Incentive Compatibility of Large Centralized Matching Markets

SangMok Lee[*]

January 25, 2014

Abstract

We revisit questions concerning the manipulability of stable matching mechanisms, which are used in practice such as the National Resident Matching Program. Our setup introduces cardinal utilities from match partners to quantify incentives to manipulate stable mechanisms. We find that most agents in large matching markets are close to being indifferent over all possible stable matchings. In one-to-one matching, the utility gained by manipulating a stable mechanism is bounded by the gap between utilities from the best and the worst stable matching partners. Thus, the main finding implies that most agents in a large market would not have a significant incentive to manipulate stable mechanisms. We introduce techniques from the theory of random bipartite graphs for the analysis of large matching markets.

Keywords: two-sided matching, stable matching mechanism, large market, random bipartite graph

JEL Class: C78, D61, D78

[*] Department of Economics, University of Pennsylvania, Philadelphia, PA, 19104. E-mail: `sangmok-at-sas.upenn.edu` This paper is a part of my dissertation at Caltech. I am especially grateful to Federico Echenique, Matt Shum, and Leeat Yariv for encouragement and guidance. For very helpful discussions, I also thank Hyoung-Jun Ahn, Itai Ashlagi, Luke Boosey, Kim Border, Andrea Bui, Chris Chambers, Guilherme Freitas, Cary Frydman, Tadashi Hashimoto, John Hatfield, Miriam Kim, Fuhito Kojima, Sera Linardi, George Mailath, Ruth Mendel, Juan Pereyra, Andy Postlewaite, Jean-Laurent Rosenthal, Tom Ruchti, Bob Sherman, Erik Snowberg, Tayfun Sönmez, and Bumin Yenmez, as well as seminar participants in various places.

1 Introduction

1.1 Overview

We study a class of algorithms called *stable matching mechanisms*, which are often used for a number of centralized labor matching markets, most famously the National Resident Matching Program (NRMP).[1] We ask why stable matching mechanisms have been so successful, despite their easy manipulability by the participants through preference misrepresentation. In particular, we analyze whether large markets, i.e., those consisting of a large number of participants, mitigate the incentives to misrepresent preferences.

In a two-sided matching market, agents of one kind match with agents of the other kind. A matching is regarded as *stable* if no agent would rather remain unmatched than matching to her current partner, and no pair of agents on opposite sides of the market prefer each other to their current partners. In a centralized matching market, a stable matching mechanism takes preference reports from participants and produces a stable matching with respect to the reported preferences. The concept of stability has been considered of central importance. Most successful mechanisms in practice implement a stable matching with respect to submitted preferences (Roth and Xing, 1994; Roth, 2002; McKinney, Niederle, and Roth, 2005). With few exceptions, stable matching mechanisms have been successful, whereas unstable mechanisms have mostly failed. For the NRMP, stability is indeed *required* as the Association of American Medical Colleges chose stability as a key property of the new algorithm (Roth, 1984).

Stable matching mechanisms, however, have one significant limitation. While the mechanisms produce stable matchings under the assumption that all participants reveal their true preferences, *no stable matching mechanism is strategy-proof* (Roth, 1982): a market participant may achieve a better matching by misrepresenting her preferences, either by changing the order of the preference lists or by announcing that some acceptable agents are unacceptable.[2] Even the current NRMP cannot entirely rule out such incentives for strategic misrepresentation. As stable matching mechanisms are not incentive compatible, the mechanisms may be manipulated by market participants, thereby not implementing the intended matchings. Moreover, each participant's decision may become difficult since she needs to best respond to other agents' strategic manipulations.

[1] See Roth and Peranson (1999) for various professional labor markets with centralized matching mechanisms.

[2] Alcalde and Barberà (1994) and Sönmez (1999) show that strategy-proofness is incompatible not only with stability but even with weaker conditions of Pareto efficiency and individual rationality.

We consider a baseline model of one-to-one matching where each firm hires one worker. This model is a tailored setup to simplify the NRMP: each year the market has thousands of participating residency programs each of which offers only a few positions.[3] We quantify incentives to manipulate a stable matching mechanism by assuming that each firm-worker pair receives utilities, one for the firm and one for the worker. In order to study the likelihood a market has a large proportion of agents who have significant incentives to manipulate, we assume that utilities are randomly drawn from some underlying distributions.

The key finding is that the proportion of market participants who can potentially achieve a significant utility gain from manipulation vanishes as the market increases. Given the tangible and intangible costs of strategic behavior in real life, we believe that this result may alleviate the concerns about manipulability, and therefore support the use of stable matching mechanisms in practice. In addition, based on our findings, market designers may more confidently advise participants to submit their true preferences.

The main question has already been addressed in several previous studies, especially in Roth and Peranson (1999), Immorlica and Mahdian (2005), and Kojima and Pathak (2009).[4] The previous studies consider a certain stable mechanism, where it is a dominant strategy for workers to reveal their true preferences, and focus on firms' incentives to misrepresent their preferences. In their model, the proportion of firms that have any incentive to misrepresent their preferences converges to zero as the market increases.

These papers rightly pointed out that stable mechanisms are hard to manipulate in large markets, and they certainly deserve credit for showing the incentive compatibility of stable mechanisms in large markets. However, as Roth and Peranson point out, the previous result is based on an assumption, *limited acceptability*: agents on one side (e.g., workers) consider only a vanishingly small proportion of agents on the other side acceptable as the market size increases.[5] Market participants often have similar preferences. For example, medical school graduates prefer top-tier hospitals, and hospitals prefer candidates with strong recommendation letters and high test scores on their medical exams. The limited acceptability assumption, combined with the commonality of preferences, may lead to many participants

[3] In the NRMP 2013, more than 4,600 residency programs are matched with more than 28,000 doctors. On average, each program hires about 6 doctors. The summary data is available on `http://www.nrmp.org/match-data/main-residency-match-data/`.

