Incentive Compatibility of Large Centralized Matching Markets

SangMok Lee*

August 22, 2016

Abstract

We study the manipulability of stable matching mechanisms. To quantify incentives to manipulate stable mechanisms, we consider markets with random cardinal utilities, which induce ordinal preferences over match partners. We show that most agents in large matching markets are close to being indifferent overall stable matchings. In one-to-one matching, the utility gain by manipulating a stable mechanism does not exceed the gap between utilities from the best and worst stable partners. Thus, most agents in a large market would not have significant incentives to manipulate stable mechanisms. The incentive compatibility extends to many-to-one matching when agents employ truncation strategies and capacity manipulations in a Gale-Shapley mechanism.

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 part of my dissertation at Caltech. I am especially grateful to Federico Echenique, Matt Shum, and Leeat Yariv for encouragement and guidance, and Jan Tilly for excellent research assistance. I thank the editor, Philipp Kircher, and three anonymous referees for thoughtful and detailed comments. For very helpful comments, 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

In the practice of market design, *stable matching mechanisms* are often used in two-sided markets. They are used to match agents of one kind (firms) with agents of another kind (workers), most famous being the National Resident Matching Program (NRMP).[1] We ask why stable matching mechanisms have been so successful, despite that they are not strategy-proof. In particular, we analyze whether the presence of a large number of participants can mitigate the incentives to misrepresent preferences.

The concept of *stability* has been considered of central importance. A matching is regarded as stable if no agent would rather remain unmatched than matching to her current partner, and if no pair of agents prefers each other to their current partners. A stable mechanism takes preference reports from participants and produces a stable matching on reported preferences. Most stable mechanisms have been successful (Roth and Xing, 1994; Roth, 2002; McKinney et al., 2005). For the NRMP, the Association of American Medical Colleges even required stability as a key property of their matching algorithm (Roth, 1984).

However, *no stable mechanism is strategy-proof* (Roth, 1982). A market participant can achieve a better match by misreporting her preferences, either by changing the order of her preference list or by announcing that some agents are unacceptable.[2] The mechanism in the NRMP may be manipulated by participants, thereby not implementing the intended matching. Also, each participant's decision may become difficult as she must best respond to other agents' strategic manipulations.

Reconciling the success and limitation of stable mechanisms is important. The most well-known findings are by Roth and Peranson (1999), Immorlica and Mahdian (2005), and Kojima and Pathak (2009).[3] These studies consider a large market approach motivated by more than 4,000 residency programs and 25,000 doctors in the NRMP.[4] They consider the worker-optimal stable mechanism, which makes truth-telling a dominant strategy for workers. The expected proportion of firms that have any incentive to misrepresent their preferences converges to zero as the market increases.

[1] See Roth and Peranson (1999) for other various professional labor markets implementing stable matching mechanisms.

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

[3] See Kojima (2015) for a summary of the relationship between the previous three papers and the current paper. Also, Feldin (2003) takes a similar approach with assuming zero commonalities of preferences.

[4] The recent data on residency and fellowship matching are available on `http://www.nrmp.org/match-data/main-residency-match-data/` and `http://www.nrmp.org/match-data/fellowship-match-data/`.

We believe that the previous papers rightly point out a large market effect on a stable mechanism. However, the previous explanation of the large market effect seems inconsistent with the markets in practice. The previous explanation relies on the assumption of *short preference lists*: agents on one side (e.g., workers) consider only a vanishingly small proportion of agents on the other side acceptable.[5] In the NRMP, participants often have very similar preferences (Agarwal, 2015). Medical school graduates prefer top-tier hospitals, and hospitals prefer candidates with strong recommendation letters and high medical exam scores. If agents have short *and* similar preferences, only a small number of agents on one side would be accepted by agents on the other side. In this situation, many agents would remain unmatched anyhow, so they would not manipulate a stable mechanism (See Section **??** in the online appendix for simulation results).[6] This explanation of the large market effect seems inconsistent with only 4.3% vacant positions in the NRMP 2015.

At first glance, the assumption of short preference lists may appear to be harmless. In the NRMP, submitted preferences tend to have few commonalities so that short preference lists do not leave many agents unmatched. Nevertheless, we believe that the short preference lists with few commonalities should not be an assumption; rather, such observations must be an implication of the non-manipulability. In practice, residency programs interview doctors before submitting preferences. Since interviews are costly, the programs interview only selective doctors and place them on their rank order lists. The selection takes into account how likely a doctor is achievable with a possible manipulation of the mechanism. As such, any restriction on preferences in a model, motivated by observed preferences in practice, may assume to some extent the non-manipulability of the mechanism.

We give a very different explanation of the large market effect on stable mechanisms. The key modeling strategy is to consider *random cardinal utilities*, by which ordinal preferences are determined. We use cardinal utilities to quantify the incentives to manipulate stable mechanisms. We ask how likely a market has a large proportion of agents with significant incentives to manipulate. In one-to-one matching markets, we find that most agents in large markets are close to indifferent between stable matchings. It is known in the literature that each agent can gain utilities from manipulation up to the maximum wedge between utilities from stable matchings. Thus, the maximum utility gain from manipulation vanishes as the market increases. The results partially extend to many-to-one matching markets. We show

[5]For instance, each worker considers only the 30 most preferred firms to be acceptable. The limit can grow, but the proportion of acceptable firms must vanish.

[6]By Rural Hospital Theorem and Demange et al. (1987), an unmatched agent in a stable matching cannot effectively manipulate a stable mechanism.

that firms have vanishing incentives to manipulate the worker-optimal stable mechanism when they are assumed to play truncation strategies and capacity misrepresentations.[7]

The baseline model of one-to-one matching 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 continuous distributions with bounded supports in $\mathbb{R}_+$:

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

Common values, C_f and C_w, represent 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). While agents in a large market typically have multiple stable partners, most agents are close to being indifferent between them (Theorem 1). In a large market, the utility of firm f in any stable matching becomes

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

That is, firms and workers match assortatively in the common value dimension. An agent gets as high a vertical match as possible given her vertical quality. At the same time, agents find very good matches in the private value dimension.

As a consequence of the main result, we identify an ϵ-Nash equilibrium behavior in which most participants report their true preferences (Theorem 2).[8] When a stable mechanism is applied to a one-to-one matching market, an agent can achieve from manipulation up to her most preferred stable partner on true preferences (Demange et al., 1987). As such, our main finding implies that, when all agents tell the truth, the expected proportion of agents with significant incentives to manipulate vanishes. To find an ϵ-Nash equilibrium behavior, we let agents with significant incentives manipulate the mechanism in a way that does not increase other agents' incentives. The rest of the agents continue to have insignificant incentives to misreport preferences.

