics*, 2(4), 530–549.

Plott, C. (1974): “On game solution and revealed preference theory,” *Social Science Working Paper No. 35, Caltech.*

Priest, G. L., and B. Klein (1984): “The Selection of Disputes for Litigation,” *The Journal of Legal Studies*, 13(1), 1–55.

Rabe, G., and D. Champion (2002): “Criminal Courts: Structure, Process, and Issues,” *No.: ISBN 0-13-780388-5*, p. 494.

Radzik, T. (1991): “Saddle point theorems,” *International Journal of Game Theory*, 20, 23–32.

Ray, I., and S. K. Snyder (2003): “Observable Implications of Nash and Subgame - Perfect Behavior in Extensive Games,” *Working Paper, University of Birmingham.*

Ray, I., and L. Zhou (2001): “Game Theory via Revealed Preferences,” *Games and Economic Behavior*, 37(2), 415 – 424.

Reinganum, J. (1988): “Plea bargaining and prosecutorial discretion,” *The American Economic Review*, 78(4), 713–728.

Richter, M. (1971): "Rational choice," in *Preferences, Utility, and Demand*, pp. 29–58. Harcourt Brace Jovanovich, New York.

Roth, A. (1982): "The economics of matching: Stability and incentives," *Mathematics of Operations Research*, 7(4), 617–628.

——— (1984): "The evolution of the labor market for medical interns and residents: a case study in game theory," *The Journal of Political Economy*, 92(6), 991–1016.

——— (2002): "The economist as engineer: Game theory, experimentation, and computation as tools for design economics," *Econometrica*, 70(4), 1341–1378.

——— (2008): "Deferred acceptance algorithms: History, theory, practice, and open questions," *International Journal of Game Theory*, 36(3), 537–569.

Roth, A., and E. Peranson (1999): "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design," *American Economic Review*, 89(748), 80.

Roth, A., and U. Rothblum (1999): "Truncation strategies in matching marketsin search of advice for participants," *Econometrica*, 67(1), 21–43.

Roth, A., and M. Sotomayor (1990): *Two-sided matching*. Cambridge Univ. Pr.

Roth, A., and J. Vande Vate (1991): "Incentives in two-sided matching with random stable mechanisms," *Economic theory*, 1(1), 31–44.

Roth, A., and X. Xing (1994): "Jumping the gun: imperfections and institutions related to the timing of market transactions," *The American Economic Review*, 84(4), 992–1044.

Samet, D. (2011): "Matching of like rank and the size of the core in the marriage problem," Discussion paper, Mimeo.

Samuelson, P. A. (1938): "The Empirical Implications of Utility Analysis," *Econometrica*, 6(4), 344–356.

Sen, A. K. (1971): "Choice Functions and Revealed Preference," *Review of Economic Studies*, 38(115), 307–17.

Serfling, R. (1980): *Approximation theorems of mathematical statistics*, vol. 371. Wiley New York.

Shapley, L. (1964): "Some Topics in Two Person Games," in *Advances in Game Theory*, ed. by M. Dresher, L. Shapley, and A. Tucker, pp. 1–28. Princeton University Press.

Sönmez, T. (1999): "Strategy-proofness and Essentially Single-valued Cores," *Econometrica*, 67(3), 677–689.

Sönmez, T., and T. Switzer (2011): "Matching with (Branch-of-Choice) Contracts at United States Military Academy," *Boston College Working Papers in Economics*.

Sprumont, Y. (2000): "On the Testable Implications of Collective Choice Theories," *Journal of Economic Theory*, 93(2), 205 – 232.

Stuntz, W. J. (2004): "Plea Bargaining and Criminal Law's Disappearing Shadow," *Harvard Law Review*, 117(8), 2548–2569.

Suzumura, K. (1976): "Remarks on the Theory of Collective Choice," *Economica*, 43(172), 381–390.

Wilson, R. (1970): "The finer structure of revealed preference," *Journal of Economic Theory*, 2(4), 348–353.

Xu, Y., and L. Zhou (2007): "Rationalizability of choice functions by game trees," *Journal of Economic theory*, 134(1), 548–556.