[4] Feldin (2003) also considers a similar model and studies properties of stable matchings. Since the paper considers a market with no commonality of preferences, it does not have the issue of a large fraction of unmatched participants discussed in the next paragraph.

[5] For instance, if the limit is 30, each worker considers only the 30 most preferred firms to be acceptable, and all other firms to be unacceptable. The limit can grow, but the proportion of acceptable firms must converge to zero.

in a large market remaining unmatched in stable matchings.

In this paper, we show incentive compatibility of stable mechanisms as a pure property of market size, without resorting to the limited acceptability assumption. The key modeling strategy is that we consider *random cardinal utilities* by which ordinal preferences are determined. With cardinal utilities, we can ask if agents have *significant* incentives to manipulate a mechanism. We consider utilities which are randomly drawn from some underlying distributions. Thus, we can ask how likely it is that a realized market contains a large proportion of agents with significant incentives to manipulate. This modeling approach is very distinct from the previous large market approach.

The model is a sequence of 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 that are randomly drawn from some underlying distributions on $\mathbb{R}_+$.[6] We formulate utility as a strictly increasing function of *common value* and *independent private value*. That is, when a firm f is matched with a worker w, the firm and the worker receive

$$\begin{aligned} U_{f,w} &= U(C_w\,,\,\zeta_{f,w}), \quad \text{and} \\ V_{f,w} &= V(C_f\,,\,\eta_{f,w}). \end{aligned} \tag{1}$$

Common values, C_f and C_w, are intrinsic values of f and w, which induce vertical preferences (e.g., top-tier vs. low-tier hospitals). Private values, $\zeta_{f,w}$ and $\eta_{f,w}$, are idiosyncratic utilities, which induce horizontal preferences (e.g., geographical preferences).

We find that while agents in a large market typically have multiple stable matching partners, most agents are close to being indifferent to all possible stable matchings (Theorem 1). The main result has a very simple intuition. Utility is given as a function of common value and private value. In the common-value dimension, firms and workers match assortatively: an agent gets as high a vertical match as possible given her own vertical quality. Then, given their vertical level, agents sort out very good matches in the private-value dimension (i.e., very good horizontal fits). Thus, the utility of firm f in any stable matching is

$$\begin{aligned} U_f^* \;\approx\; & U(\text{common value of a worker in the same position as } f, \\ & \text{maximum of the support of the workers' private values}). \end{aligned}$$

It is well-known in the matching literature that when a stable matching mechanism is applied in one-to-one matching market, the best an agent can achieve (by misrepresenting her

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

preferences) is matching with her best stable matching partner with regard to true preferences (Demange, Gale, and Sotomayor, 1987). As a result, 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 a significant incentive to manipulate the mechanism vanishes as the market size increases.

Furthermore, and as a consequence of our main result, we identify ϵ-Nash equilibrium behavior in which most market participants report their true preferences (Theorem 2).[7] Under the equilibrium behavior, some of those agents with significant incentives misrepresent their preferences. Nevertheless, the rest of the participants still have no significant incentive to respond to such manipulations.

From a methodological standpoint, we introduce new techniques from random bipartite graph theory to matching models.[8] To prove the main theorem, in each market realization, we 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 private values is significantly lower than the upper bound of the support. Then, 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 such a relatively large subset of nodes becomes small as the initial set of nodes increases (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 size increases.

We extend our model to markets with incomplete information. We consider that common values are known to all market participants but private values are only privately known. Our main result holds: most participants in a large market have a very small expected

[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.

[8] Two recent papers, Ashlagi and Roth (2013) and Ashlagi, Gamarnik, Rees, and Roth (2012), also introduce a random graph model for an analysis of large kidney exchange matching markets. They use the Erdos-Rényi theorem, which is different from the current paper's technique.

difference between utilities from the best and the worst stable matchings. (Theorems 4 and E.1). The direct implication on incentive compatibility also holds: When a stable matching mechanism is applied and all agents reveal their true preferences, most participants do not have significant incentives of manipulation. Also we identify ϵ-Nash equilibrium for a market with pure private-value utilities (i.e., $U_{f,w} = \zeta_{f,w}$). In this case, it is most likely that all participants in a large market reveal their true preferences as an ϵ-Bayesian-Nash equilibrium behavior.

1.2 Related Literature

The closest studies to this paper are Roth and Peranson (1999), Immorlica and Mahdian (2005), and Kojima and Pathak (2009). These studies consider a stable matching mechanism, called the worker-optimal stable mechanism, which implements a stable matching favorable to workers. As it is a dominant strategy for workers to truthfully reveal their preferences (Roth, 1982; Dubins and Freedman, 1981), the papers focus on firms' incentives to misrepresent their preferences.

Roth and Peranson (1999), based on a computational analysis, have shown that the proportion of firms that have any incentive to manipulate the mechanism converges to zero as the market size increases. This convergence is theoretically proven by Feldin (2003) for one-to-one matching without commonality of preferences (cf., pure private-value utilities), by Immorlica and Mahdian (2005) for one-to-one matching with general preferences, and by Kojima and Pathak (2009) for many-to-one matching.[9]

As we mentioned before, the above results are derived under the limited acceptability assumption. If market participants have a commonality of preferences, the assumption may lead to a large proportion of firms that are unacceptable by most workers as the market size increases, so most firms would remain unmatched in stable matchings. Figure 1 presents this phenomenon with simulations of one-to-one matching in which each worker considers only up to thirty of the most preferred firms acceptable.[10] We generate utilities for firms as $U_{f,w} = \lambda\, C_w + (1-\lambda)\, \zeta_{f,w}$, and similarly for workers. The value of each component is drawn from the uniform distribution over $[0,1]$, and $\lambda \in (0,1)$ represents the degree of

[9] A recent paper by Ashlagi, Kanoria, and Leshno (2013) also studies a closely related setup, but the following discussion on the limited acceptability assumption is irrelevant to their paper. The paper focuses on the effect of unequal number of firms and workers on stable matchings and incentive compatibility of stable matching mechanisms.