We extend our results to large many-to-one matching markets. The assortative feature

[7] In a truncation strategy, a firm submits a preference list of the first few workers in the same order as the true preference list.

[8] Under an ϵ-Nash equilibrium, agents approximate the best response to other agents' strategies. No one can gain more than ϵ by switching to an alternative strategy.

of stable matchings extends to large many-to-one matching markets (Theorem 3). We find an ϵ-Nash equilibrium in a restricted environment in which the worker-optimal stable mechanism is applied, and firms only play truncation strategies and capacity misrepresentations (Theorem 4).

We believe in showing the *right* way to characterize the large market effect with the following three reasons.

First, our characterization of a large market effect is intuitively compelling. In the NRMP 2015, about half of the matched applicants matched to their top choices, and more than 85% of matched applicants matched to their top four choices. It is very unlikely that such a large proportion of workers in a labor market match to their truly most preferred firms. In practice, agents have non-zero costs of doing interviews and submitting applications. As such, agents interview only selective potential partners with judgments on the attainability of a match. The assortative feature of stable matching makes such judgments in the common value dimension easy.

Second, our model is consistent with two prominent features of the NRMP: a strong commonality of preferences and non-increasing proportion of vacant positions. A recent empirical analysis by Agarwal (2015), based on anecdotal evidence, even assumes a complete commonality of preferences for residency programs. The estimated doctors' preferences also show a clear dependency on commonly observable characteristics of residency programs. The short preference lists would lead to an increasing proportion of vacant positions. However, in the NRMP 2015, only 4.3% of residency positions are vacant. The proportion is substantially lower than the 12.2% of vacant positions in fellowship matching, even though the fellowship matching markets for each sub-specialty tend to be much smaller.

Last, the speeds of convergence in our results are fast. In a market the size of the NRMP, our approach gives more compelling results than the previous studies. Consider linear utilities, $U_{f,w} = \lambda C_w + (1-\lambda)\zeta_{f,w}$, in which $C_w, \zeta_{f,w} \sim U[0,1]$. For any arbitrary $\epsilon, \theta > 0$, the chance of having a market realization in which a large ($> \theta$) proportion of agents has significant ($> \epsilon$) incentives to manipulate a stable mechanism approaches quickly to zero. The probability vanishes with speed $o(e^{-n^{1/2}})$, which is faster than $O(1/n)$, the corresponding speed in Immorlica and Mahdian (2005) and Kojima and Pathak (2009). (See Remark 1 in the online appendix of Kojima and Pathak (2009)).[9] In Section 5, we also simulate how fast the maximum expected gain from manipulation vanishes. For the size

[9]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 there exists a constant M such that $|x_n| \leq M|y_n|$. We denote by $x_n = o(y_n)$ if $x_n/y_n \to 0$.

26,000, the maximum expected gain is about 0.0034, 0.0013, and 0.0004 for $\lambda = 1/4, 1/2,$ and $3/4$, respectively. In many-to-one markets, in which each firm hires up to 8 workers, the maximum gain for each firm is less than 0.014, for all three levels of commonality.

The large market approach has been extended to richer models. Ashlagi et al. (2014) and Kojima et al. (2013), for instance, develop 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). It turns out that large markets with couples are most likely to have stable matchings. If a mechanism produces a stable matching (whenever one exists), it is an approximate equilibrium for all agents to submit their true preferences.

There is an extensive large market literature beyond two-sided matching. Among many others, Roberts and Postlewaite (1976) and Jackson (1992) study general equilibrium models, and Gresik and Satterthwaite (1989) and Rustichini et al. (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 the asymptotic equivalence of all symmetric, strategy-proof, and ordinal efficient mechanisms. Hashimoto (2013) proposes a generalized random priority mechanism, which approximates any incentive compatible mechanism. In mechanism design, Kearns et al. (2014) 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 a baseline model of one-to-one matching markets with random utilities. We show that all stable matchings tend to be assortative and find a truth-telling equilibrium behavior. In Section 3, we extend our model to many-to-one matching. In Section 4, we illustrate the intuition of the proof using a random bipartite graph model. Section 5 includes the results on the speed of convergence. An extension to incomplete information is found in Section 6. We relegate all detailed proofs to the online appendix.

2 One-to-one Matching

2.1 Setup

We first build a setup based on the standard one-to-one matching model. We introduce latent utilities, which in turn generate ordinal preferences.

2.1.1 One-to-one Matching

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

$$\succ_f = w_1, w_2, w_3, \emptyset$$

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 $w_1 \succ_f w_2$ to mean that f prefers w_1 to w_2. We call a worker w **acceptable** to f if w appears in the firm's preference list $\succ_f$; 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 $F \cup W$ to 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 $\mu(w) \in F$. 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 $f \succ_w \mu(w)$. 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 standard one-to-one matching market with strict preference lists has a firm-optimal stable matching μ_F and a worker-optimal stable matching μ_W (Gale and Shapley, 1962). 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 $\mu \succeq_W \mu_F$.

A **matching mechanism** M is a function $\succ \longmapsto M(\succ)$ from the set of all preference profiles to the set of all matchings. A mechanism M is **stable** if $M(\succ)$ is a stable matching with respect to preference profile $\succ$. We denote by M_F and M_W firm-optimal and worker-optimal stable matching mechanisms.

A mechanism is **strategy-proof** if it is a dominant strategy for every agent to state her

true preference list. That is, for every preference profile $\succ$ and agent $i \in F \cup W$,

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

No stable matching mechanism is strategy-proof (Roth, 1982). Whenever there is more than one stable matching, at least one agent can profitably misrepresent her preferences (Roth and Sotomayor, 1990) by switching potential partners in her preference list or announcing some acceptable partners unacceptable.[10] Even the worker-optimal matching mechanism (e.g., the worker-proposing Gale-Shapley algorithm), which makes it a dominant strategy for workers to state true preferences (Roth, 1982; Dubins and Freedman, 1981), may not rule out firms' incentives to misrepresent their preference lists.

2.1.2 Random Utilities

We assume that preferences are induced by underlying cardinal utilities that are drawn from underlying probability distributions. This approach allows us to measure incentives to manipulate a stable matching mechanism.

A random market is a tuple $\langle F, W, U, V \rangle$. $U = [U_{f,w}]$ and $V = [V_{f,w}]$ are two $n \times n$ random matrices representing utilities. 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 use u and v to denote realized matrices of U and V, respectively.

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 C_w and C_f are *common values*, $\zeta_{f,w}$ and $\eta_{f,w}$ are *independent private values*, and $U(.,.)$ and $V(.,.)$ are continuous and strictly increasing functions from $\mathbb{R}^2_+$ to $\mathbb{R}_+$.

Common values are two random vectors

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

in which each C_w and C_f is drawn from distributions with positive density functions and

[10] Indeed, the conditions on a preference profile to yield a unique stable matching seem very restrictive (Eeckhout, 2000; Clark, 2006), so most preference profiles admit agents with strategic incentives.

bounded supports in $\mathbb{R}_+$. Independent private values are two $n \times n$ random matrices

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

in which each $\zeta_{f,w}$ and $\eta_{f,w}$ is randomly drawn from continuous distributions with bounded supports in $\mathbb{R}_+$. We assume without loss of generality that all common values and private values are uniformly distributed over $[0,1]$.[11] We normalize the utility of remaining unmatched equal to 0 so that all firms and workers are mutually acceptable to each other.

Example 1 (Linear utilities)**.** *For each pair (f,w), utilities are defined as*

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

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

The common value component introduces vertical preferences. Firms with high common values tend to be ranked highly by workers, and vice versa, agents on each side of the market have a commonality of preferences. In practice, commonality is prevalent. In the NRMP, medical school graduates often consider the *US News and World Report* as a guide for prestigious hospitals, and all hospitals want to hire candidates with strong recommendation letters. The private-value component introduces idiosyncratic horizontal preferences.

All agents have distinct utilities with probability one. For each market realization $\langle F, W, u, v\rangle$, each firm $f \in F$ is associated with a strict preference list

$$\succ_f = w, w', \ldots, w'' \quad \text{if and only if} \quad u_{f,w} > u_{f,w'} > \cdots > u_{f,w''}.$$

Similarly, each worker $w \in W$ is associated with a strict preference list $\succ_w$.

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.

[11] Whatever distribution we assume, there always exists a change of variables that delivers the uniform distribution, and we can transform utility functions monotonically.

To get some intuitions of the main results, we sometimes consider two extreme cases, which are not included in our main model. The pure common value model assumes $U_{f,w} = C_w$ and $V_{f,w} = C_f$. All firms (and workers) have an identical preference list for workers (respectively, for firms). The pure private value model assumes $U_{f,w} = \zeta_{f,w}$ and $V_{f,w} = \eta_{f,w}$. A firm's ordering of workers is equally likely to be any permutation of the n workers.

2.2 Results

We show incentive compatibility of all stable matching mechanisms in the simplified model. It turns out that agents in stable matchings of a large market match assortatively in the common value dimension and manage to obtain favorable matches in the private value dimension. An important implication is that each agent receives similar payoffs from all stable matchings. In the case of one-to-one matching, agents with near indifference between stable matchings would not have significant incentives to misreport their preferences to stable mechanisms (Demange et al. (1987)).[12]

2.2.1 Assortative Feature of Stable Matchings

We show that agents in a stable matching of a large market match assortatively in the common value dimension and manage to obtain favorable matches in the private value dimension. Thus, the utility of firm f in any stable matching is close to

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

Take a firm $f \in F$ who submits a preference list to a stable matching mechanism in a market instance $\langle F, W, u, v\rangle$. For every $\epsilon > 0$, we define the set of firms whose utilities from all stable matchings are within ϵ difference from their reference utilities:

$$A_F(\epsilon; u, v) := \left\{ f \in F \mid U(c_f, 1) - \epsilon < U_f^{\mu_W} \leq U_f^{\mu_F} < U(c_f, 1) + \epsilon \right\}.$$

[12]While this approach is convenient to use for one-to-one matching, it is not applicable for many-to-one matchings, where a firm may become better off than even in the firm-optimal stable matching. For many-to-one matching, we will take a more technical approach and show the vanishing gains directly from the assortative feature, without resorting to Demange et al. (1987).

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

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|A_F(\epsilon; U, V)|}{n}\right] = 1.^{13,14}$$

To gain some intuition, let us 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. 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. One the other hand, in the pure private-value model ($U_{f,w} = \zeta_{f,w}$ and $V_{f,w} = \eta_{f,w}$), even the worst stable matching partners give utilities asymptotically close to the upper bound (Pittel (1989); Lee and Yariv (2014); Ashlagi et al. (Forthcoming)).[15]

One important implication of the assortative feature is that all stable matchings yield very similar utilities. 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. In the case of one-to-one matching, the gain from manipulating a stable mechanism is bounded by the difference between utilities from the most and the least preferred stable matching partners. The best a firm can achieve by manipulating a stable mechanism is the firm-optimal stable matching partner on true preferences; likewise, the best a worker can achieve is matching with the worker-optimal stable matching partner (Demange et al. (1987)). Therefore, most agents who are near indifferent between stable matchings would not have significant incentives to misreport their preferences.

[13] Alternatively, we can write the theorem as follows:
For any $\epsilon, \delta, \theta > 0$, there exists N such that

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

[14] We omit the corresponding definitions and the theorem for workers.

[15] Pittel (1989) does not consider utilities, but a model with pure random ordinal preferences. For each firm, let the most preferred worker be ranked 1, the next worker 2, and so on. The sum of the rank numbers of firms' worker-optimal stable matching partners is asymptotically equal to $n^2 \log^{-1} n$. The average rank number, with a normalization of division by the market size n, converges to 0. Compte and Jehiel (2008) also observe that firms' utilities from the firm-optimal stable matching become utilitarian-efficient as the market size increases. Ashlagi et al. (Forthcoming) show that unequal numbers of agents on two sides lead to near indifference between stable matchings even in small markets.

We emphasize that our main result requires enough randomness in preferences. To illustrate this idea, consider a large replica of a small market with multiple stable matchings (Alkan and Gale, 2003; Bodoh-Creed, 2013; Azevedo and Hatfield, 2013).[16] We consider each firm or worker as a type, and each firm-worker pair receives utilities according to the pair of types. The wedge between utilities from stable matchings remains the same regardless how many times we replicate the finite market. Also, it is necessary that the numbers of agents on both sides grow. Azevedo and Leshno (Forthcoming) considers markets where only the worker side gets large. Each firm has non-vanishing market power so that it can manipulate stable mechanisms.

2.2.2 Incentive Compatibility of Stable Mechanisms

The previous theorem suggests incentive compatibility of stable matching mechanisms: at any fixed costs, the proportion of agents who have no incentive to manipulate a stable matching mechanism converges to one as the market size increases. However, this requires the condition that all other agents truthfully reveal their preferences. This condition is not guaranteed to hold, as some agents may have large incentives to misrepresent their preferences.

We want to show truthful revelation as an equilibrium behavior of a game induced by a stable matching mechanism. We consider an ϵ-Nash equilibrium, in which agents best-respond to other agents' strategies approximately such that no one can gain more than ϵ by switching to an alternative strategy.