[10] These simulations are based on preferences generated by our own model, rather than the previous studies' model. We observe similar effects of the limited acceptability assumption simulations based on the previous studies' model as well (see Section F in the supplemental appendix).

commonality of preferences. Each graph in the figure depicts the proportion of firms (or workers) unmatched in stable matchings averaged over ten repetitions.[11] With commonality of preferences, the proportion of unmatched agents in stable matchings increases as the market size increases.

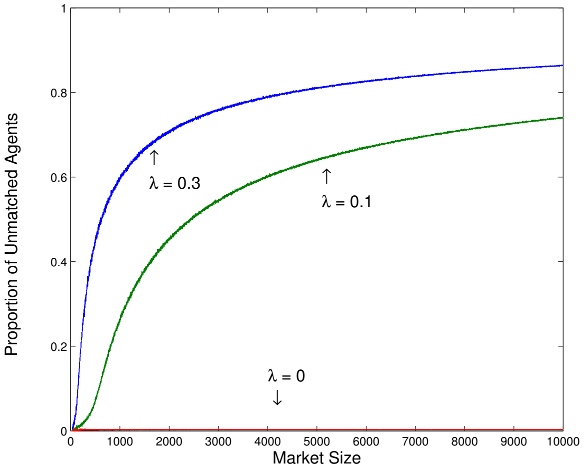

Figure 1: Proportion of agents unmatched in stable matchings.

The main advantage of our approach is that non-manipulability of stable matching mechanisms is a pure property of market size, without the limited acceptability assumption. The cardinal utilities allow us to consider a weaker notion of incentive compatibility, assuming that an agent manipulates only when it is significantly beneficial.

The large market approach on non-manipulability of stable mechanisms is not limited to the standard matching model. Ashlagi, Braverman, and Hassidim (2011) and Kojima, Pathak, and Roth (2013), 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 size increases. With some additional regularity conditions, when a mechanism produces a stable matching with high probability, it is an approximate equilibrium for all market participants to submit their true preferences.

We emphasize that not every large market model leads to non-manipulability of stable

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

mechanisms. As a simple example, consider a large matching market which is a replica of a small market with multiple stable matchings (Alkan and Gale, 2003; Bodoh-Creed, 2013; Azevedo and Hatfield, 2013).[12] Each firm or worker is considered as a type of firm or worker, and each firm-worker pair receives utilities according to the pair of types. In this replica economy, every copy of an agent would be matched to a copy of whom the original agent was matched. Thus the gap between utilities from firm-optimal and worker-optimal stable matchings remains the same regardless of how many times we replicate the finite market. The incentives for manipulation in a large replica market are essentially the same as the incentives in the original market.

Another large market model which does not lead to non-manipulability of stable mechanisms considers a finite number of firms that are matched with a continuum of workers (Azevedo and Leshno, 2013). In this setup, each firm has non-vanishing market power, so it can manipulate stable mechanisms, especially by misrepresenting capacities. Accordingly, Azevedo (2014) studies the welfare effects of firms' manipulations when a firm pays its employees equally (uniform wages) or when a firm pays different wages to different workers (personalized wages). In our paper, we do not take Azevedo-Leshno's approach because the NRMP has large numbers of participants on both sides of the market.

The literature on non-manipulability in a large market is huge, and contains both classical papers as well as more recent contributions. Among many others, Roberts and Postlewaite (1976) and Jackson (1992) study general equilibrium models, and Gresik and Satterthwaite (1989) and Rustichini, Satterthwaite, and Williams (1994) study double auctions. In the problems of allocating indivisible objects without monetary transfer, Kojima and Manea (2010) and Che and Kojima (2010) study incentives in the probabilistic serial mechanism; Liu and Pycia (2013) show asymptotic equivalence of all sensible, symmetric, strategy-proof, and ordinal efficient mechanisms, and Hashimoto (2013) proposes a generalized random priority mechanism, which approximates any incentive compatible mechanism. In general mechanism design, Kearns, Pai, Roth, and Ullman (2012) considers a mechanism with agents concerned about keeping their types private vis-a-vis other market participants; Azevedo and Budish (2013) study a notion of approximate strategy-proofness.

The rest of this paper is organized as follows. In Section 2, we introduce our model–a

[12] Bodoh-Creed (2013) considers considers a large, but finite, number of replications, whereas Alkan and Gale (2003) and Azevedo and Hatfield (2013) consider the case of a continuum of replications: i.e., a continuum of agents with a finite number of types. These papers consider much more general models than a simple replica economy, and the main questions differ from incentive compatibility. See also, e.g. Gretsky, Ostroy, and Zame (1992) and Azevedo, Weyl, and White (2013), for assignment models with a continuum of agents.

sequence of one-to-one matching markets with random utilities. In Section 3, we state the main theorem informally and then formally, and find a truth-telling equilibrium behavior. In Section 4, we illustrate the intuition of the proof using a random bipartite graph model. We extend our model to incomplete information in Section 5 and to many-to-one matching in Section E. All detailed proofs are relegated to the supplementary appendix.