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.*

The equilibrium strategy that we identify is simple. Most agents merely report their true preferences. Agents who misreport their preferences use *truncation strategies*: submitting a preference list of the first k ($k \leq n$) in the same order as the true preference list. Truncations are natural strategies. Agents do not carefully devise the order of the reported preference list.

The equilibrium is based on two important properties of truncations strategies. First, truncation strategies are *undominated* (Roth and Vande Vate (1991)): for any given submitted preferences by other agents, an agent always has a best response that is a truncation of

[16] The models studied in these papers are much more general than a simple replica, and the main questions differ from incentive compatibility.

her true preference list. Thus, each agent can restrictively play a simple truncation strategy without loss in payoffs. Second, when some agents play truncation strategies and benefit, it reduces the gap in utility from different stable matchings for all agents. Let $\succ$ be a true preference profile and $\succ'$ differ from $\succ$ in that a coalition of firms and workers profitably misreport their preferences using truncations. The final matching outcome, which is stable given agents' reported preferences, is stable on true preferences as well, because reversing the truncation from the reported preferences to true ones do not create blocking pairs. Since all stable matchings under $\succ'$ are also stable under $\succ$, the gap in utility from different stable matchings is reduced by truncations.

In the ϵ-Nash equilibrium, a small proportion of agents who have potential gains from manipulations larger than ϵ submit truncations of their true preferences. There exist profitable truncations by these agents such that 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 difference in utilities from stable matchings under the announced preferences, so they will have no significant incentive to respond to others' truncations.

3 Many-to-one Matching

We next consider a large many-to-one matching in which each firm hires up to a fixed number of workers. As in the one-to-one matching, we introduce latent utilities, which in turn generate ordinal preferences. We show that agents in large many-to-one markets are most likely to have only a vanishingly small utility gain by misreporting their preferences, given that all other agents reveal their true preferences. By restricting agents to a class of simple strategies (truncation strategies with capacity constraints), we show that most agents reveal their true preferences as an equilibrium behavior.

3.1 Setup

3.1.1 Many-to-one Matching

Let F be the set of n firms and W be the set of m workers. Each firm has a **capacity**, the number of workers it can hire. We denote firms' capacities by $\mathbf{q} = \langle q_f \rangle_{f \in F}$. Each worker w has a strict preference list $\succ_w$ over firms. Similarly, each firm f has a strict preference list $\succ_f$ of "individual" workers. Let P_f be firm f's preferences over the set of all subsets of W.

We assume that P_f is **responsive** to $\succ_f$ (Roth (1985)). That is, given any $W' \subset W$ with $|W'| < q_f$, (1) for every $w \notin W'$, $W' \cup \{w\} P_f W'$ if and only if w is acceptable, and (2) for every $w, w' \notin W'$, $W' \cup \{w\} P_f W' \cup \{w'\}$ if and only if $w \succ_f w'$.

A **matching** μ is a function from $F \cup W$ to the set of subsets of $F \cup W$ such that (i) for every $w \in W$, $\mu(w) \subset F$ and $|\mu(w)| \leq 1$, (ii) for every $f \in F$, $\mu(f) \subset W$ and $|\mu(f)| \leq q_f$, and (iii) for every firm-worker pair (f, w), $\mu(w) = \{f\}$ if and only if $w \in \mu(f)$. We often write $\mu(w) = f$ when $\mu(w) = \{f\}$, and $\mu(w) = w$ when $\mu(w) = \emptyset$.

A matching μ is **individually rational** if each worker is matched to an acceptable firm and each firm is matched to "individually" acceptable workers. A matching μ is **blocked** by a firm-worker pair (f, w) if $f \succ_w \mu(w)$, and either $w \succ_f w'$ for some $w' \in \mu(f)$ or $|\mu(f)| < q_f$ and w is acceptable to f. A matching is **stable** if it is individually rational and has no blocking pair.

The set of stable matchings is uniquely determined by $\succ$: preference lists over *individual* partners (see Lemma 5.6 in Roth and Sotomayor (1990)). As such, several properties of stable matchings in one-to-one matching carry over to many-to-one matching. The set of stable matchings is nonempty, there exist firm-optimal and worker-optimal stable matchings, and firms and workers have opposite preferences over two distinct stable matchings (see Gale and Shapley (1962), and Lemma 5.6, Corollary 5.9, and Theorem 5.29 in Roth and Sotomayor (1990)). We will denote many-to-one matching markets by using preference profiles over individual partners only.

With the same reason and following the literature, we consider stable mechanisms that take preference lists of *individual* partners and capacities. That is, a **mechanism** M is a function $(\succ, \mathbf{q}) \mapsto M(\succ, \mathbf{q})$. In particular, each firm submits preferences over individual workers only. A mechanism M is called **stable** if $M(\succ, \mathbf{q})$ is a stable matching with respect to $(\succ, \mathbf{q})$.

In many-to-one matching, incentive compatibility is even harder to achieve for stable matching mechanisms than it is in one-to-one matching. No stable matching mechanism, including the firm-optimal stable mechanism, makes it a dominant strategy for all firms to state true preferences and capacities. In some situations, firms can manipulate a stable mechanism and become even better off than they would have been in the firm-optimal stable matching (Roth, 1985).

3.1.2 Random Utilities

A random market is a tuple $\langle F, \mathbf{q}, W, U, V\rangle$, in which $U = [U_{f,w}]$ and $V = [V_{f,w}]$ are two $n \times m$ random matrices representing individual utilities. Similar to one-to-one matching, the individual utilities are defined as

$$U_{f,w} := U(C_w, \zeta_{f,w}) \quad \text{and}$$
$$V_{f,w} := V(C_f, \eta_{f,w}).$$

As before, C_w and C_f are common values, $\zeta_{f,w}$ and $\eta_{f,w}$ are independent private values, and $U(.,.)$ and $V(.,.)$ are continuous and strictly increasing functions from $\mathbb{R}^2_+$ to $\mathbb{R}_+$. As in the one-to-one matching, we assume without loss of generality that common values and private values are uniformly distributed over $[0, 1]$.

We assume that all firms have capacities up to q. A responsive ordinal preferences will be represented by an aggregate function $\Phi : \mathbb{R}^q_+ \to \mathbb{R}_+$ such that the utility a firm f receives from a subset of workers $\{w_1, w_2, \ldots, w_k\} \subset W$ with $k \leq q_f$ is

$$U_{f,\{w_1,w_2,\ldots,w_k\}} := \Phi(U_{f,w_1}, U_{f,w_2}, \ldots, U_{f,w_k}, 0, \ldots, 0).$$

The aggregate function Φ is symmetric (i.e., the value from any q-tuple remains the same for any permutation), continuous, and strictly increasing in every argument.