2 Model

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

2.1 Standard One-to-one 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 $\mu \succeq_W \mu_F$.

We let M denote a function $\succ \longmapsto M(\succ)$ from the set of all preference profiles to the set of all matchings. We call the function M a **matching mechanism** and state that a mechanism M is **stable** if $M(\succ)$ is a stable matching with respect to preference profile $\succ$. We also let M_F and M_W denote firm-optimal and worker-optimal stable matching mechanisms. A matching mechanism induces a game in which each agent states a preference list. For every preference profile $\succ$ and an agent $i \in F \cup W$, if

$$M(\succ) \ \succeq_i \ M(\succ'_i, \succ_{-i}) \quad \text{for every } \ \succ'_i$$

then we say that it is a dominant strategy for agent i to state her true preference list. A mechanism is called **strategy-proof** if it is a dominant strategy for every agent to state her true preference list.

We illustrate how an agent can manipulate a mechanism. Table 1 lists the preferences of firms and workers. For instance, firm 1 most prefers worker 3, followed by worker 1 and worker 2. All firms and workers are mutually acceptable. 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: An example of a two-sided matching market.

Suppose a stable matching mechanism produces the second stable matching marked by $[\cdot]$. In that case, suppose firm 1 misrepresents its 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 $\langle\cdot\rangle$. Firm 1 is better off because firm 1 is matched with worker 1 rather than worker 2.

2.2 Random Utilities

In order to measure incentives to manipulate a stable matching mechanism, we assume that preferences are induced by underlying utility functions. Moreover, to measure the likelihood of manipulations, 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

$$\begin{aligned} U_{f,w} &= U(C_w, \zeta_{f,w}) \quad \text{and} \\ V_{f,w} &= V(C_f, \eta_{f,w}), \end{aligned}$$

where $U(.,.)$ and $V(.,.)$ are continuous and strictly increasing functions from $\mathbb{R}_+^2$ to $\mathbb{R}_+$.

We call C_w and C_f *common values* and $\zeta_{f,w}$ and $\eta_{f,w}$ *independent private values.* Common values are defined as random vectors

$$C_W := \langle C_w \rangle_{w \in W} \quad \text{and} \quad C_F := \langle C_f \rangle_{f \in F}.$$

Each C_w and C_f is drawn from distributions with positive density functions and 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 $\eta_{f,w}$ is randomly drawn from continuous distributions with bounded supports in $\mathbb{R}_+$.[13] We assume that the utility of remaining unmatched is equal to 0 so that all firms and workers are mutually acceptable to each other.[14]

[13] We can allow two private values to be an i.i.d sample from a joint distribution, so that private values for each pair can be positively correlated. For example, parents may want to send their children to schools in which siblings are already attending from previous years. Also those students may get priority in the schools' preference list..

[14] The main result holds without the assumptions of balanced market and all acceptable firms and workers. The set of unmatched firms and workers is the same for all stable matchings (McVitie and Wilson (1970)), so

Example (Linear utilities)**.** *For each pair (f, w), utilities are defined as*

$$\begin{aligned} U_{f,w} &= \lambda C_w + (1-\lambda)\zeta_{f,w}, \quad \textit{and} \\ V_{f,w} &= \lambda C_f + (1-\lambda)\eta_{f,w}. \end{aligned}$$

where $\lambda \in (0, 1)$. All four components ($C_w, C_f, \zeta_{f,w}, \eta_{f,w}$) have the uniform distribution over $[0, 1]$.

The common-value component and the private-value component introduce vertical preferences and horizontal preferences. The vertical preferences induce commonality of preferences as firms with high common values tend to be ranked highly by workers, and vice versa. In practice, commonality is prevalent. In the NRMP, medical school graduates often consider the *US News and World Report* as a guidance for prestigious hospitals, and all hospitals want to hire candidates with strong recommendation letters. At the same time, participants have idiosyncratic horizontal preference such as geographical preferences.

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, given a realization $\langle F, W, u, v\rangle$, for each $f \in F$ we can derive a strict preference list $\succ_f$ as

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

if and only if

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

Similarly, we derive $\succ_w$ for each $w \in W$.

We study the 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 doing so is not confusing.

Along with our main model, we also consider two extreme cases: pure common-value model ($U_{f,w} = C_w$ and $V_{f,w} = C_f$) and pure private value model ($U_{f,w} = \zeta_{f,w}$ and $V_{f,w} = \eta_{f,w}$). In the pure common-value model, all firms have an identical preference list for workers, and all workers have an identical preference list for firms. In the pure private-value model, all

participants who remain unmatched have no difference in utilities from all stable matchings, and participants who are matched in stable matchings will have small differences in utilities. Relaxing the assumptions does not affect the equilibrium analysis either (cf., Footnote 17).

utilities are independent and identically distributed (i.i.d.), so a firm's ordering of workers is equally likely to be any permutation of the n workers. Similarly, a worker's ordering of firms is equally likely to be any permutation of the n firms.

In particular, the pure private-value model is an interesting case to study as it has no commonality of preferences. Strong 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). Holzman and Samet (2013) also proposes commonality as a source establishing a small core: the small difference between the stable matchings favorable to firms and to workers. With the pure private-value model, we can highlight that non-manipulability of stable matching mechanisms is a property that can be solely derived by market size. Commonality may contribute to, but is not necessary for the non-manipulability.

3 Main Results

We give two statements of the main theorem: one informal and one formal. Later, we find an equilibrium behavior of the game induced by a stable matching mechanism in which most agents reveal their true preferences.

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 over all possible stable matchings.

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

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 size increases.*

No stable matching mechanism is strategy-proof (Roth, 1982). For instance, when the worker-optimal matching mechanism (e.g., the 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 can be a firm that may improve its position by misrepresenting its preference list. Noting that a matching mechanism is defined over all possible preference profiles, we may hope that a stable matching mechanism is not manipulable in most 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 conditions on a preference profile to yield a unique stable matching seem very restrictive (Eeckhout, 2000; Clark, 2006).