We study the properties of stable matchings in a sequence of random markets

$$\langle F_n, W_{m_n}, U_{n\times m_n}, V_{n\times m_n}\rangle_{n=1}^{\infty}$$

with simplifying assumptions: all firms have an equal capacity q (fixed for all market size n) and $m_n = q \times n$. This assumption eases our expositions, and the results remain the same in an alternative setup, in which capacities are i.i.d samples from an underlying distribution over $\{1, 2, \ldots, q\}$.

3.2 Results

Similar to one-to-one matching, stable matchings in large many-to-one markets match agents assortatively in the common value dimension. From these assortative stable matchings, we show that the gain from manipulating the worker-optimal stable mechanism by playing truncation strategies and capacity misreports vanishes.

3.2.1 Assortative Feature of Stable Matchings

Define $\phi(u) := \Phi(u, u, \ldots, u)$.

We show that a firm $f \in F$ with common value c would expect to achieve an approximate payoff of $\phi \circ U(c, 1)$.

Take a market instance $\langle F, W, u, v\rangle$. For every ϵ, we define the set of firms whose utilities from all stable matchings are within ϵ difference from their reference levels:

$$A_F^M(\epsilon; u, v) := \{f \in F \mid \phi \circ U(c_f, 1) - \epsilon < U_f^{\mu_W} \leq U_f^{\mu_F} < \phi \circ U(c_f, 1) + \epsilon\}.^{17}$$

The first theorem of one-to-one matching remains the same.

Theorem 3. *For any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|A_F^M(\epsilon; U, V)|}{n}\right] = 1.$$

As in one-to-one matching, agents in large many-to-one markets match assortatively in the common value dimension and obtain favorable matches in the private-value dimension. However, the above theorem does not directly imply incentive compatibility of stable mechanisms. In many-to-one matching, the gain from manipulating a stable mechanism is *not* bounded by the gap in utilities from stable matchings. As such, we derive incentive compatibility (in a limited environment, which we explain in the next section) directly from the assortative feature of stable matchings through careful reasonings. Given a market realization, take a firm f with a common value c_f. Let c_w be the lowest common value among the workers matched to f. The assortative feature of stable matchings suggests that the firm f probably matches to a worker with very good private value. As such, a significant gain from manipulation is achievable only by matching to a worker with a common value significantly higher than c_w. However, those workers are probably out of reach for the firm f as they are matched to other firms with common values significantly higher than c_f.

3.2.2 Incentive Compatibility of the Worker-optimal Stable Mechanism

We derive truthful revelation as an equilibrium behavior of a game induced by a stable matching mechanism in a restricted environment: the worker-optimal stable mechanism is applied, and firms' may manipulate it by using truncation strategies and capacity misrepresentations only. The restricted environment is motivated by the NRMP. The worker-optimal

[17] The superscript M refers to *many-to-one* matching.

stable mechanism is the core of the current matching algorithm, and truncations strategies and capacity misrepresentations are the most easily conceivable manipulations.

Theorem 4. *Suppose that the worker-optimal stable mechanism is applied, and firms may manipulate the mechanism by truncation strategies and capacity misrepresentations only. 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.*

We chose to restrict the environment because of our limited understanding of how some agents' manipulations affect other agents' incentives in the literature. Previously in one-to-one matching, truncation strategies are undominated, and an agent's truncation reduces other agents' incentives for manipulations. But in many-to-one matching, truncation strategies may be dominated (not every gain by manipulations is achievable by playing truncation strategies). We may consider *dropping strategies*, a class of undominated strategies in many-to-one matching (Kojima and Pathak, 2009), but an agent's dropping strategy may increase other agents' incentives (see Example 2 in Ashlagi and Klijn (2012)).

In the restricted environment, we can trace how some agents' manipulations affect other agents' incentives to misreport their preferences. The worker-optimal stable matching mechanism makes it a dominant strategy for every worker to submit her true preferences (Theorem 5.16 in Roth and Sotomayor (1990)). Solely firms may want to manipulate the mechanism by playing truncation strategies combined with capacity misrepresentations. On the other hand, each firm's truncation strategy combined with capacity misrepresentations make all other firms weakly better off (Theorems 5.34 and 5.35 in Roth and Sotomayor (1990)). Every firm's profitable manipulation makes all firms even better off than at the stable matching. As most firms in the stable matching have favorable matches in the private value dimension, their payoff increases must come from re-matching workers with higher common values. Such improvement is likely to be infeasible.

4 Intuition Behind the Proofs

Our proof is based on a new technique from random bipartite graph theory for matching models. To prove the main theorem, in each market realization, we count the number of firms and workers satisfying certain conditions. 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 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 et al., 2001). That is, 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 describe the techniques in greater depth in the following subsections and relegate a detailed proof to the online 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 we 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 5 (Dawande et al. (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*

$$P\left(\beta_n/2 \leq B \leq \beta_n\right) \to 1, \quad as \quad n \to \infty.$$

4.2 Intuition of the Proofs

We demonstrate how to use techniques from the theory of random bipartite graphs in a simplified model, called *a three-tier market* of one-to-one matching 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 $[0, 1]$. 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 $u_{W2} + 1 - \epsilon$: the maximal utility that a firm can achieve by matching with workers in Tier 2. That is, firms in Tier 1 achieve high levels of utility due to the existence of workers in Tier 2. Although not necessarily 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

$$\bar{F} := \left\{ f \in F_1 \mid u_f^{\mu_W} \leq c_{W2} + 1 - \epsilon \right\},$$

and show that

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

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). 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:

$$\zeta_{f,w} \leq 1 - \epsilon \quad \text{or} \quad \eta_{f,w} \leq 1 - (c_{W1} - c_{W2}).$$

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 1).