However, the gain from manipulation is bounded. Whenever a stable matching mechanism is applied, the best a firm can achieve is matching with the firm-optimal stable matching partner with respect to true preferences; likewise, the best a worker can achieve is matching with the worker-optimal stable matching partner (Demange, Gale, and Sotomayor, 1987). Consider the example with three firms and three workers in Table 1. 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. As stable matching mechanisms guarantee market participants to be matched *at best* with their most favorite stable matching partners, the gain from manipulation is bounded by the difference between utilities from the most and the least preferred stable matching partners.

As a result, the main theorem implies that agents in a large market are most likely to have only a vanishingly small 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, most market participants find no incentive to manipulate a stable matching mechanism.

Formal Statement

Given a market instance $\langle F, W, u, v\rangle$ and a matching μ, we let u_f^μ and v_w^μ denote utilities from the matching outcome: i.e., $u_f^\mu := u_{f,\mu(f)}$ and $v_w^\mu := v_{\mu(w),w}$. 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:

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

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 any $\epsilon, \delta, \theta > 0$, there exists N such that*

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

As $\frac{|A_F(\epsilon;U,V)|}{n}$ is bounded above by 1 with probability 1, we can rewrite Theorem 1 as the convergence in mean, which leads to the informal statement of Theorem 1 (see Remark A.1 in the supplemental appendix).

In the case of linear utilities the order of convergence in Theorem 1 is $o(e^{-n^c})$ with any constant $c \in (0, 1/2)$.[15] We relegate formal analysis on the order of convergence to Section C in the supplemental appendix.

3.2 Equilibrium Analysis

Previously, we showed that most agents have no significant incentive to manipulate a stable matching mechanism as the market size increases. However, the result requires the condition that all other market 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 the 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 consider an ϵ-Nash equilibrium, in which agents are approximately best responding to other agents' strategies such that no one can gain more than ϵ by switching to an alternative strategy. We first state this finding as a theorem, and then describe the appealing aspects of the equilibrium behavior and the intuition behind the proof.

Theorem 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 theorem is based on simple equilibrium behavior. Most agents simply reveal their true preferences. Agents misrepresenting their preferences use *truncation strategies*: an agent

[15] Given two sequences $\langle x_n \rangle_{n=1}^{\infty}$ and $\langle y_n \rangle_{n=1}^{\infty}$, we denote by $x_n = o(y_n)$ if $x_n / y_n \to 0$, as $n \to \infty$.

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.

We find an equilibrium in which all agents play truncation strategies against others' truncation strategies. In the equilibrium, agents who played truncation strategies are playing best responses among all possible strategies. Therefore, the equilibrium with the restriction of truncation strategies is also an equilibrium without the restriction.[16]

To derive an equilibrium behavior, we use a key property of truncation strategies: truncations only reduce the gap in utility from different stable matchings. Let $\succ$ be a true preference profile and $\succ'$ differ from $\succ$ in that some coalition of firms and workers profitably misstate their preferences using truncations. If there exists a matching μ stable under $\succ$ remaining individually rational under $\succ'$, then μ is indeed stable under $\succ'$ because no blocking pair has been generated by truncations. Thus, all participants are matched in stable matchings under $\succ'$ because the set of matched agents is the same for all stable matchings (McVitie and Wilson, 1970).[17] Then, any stable matching μ' under $\succ'$ is also stable under $\succ$ since we do not create any blocking pair by extending preferences from $\succ'$ to $\succ$. Since all stable matchings under $\succ'$ are also stable under $\succ$, the gap in utility from different stable matchings is reduced by truncations.

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. 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 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

[16] Truncation strategies make it easy to construct an equilibrium because an agent's truncation reduces other agents' gap in utilities from stable matchings. It is difficult to construct an equilibrium with non-truncation strategies as an agent's manipulation may increase the gap in utility from stable matchings for other market participants (Appendix G)

[17] We have implicitly used the property 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 to impose an additional condition that agents truncate only when truncations are strictly profitable. In particular, unmatched agents in stable matchings under $\succ$ will remain unmatched when she truncates her preference list. If unmatched agents do not truncate their preferences, we have the same result: all stable matchings under $\succ'$ are stable under $\succ$. The proof is easy to derive, so we omit it here.

utilities from stable matchings will have even less difference in utilities from stable matchings under the announced preferences. Thus, these participants have no significant incentive to respond to others' truncations, thereby submitting their true preferences. Lastly, Theorem 1 guarantees that most participants reveal their true preferences.

4 Intuition Behind the Proof of Theorem 1

We first consider the pure common-value model ($U_{f,w} = C_w$ and $V_{f,w} = C_f$). In this model, 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 forms a unique stable matching.

The Theorem 1 holds also for the pure private-value model. 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 (1989) shows that the sum of the rank numbers of firms' worker-optimal stable matching partners is asymptotically equal to $n^2 \log^{-1} n$.[18] Then, the rank number of each firm's worker-optimal stable matching partner 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. Since the private values are randomly drawn from distributions with bounded supports, even the worst stable matching partners give utilities asymptotically close to the upper bound. Therefore, all stable matchings yield very similar utilities. We relegate a detailed proof to Section B.2 in the supplemental appendix.[19]

For our general model with both common values and private values, firms and workers in a large market match assortatively in the common-value dimension and manage to get very good matches in the private-value dimension. Thus, the utility of firm f in any stable

[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 the pure private-value model.

[19] We can also observe from Claim 1 in Compte and Jehiel (2008) that utilities from the best stable matching become close to 1 as the market size increases.

matching is close to

$$U(\text{common value of a worker in the same position as } f, \\ \text{maximum of the support of the workers' private values}).$$

Although the main theorem has this very simple intuition, the proof is not straightforward from the proofs of the pure common-value model and of the pure private-value model. Now, common values and private values are entangled in a single utility structure.