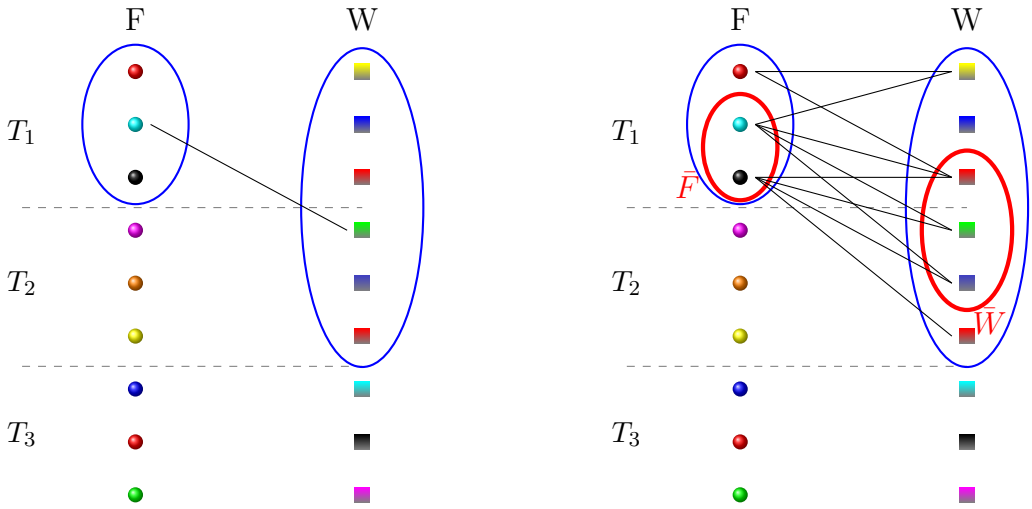


Figure 1: For each realized utility, we draw a bipartite graph with firms in Tier 1 and workers in tiers up to 2 as the partitioned set of nodes (left). Firms in Tier 1 receiving low utilities ($\bar{F}$) and workers in tiers up to 2 matched with firms, not in Tier 1 ($\bar{W}$) form a biclique (right).

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 $f \in \bar{F}$,

$$u_f^{\mu_W} \leq c_{W2} + 1 - \epsilon.$$

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

$$v_w^{\mu_W} \leq c_{F2} + 1.$$

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,

$$\zeta_{f,w} > 1 - \epsilon \quad \text{and} \quad \eta_{f,w} > 1 - (c_{F1} - c_{F2}),$$

and therefore

$$u_{f,w} > c_{W2} + 1 - \epsilon \quad \text{and} \quad v_{f,w} > c_{F1} + 1 - (c_{F1} - c_{F2}) = c_{F2} + 1.$$

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. Suppose that 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, by Theorem 5, a maximum balanced biclique has a size on the order of $\log(n)$ with a sequence of probabilities converging to 1 as n increases. Also, $\bar{W}$ contains at least $n/3$ workers as there are $2n/3$ workers in tiers up to 2, but only $n/3$ firms in tier 1. Therefore, $\bar{F}$ must have a size, at most, on the order of $\log(n)$. The biclique $\bar{F} \cup \bar{W}$ would otherwise contain a balanced biclique with a size bigger than the order of $\log(n)$, violating Theorem 5. Lastly, $\mathbb{E}\left[\frac{|\bar{F}|}{n/3}\right] \to 0$ 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 $[0, 1]$. To study an asymptotic payoff from any stable matching for a firm f with common value $\bar{c} > 0$, we 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 1-4, where agents in the same tier have common values in the same subinterval. As before, we find an asymptotic lower bound of utilities for 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. As such, utilities of Tier 1 firms will be bounded above by $U(\tilde{c}, \bar{u}) - \epsilon$: a level slightly lower than the maximum utility from workers in Tier 3, rather than Tier 2. 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 $\bar{c}$.

5 The Speed of Convergence

We study the speed of convergence in our results. As the speed depend on the utility functions and the distributions of the common and private values, we consider the case of linear utilities (see Example 1) and deterministic common values:

$$\langle c_{f_1}, c_{f_2}, \ldots, c_{f_n} \rangle = \langle c_{w_1}, c_{w_2}, \ldots, c_{w_n} \rangle = \left\langle 1 - \frac{1}{n}, 1 - \frac{2}{n}, \ldots, \frac{1}{n}, 0 \right\rangle.^{18}$$

We prove that

Theorem 6. *For* $\epsilon_n = 6\lambda n^{-1/4}$,

$$P\left(\frac{|A_F(\epsilon; U, V)|}{n} > \theta_n \right) \leq \delta_n,$$

where $\theta_n = O(n^{-1/4})$ *and* $\delta_n = o(e^{-n^{1/2}})$.

To get some intuitions on the convergence speed, consider an $n \times n$ random matrix in which each element is independently either 0 with probability p or 1 otherwise. How large can a square sub-matrix, in which all elements are zero, be? As n increases, it becomes very difficult to have a large sub-matrix with all zero elements. Dawande et al. (2001) shows that the maximum size of sub-matrices with all zero elements is within $O(\log n)$ with probability $1 - o(e^{-n^{1/2}})$. Our main theorem counts the number of agents with strong incentives for manipulation. The number corresponds to the size of a biclique in a random graph, and each biclique, in turn, corresponds to a sub-matrix with all zero elements in a random matrix. As such, the chance of having a large number of agents with strong incentives for manipulation vanishes quickly.

In a market the size of the NRMP, our large market approach seems to give more compelling results than the approach in the previous studies.

First, the convergence speed in probability is faster. For any arbitrary $\epsilon, \theta > 0$, the chance for a market having a large ($> \theta$) proportion of agents with significant ($> \epsilon$) incentives to manipulate a stable mechanism vanishes with speed $o(e^{-n^{1/2}})$. This convergence speed in probability is faster than $O(1/n)$, the corresponding speed in Immorlica and Mahdian (2005) and Kojima and Pathak (2009) (see Remark 1 in the online appendix of Kojima and Pathak

[18] The assumption of deterministic common values is without loss of generality because the distribution of deterministic common values and the empirical distribution of common values from the uniform distribution converge to each other at an exponential rate (see Fact **??** in the online appendix).

(2009)). A real life market of a reasonable size would be explained by our result with a higher probability than the previous studies.