Basically, we want to count participants whose utilities from all stable matchings are close to the utility levels specified in the above equation. We therefore need a counting method 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 their 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. We describe the techniques in greater depth in the following subsections and relegate a detailed proof to Section B.1 in the supplemental appendix.

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 $i \in U_1$ and $j \in U_2$, $(i, j) \in E$. In other words, a biclique is a complete bipartite subgraph of G. We say that a biclique is **balanced** if the size of U_1 is equal to the size of U_2 (i.e., $|U_1| = |U_2|$), and refer to a balanced biclique with the maximum size as **a maximal balanced biclique**.

Given a partitioned set $V_1 \cup V_2$, we randomly construct bipartite graphs so that each pair of nodes, one in V_1 and the other in V_2, is included in E independently with probability p. By abuse of notation, we denote a random bipartite graph by $G(V_1 \cup V_2, p)$.

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 $\beta_n =$*

$2\log n/\log\frac{1}{p}$. *If a maximal balanced biclique of this graph has size* $B \times B$*, then*

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

4.2 Intuition of the Proof

We demonstrate how to use techniques from the theory of random bipartite graphs in a simplified model, called *a three-tier market* with *linear utilities.*

In a three-tier market, firms and workers are partitioned into three tiers and endowed with tier-specific common values. That is, F is partitioned into F_1, F_2, and F_3, and W is partitioned into W_1, W_2, and W_3. Common values are uniquely determined by tiers such that

$$c_{F1} > c_{F2} > c_{F3} \quad \text{and} \quad c_{W1} > c_{W2} > c_{W3}.$$

Private values, $\zeta_{f,w}$ and $\eta_{f,w}$, are randomly drawn from uniform distributions over $[0,\bar{u}]$ and $[0,\bar{v}]$, respectively. 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 each other, then they receive utilities

$$U_{f,w} = c_{Wl} + \zeta_{f,w} \quad \text{and} \quad V_{f,w} = c_{Fk} + \eta_{f,w}.$$

We find an asymptotic lower bound on utilities that firms in tier 1 receive in a stable matching. The lower bound is defined as the level arbitrarily close to the maximal utility that a firm can achieve by matching with workers in tier 2: i.e., $u_{W2} + \bar{u} - \epsilon$. That is, firms in tier 1 achieve high levels of utility due to the existence of workers in tier 2. Although not necessarily being matched with workers in tier 2, firms in tier 1 would otherwise form blocking pairs with workers in tier 2. Formally, we define the set of firms in tier 1 that fail to achieve the specified utility level in the worker-optimal stable matching as

$$\bar{F} := \left\{ f \in F_1 \mid u_f^{\mu_W} \le c_{W2} + \bar{u} - \epsilon \right\},$$

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 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:

$$\zeta_{f,w} \le \bar{u} - \epsilon \quad \text{or} \quad \eta_{f,w} \le \bar{v} - (c_{W1} - c_{W2}).$$

We define the set of workers in tiers up to 2 matched with firms not in tier 1 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 2).

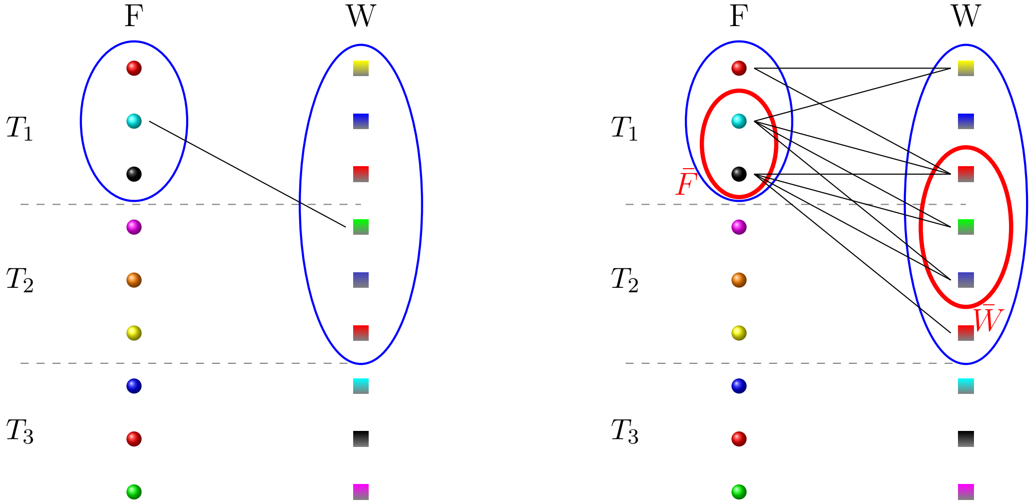

Figure 2: For each realized utility, we draw a bipartite graph with firms in tier 1 and workers in tiers up to 2 as the partitioned set of nodes (left). Firms in tier 1 receiving low utilities ($\bar{F}$) and workers in tiers up to 2 matched with firms not in tier 1 ($\bar{W}$) form a biclique (right).

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 $f \in \bar{F}$,

$$u_f^{\mu_W} \le c_{W2} + \bar{u} - \epsilon.$$

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

$$v_w^{\mu_W} \le c_{F2} + \bar{v}.$$

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,

$$\zeta_{f,w} > \bar{u} - \epsilon \quad \text{and} \quad \eta_{f,w} > \bar{v} - (c_{F1} - c_{F2}),$$

and therefore

$$u_{f,w} > c_{W2} + \bar{u} - \epsilon \quad \text{and} \quad v_{f,w} > c_{F1} + \bar{v} - (c_{F1} - c_{F2}) = c_{F2} + \bar{v}.$$

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

This construction of a bipartite graph fits into a random bipartite graph model. 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 a maximum balanced biclique has a size on the order of $\log(n)$ with a sequence of probabilities converging to 1 as n increases. 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 Theorem 3. Lastly, $E\left[\frac{|\bar{F}|}{n/3}\right] \to 0$ follows immediately from $\log(n)/n \to 0$.