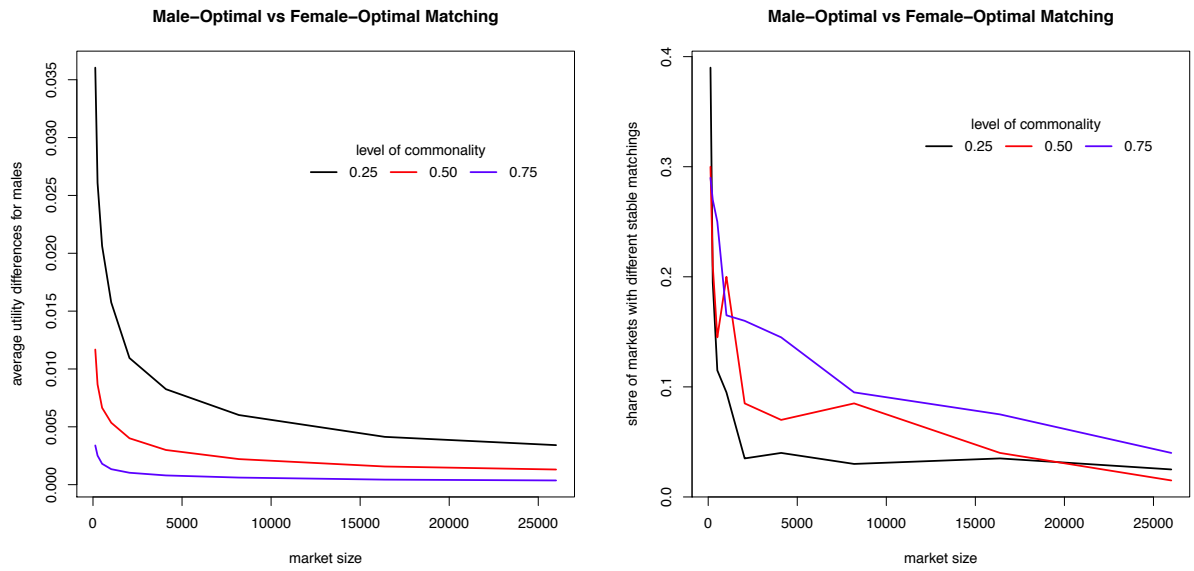
Second, while we could consider only slowly vanishing ϵ_n in theory,[19] a simulation study shows that the average maximum gain from manipulations vanishes quickly. We simulate markets with sizes of $2^7, 2^8, \ldots, 2^{14}(= 16,384)$ and $26,000$. For each market size n, we simulate 100 markets of linear utilities (Example 1). All common and private values are drawn from the uniform distribution on $[0,1]$. For each realized market, we compute each firm's maximum utility gain by manipulation (i.e., the wedge between utilities from firm-optimal and worker-optimal stable matchings), averaged over all firms.

Figure 2 shows the simulation results for one-to-one matching markets. In Panel (a), the solid black line, the long dashed line, and the short dashed line depict the results for the markets with the commonality levels of $1/4, 1/2$, and $3/4$, respectively. Even with a market size of $2^{11}(= 2,048)$, the average maximum gain by manipulation comes close to 0.01, which is 1% of the utility range. When the market size reaches $26,000$, the average maximum gain by manipulation measures around 0.0034, 0.0013, 0.0004, respectively. We compare this convergence speed with the speed of ϵ_n in the previous studies' ϵ_n-Nash equilibrium. For a one-to-one markets, ϵ_n corresponds to the chance that a market of size n contains multiple stable matchings. As such, for each market size n, we simulate 200 markets of linear utilities with the maximum preference length equals to 30. The scale of the vertical axis of Panel (b) shows that the proportion of markets with multiple stable matchings vanishes at a slower speed.[20]

Similarly, we simulate the speeds of vanishing incentives in many-to-one markets. We simulate markets of sizes $2^3, 2^4, \ldots, 2^{14}$, or $26,000$, in which each firm hires $2, 4$, or 8 workers. For each firm, we find the sum of payoffs from matched workers and divide the sum by the firm's capacity. Thus, the utility is in the range of $[0,1]$. We compute $\Phi \circ U(c_f, 1) - U_f^{\mu_W}$: the wedge between each firm's maximum conceivable utility given the firm's common value and the utility from the worker-optimal stable matching. Figure 3 illustrates the numerical results on the average maximum gain. The top left, top right, and bottom panels correspond to the markets in which each firm hires 2, 4, and 8 workers, respectively. In each panel, the solid black line, the long dashed line, and the short dashed line depict the results for the

[19] Still, the speed $O(n^{-1/4})$ is faster than $O(1/\log n)$, the corresponding speed in the environment of pure private-values, $U_{f,w} = \zeta_{f,w}$ and $V_{f,w} = \eta_{f,w}$ (see Footnote 15).

[20] It is important to note that this comparison depends on the maximum preference length. We can scale down or up ϵ_n in Panel (b) by choosing a lower or higher maximum lengths. For example, if the maximum length is 1, no market has multiple stable matchings: $\epsilon_n = 0$ for all n. If the maximum length is $26,000$, most large markets would have multiple stable matchings: $\epsilon_n \approx 1$ for $n \approx 26,000$.



(a) ϵ_n = The expected maximum gain by manipulation.

(b) ϵ_n = The chance of multiple stable matchings (the length of preference lists ≤ 30).

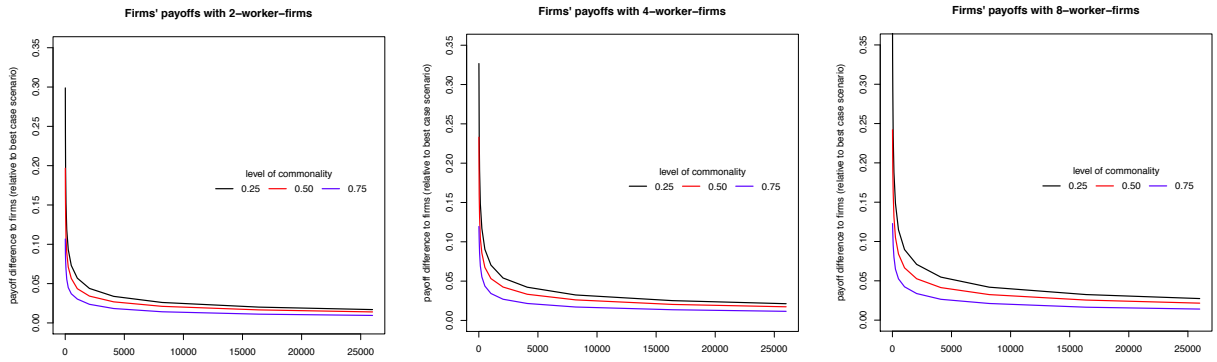
Figure 2: The comparison between the convergence speeds of ϵ_n.

markets with the commonality levels of $1/4, 1/2$, and $3/4$, respectively. In all cases, when the number of workers exceeds $2^{12}(= 4,096)$, the average maximum gain is bounded above by 0.05. The maximum gain falls below 0.014 in all three panels as the number of workers becomes $26,000$.

6 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 consider a market with incomplete information for the case of one-to-one matching and explore a possibility to find a stronger truth-telling incentive. We consider a model in which common values are known to all participants, but private values are known to only the agent who receives the utilities. We will extrapolate findings from the case of complete information to study the incentive compatibility of stable matchings under



(a) Each firm hires 2 workers. (b) Each firm hires 4 workers. (c) Each firm hires 8 workers.

Figure 3: The convergence speeds of vanishing incentives in many-to-one matching

incomplete information.

Let Π_f be a firm f's information about U and V:

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

Take any market realization $\langle F, W, u, v \rangle$. For each firm f, we define its expected utilities from the extreme stable matchings as $EV_f^{\mu_F} := \mathbb{E}\left[V_f^{\mu_F} \,|\, \pi_f\right]$ and $EV_f^{\mu_W} := \mathbb{E}\left[V_f^{\mu_W} \,|\, \pi_f\right]$. Then for each ϵ, we denote the set of firms whose expected utilities from all stable matchings are within ϵ difference of their reference utilities:

$$A_F^E(\epsilon; u, v) := \left\{ f \in F \,|\, U(c_f, 1) - \epsilon < EV_f^{\mu_W} \leq EV_f^{\mu_F} < U(c_f, 1) + \epsilon \right\}.$$

The first theorem, assortative expected payoffs from stable matchings, remains the same.

Theorem 7. *For any* $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{E}\left[\frac{|A_F^E(\epsilon; U, V)|}{n} \right] = 1.$$

The intuition behind this theorem is very simple: an expected value is an average of all realized values. It is likely that most agents have insignificant differences in utilities from stable matchings and their reference levels (Theorem 1). The expected difference in utilities from stable matchings is a convex combination of differences realized in all market instances. For most agents, the expected differences would be negligible as well.

In each market realization, the expected utility gain by manipulation is bounded by the gap between expected utilities from two extreme stable matchings. Thus, as in the case of complete information, the above theorem implies that each agent has vanishing incentives to misreport her preferences, conditioned on all other agents' truth-telling.

However, it turns out to be difficult to find a formal equilibrium of truth-telling. In the literature, we have a limited understanding of strategic behavior under incomplete information: a tractable class of undominated strategies is not known, and it is not feasible to trace how all possible manipulations of some agents affect others' incentives. We managed to find a truth-telling equilibrium only in a special environment: a pure private-value model in which $U_{f,w} = \zeta_{f,w}$. In this environment, an ϵ-Bayesian-Nash equilibrium in which *every* participant reveals her true preference list exists with high probabilities.

Theorem 8. *Suppose a stable matching mechanism is applied to the 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 7. In the pure private value model, asymptotically *every* participant has an insignificant expected gain from manipulation, conditioned on others' truth-telling.[21] Given this observation, consider the following strategy profile. Agents, who have small expected utility gains by manipulation, conditioned on others' truth-telling, submit their true preferences. Any agent, who expects a large utility gain, conditioned on all others' truth-telling, plays any best-response to all other agents' truth-telling. The strategyprofile is an ϵ-Bayesian-Nash equilibrium: participants are approximately best-responding to other agents' strategies (that may not be truth-telling). If agents play the equilibrium strategy, it is most likely that all agents tell the truth in a large market. That is, with a high probability, every agent, whose truth-telling is a best-response to all other agents' truth-telling, is indeed best-responding to other agents' equilibrium strategies.

[21] We can state Theorem 7 as "for any $\epsilon, \delta, \theta > 0$, there exists N such that $P\left(\frac{|A_F^E(\epsilon;U,V)|}{n} > 1-\theta\right) > 1-\delta$, for every $n > N$." If utilities are pure private-values, we could write $P\left(\frac{|A_F^E(\epsilon;U,V)|}{n} = 1\right) > 1-\delta$.

7 Conclusion

We demonstrate an assortative feature of stable matchings as the number of market participants increases. An important implication of one-to-one matching markets is that the proportion of agents who have significant incentives to manipulate stable matching mechanisms vanishes in large markets. Moreover, with high probability, the truthful reporting is an ϵ-Nash equilibrium of the game induced by a stable matching mechanism. These implications hold in many-to-one matching when the worker-proposing Gale-Shapley is applied, and firms play truncation strategies and capacity misrepresentations. We prove our results using techniques from the theory of random bipartite graphs.

References

Agarwal, N. (2015): "An empirical model of the medical match," *American Economic Review*.

Alcalde, J. and S. Barberà (1994): "Top dominance and the possibility of strategy-proof stable solutions to matching problems," *Economic Theory*, 4, 417–435.

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

Ashlagi, I., M. Braverman, and A. Hassidim (2014): "Stability in large matching markets with complementarities," *Operations Research*, 62, 713–732.

Ashlagi, I., Y. Kanoria, and J. D. Leshno (Forthcoming): "Unbalanced random matching markets: The stark effect of competition," *Journal of Political Economy*.

Ashlagi, I. and F. Klijn (2012): "Manipulability in matching markets: conflict and coincidence of interests," *Social Choice and Welfare*, 39, 23–33.

Azevedo, E. M. and E. B. Budish (2013): "Strategy-proofness in the large," *Chicago Booth Research Paper*.

Azevedo, E. M. and J. W. Hatfield (2013): "Complementarity and multidimensional heterogeneity in large matching markets," Tech. rep., mimeo.

Azevedo, E. M. and J. D. Leshno (Forthcoming): "A supply and demand framework for two-sided matching markets," *Journal of Political Economy*.

Bodoh-Creed, A. L. (2013): "Large Matching Markets: Risk, Unraveling, and Incentive Compatibility," Tech. rep., mimeo.

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

Clark, S. (2006): "The uniqueness of stable matchings," *Contributions in Theoretical Economics*, 6, 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, 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, 485–494.

Eeckhout, J. (2000): "On the uniqueness of stable marriage matchings," *Economics Letters*, 69, 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, 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, 304–332.

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

Immorlica, N. and M. Mahdian (2005): "Marriage, honesty, and stability," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, 53–62.

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

Kearns, M., M. Pai, A. Roth, and J. Ullman (2014): "Mechanism design in large games: Incentives and privacy," in *Proceedings of the 5th conference on Innovations in theoretical computer science*, ACM, 403–410.

Knuth, D. (1976): *Mariages stables*, Les Presse De L'Universite De Montreal.

Kojima, F. (2015): "Recent Developments in Matching Theory and Its Practical Applications," *mimeo.*

Kojima, F. and M. Manea (2010): "Incentives in the probabilistic serial mechanism," *Journal of Economic Theory*, 145, 106–123.

Kojima, F. and P. Pathak (2009): "Incentives and stability in large two-sided matching markets," *The American Economic Review*, 99, 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, 1585–1632.

Lee, S. and L. Yariv (2014): "On the efficiency of stable matchings in large markets," *Available at SSRN 2464401.*

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, 878–889.

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

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

Roth, A. (1982): "The economics of matching: Stability and incentives," *Mathematics of Operations Research*, 7, 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, 991–1016.

——— (2002): "The economist as engineer: Game theory, experimentation, and computation as tools for design economics," *Econometrica*, 70, 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, 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, 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, 992–1044.

Roth, A. E. (1985): "The college admissions problem is not equivalent to the marriage problem," *Journal of economic Theory*, 36, 277–288.

Rustichini, A., M. A. Satterthwaite, and S. R. Williams (1994): "Convergence to efficiency in a simple market with incomplete information," *Econometrica*, 1041–1063.

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