To prove the main theorem (without tier structure or linear utility), we continue the proof as if we have a model with tiers assigned by common values. Suppose the common values of firms and workers are distributed uniformly over $[0, 1]$. For any $\bar{c} > 0$, take $\hat{c}$ and $\tilde{c}$ such that $0 < \hat{c} < \tilde{c} < \bar{c}$. We partition the unit interval into $[0, \hat{c})$, $[\hat{c}, \tilde{c})$, $[\tilde{c}, \bar{c})$, and $[\bar{c}, 1]$. Firms and workers are, in turn, grouped into tiers where agents in the same tier have common values in the same subinterval (firms and workers in tier 1 have the highest common values). As before, we find an asymptotic utility lower bound of firms in tier 1: i.e., firms with common values above $\bar{c}$. This time, because the common values are random, the number of firms and the number of workers in each tier are random. Moreover, agents in adjacent tiers may have arbitrarily close common values, so utilities of tier 1 firms will be bounded above by a level slightly lower than the maximum utility from workers in tier 3, rather than tier 2: i.e., $U(\tilde{c}, \bar{u}) - \epsilon$. As we choose $\hat{c}$ and $\tilde{c}$ arbitrarily close to $\bar{c}$, and ϵ arbitrarily small, the asymptotic lower bound becomes close to $U(\bar{c}, \bar{u})$: the maximal utility achievable by matching with a worker in the position of $\bar{c}$.

With a similar exercise, we find an asymptotic lower bound on utilities of workers for each level of common values. Then, it is likely that most workers with common values significantly higher than $\bar{c}$ match with firms with common values significantly higher than $\bar{c}$. This assortative feature induces an asymptotic upper bound on the utilities of firms with common values near $\bar{c}$. Therefore, we can find an asymptotic lower bound and an asymptotic upper bound, which are arbitrarily close to each other.

5 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. Expecting market participants to have this much information is obviously not realistic, but participants with limited information would be more likely to passively submit their true preferences. In this respect, we explore a model of market with incomplete information and obtain an equilibrium of a stronger truth-telling behavior than the ϵ-equilibrium under complete information.

Setup

We assume that common values are known to all participants, but private values are known to only the agent who receives the utilities. Let Π_f be a firm f's information about U and V:

$$\Pi_f = \langle C_w, U_{f,w} \rangle_{w \in W} \cup \langle C_{f'} \rangle_{f' \in F}.$$

We can extrapolate findings from the case of complete information to show the incentive compatibility of stable matchings under incomplete information. As before, we first state the theorem informally and then restate it later with formal expressions.

Theorem. *In a sequence of markets with incomplete information, for every $\epsilon > 0$, the expected proportion of firms (and workers) that have less than ϵ expected differences between utilities from μ_F and μ_W converges to one as the market size increases.*

Corollary. *In a sequence of markets with incomplete information, 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 size increases.*

The intuition behind this theorem is very simple: an expected value is an average of all realized values. The expected difference between utilities from firm-optimal and worker-optimal stable matchings is simply a convex combination of differences realized in all market instances. It is likely that most agents have insignificant differences (Theorem 1). As each agent is most likely to have an insignificant difference, the expected value of the difference would be negligible. We relegate the detailed proof to Appendix D.1.

There are two advantages to 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 in regards to the information structure. The intuition of showing the results with incomplete information through convex combinations remains valid regardless of the details of the information structure. Second, we can stress that the non-manipulability of stable matching mechanisms is a property of the two-sided matching market itself, rather than stemming from insufficient information to manipulate. 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 with this information will be small.

Formal Statement

Given a realization $\langle F, W, u, v\rangle$, we define $\Delta^E(f; u, v)$ as firm f's expected difference between utilities from firm-optimal and worker-optimal stable matchings, conditioned on π_f:

$$\Delta^E(f; u, v) := E\left[\Delta(f; U, V) \mid \pi_f\right].$$

For every $\epsilon > 0$, we have the set of firms each of which has the expected difference between utilities from stable matchings less than ϵ:

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

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 $\epsilon, \delta, \theta > 0$, there exists N such that*

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

Equilibrium analysis

For the pure private-value model (i.e., $U_{f,w} = \zeta_{f,w}$), we find an ϵ-Bayesian-Nash equilibrium in which, with probabilities converging to one, every participant reveals her true preference list. For the general model, however, we have not been able to construct a truth-telling equilibrium.

Theorem 5. *Suppose a stable matching mechanism is applied to pure private-value markets with incomplete information. For any $\epsilon, \delta > 0$, there exists N such that a market of size $n > N$ has an ϵ-Bayesian-Nash equilibrium in which with probability at least $(1-\delta)$ all agents reveal their true preference lists.*

The ϵ-Bayesian-Nash equilibrium is based on a result stronger than Theorem 4, which we could prove for the pure private-value model: Asymptotically, *all participants have insignificant expected difference between utilities from firm-optimal and worker-optimal stable matchings.* Consider a strategy profile in which participants, who have small expected differences between utilities from firm-optimal and worker-optimal stable matchings, submit their true preference lists. If an agent has a large expected difference between the utilities, the agent is assumed to play a best-response to all other agents' truth-telling strategies.

In this strategy profile, participants are approximately best-responding to other agents' strategies such that no one can gain more than ϵ by switching to an alternative strategy. Truth-telling is an approximate best-response for an agent with a small expected difference between the utilities from stable matchings, as most likely all other agents will submit their true preferences and the gain by manipulation is bounded above by the small expected difference in utilities from stable matchings. Conversely, if the agent has a large expected difference between the utilities from stable matchings, a best-response to all other agents' truth-telling is an approximate best-response to other agents' strategies, it is most likely that all agents will tell the truth. We relegate a formal proof of Theorem 5 to Appendix D.2.

6 Conclusion

This paper demonstrates an asymptotic similarity of stable matchings as the number of market participants increases. Our measure of similarity is based on utilities by which ordinal preferences are determined. Since 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 over all possible stable matchings converges to one as the market size increases. In one-to-one matching, the result implies that the expected proportion of agents who have significant incentives to manipulate stable matching mechanisms vanishes in large markets. This is because the gain from manipulation is bounded above by the gap between utilities from firm-optimal and worker-optimal stable matchings. Furthermore, 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.

References

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.

Alkan, A., and D. Gale (2003): "Stable schedule matching under revealed preference," *Journal of Economic Theory*, 112(2), 289–306.

Ashlagi, I., M. Braverman, and A. Hassidim (2011): "Stability in Large Matching Markets with Complementarities," *mimeo.*

Ashlagi, I., D. Gamarnik, M. A. Rees, and A. E. Roth (2012): "The need for (long) chains in kidney exchange," *mimeo.*

Ashlagi, I., Y. Kanoria, and J. D. Leshno (2013): "Unbalanced random matching markets.," *mimeo.*, pp. 27–28.

Ashlagi, I., and A. E. Roth (2013): "Free riding and participation in large scale, multi-hospital kidney exchange," *Theoretical Economics*, forthcoming.

Azevedo, E. (2014): "Imperfect Competition in Two-Sided Matching Markets," *Games and Economic Behavior*, forthcoming.

Azevedo, E., and E. Budish (2013): "Strategyproofness in the large," *Mimeo.*

Azevedo, E. M., and J. W. Hatfield (2013): "Complementarity and Multidimensional Heterogeneity in Matching Markets," *mimeo.*

Azevedo, E. M., and J. D. Leshno (2013): "A supply and demand framework for two-sided matching markets," *mimeo.*

Azevedo, E. M., E. G. Weyl, and A. White (2013): "Walrasian equilibrium in large, quasilinear markets," *Theoretical Economics*, 8(2), 281–290.

Bodoh-Creed, A. L. (2013): "Large Matching Markets: Risk, Unraveling, and Incentive Compatibility," Discussion paper, Mimeo, Cornell University.

Che, Y., and F. Kojima (2010): "Asymptotic equivalence of probabilistic serial and random priority mechanisms," *Econometrica*, 78(5), 1625–1672.

Clark, S. (2006): "The uniqueness of stable matchings," *Contributions in Theoretical Economics*, 6(1), 1–28.

Compte, O., and P. Jehiel (2008): "Voluntary participation and reassignment in two-sided matching," *mimeo.*

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.

Dubins, L., and D. Freedman (1981): "Machiavelli and the Gale-Shapley algorithm," *American Mathematical Monthly*, 88(7), 485–494.

Eeckhout, J. (2000): "On the uniqueness of stable marriage matchings," *Economics Letters*, 69(1), 1–8.

Feldin, A. (2003): *Core Convergence in Two-Sided Matching Markets.* Springer.

Gale, D., and L. Shapley (1962): "College admissions and the stability of marriage," *American Mathematical Monthly*, 69(1), 9–15.

Gresik, T. A., and M. A. Satterthwaite (1989): "The rate at which a simple market converges to efficiency as the number of traders increases: An asymptotic result for optimal trading mechanisms," *Journal of Economic Theory*, 48(1), 304–332.

Gretsky, N. E., J. M. Ostroy, and W. R. Zame (1992): "The nonatomic assignment model," *Economic Theory*, 2(1), 103–127.

Hashimoto, T. (2013): "The Generalized Random Priority Mechanism with Budgets," *mimeo.*

Holzman, R., and D. Samet (2013): “Matching of like rank and the size of the core in the marriage problem,” .

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.

Jackson, M. (1992): “Incentive compatibility and competitive allocations,” *Economics Letters*, 40(3), 299–302.

Kearns, M., M. Pai, A. Roth, and J. Ullman (2012): “Mechanism Design in Large Games: Incentives and Privacy,” *arXiv preprint arXiv:1207.4084*.

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. A. Pathak, and A. E. Roth (2013): “Matching with Couples: Stability and Incentives in Large Markets*,” *The Quarterly Journal of Economics*, 128(4), 1585–1632.

Liu, Q., and M. Pycia (2013): “Ordinal Efficiency, Fairness, and Incentives in Large Markets,” *mimeo.*

McKinney, C., M. Niederle, and A. Roth (2005): “The collapse of a medical clearinghouse (and why such failures are rare),” *American Economic Review*, 95(3), 878–889.

McVitie, D., and L. Wilson (1970): “Stable marriage assignment for unequal sets,” *BIT Numerical Mathematics*, 10(3), 295–309.

Pittel, B. (1989): “The Average Number of Stable Matchings,” *SIAM Journal on Discrete Mathematics*, 2(4), 530–549.

Roberts, D., and A. Postlewaite (1976): “The incentives for price-taking behavior in large exchange economies,” *Econometrica*, pp. 115–127.

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," *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.

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 M. SOTOMAYOR (1990): *Two-sided matching*. Cambridge University Press.

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," *American Economic Review*, 84(4), 992–1044.

RUSTICHINI, A., M. A. SATTERTHWAITE, AND S. R. WILLIAMS (1994): "Convergence to efficiency in a simple market with incomplete information," *Econometrica*, pp. 1041–1063.

SÖNMEZ, T. (1999): "Strategy-proofness and Essentially Single-valued Cores," *Econometrica*, 67(3), 677–689.