Interdependent Utility and Truthtelling in Two-Sided Matching

Xiao Yu Wang*

July 18, 2013

Abstract

Mechanisms which implement stable matchings are often observed to work well in practice, even in environments where the stable outcome is not unique, information is complete, and the number of players is small. Why might individuals refrain from strategic manipulation, even when the complexity cost of manipulation is low? I study a two-sided, one-to-one matching problem with no side transfers, where utility is interdependent in the following intuitive sense: an individual's utility from a match depends not only on her preference ranking of her assigned partner, but also on that partner's ranking of her. I show that, in a world of complete information and linear interdependence, a unique stable matching emerges, and is attained by a modified Gale-Shapley deferred acceptance algorithm. As a result, a stable rule supports truthtelling as an equilibrium strategy. Hence, these results offer a new intuition for why stable matching mechanisms seem to work well in practice, despite their theoretic manipulability: individuals may value being liked.

JEL Classification Codes: C78, D82

Keywords: Two-sided Matching; Interdependent Utility; Stability

**I thank Gabriel Carroll, Sebastian Di Tella, Parag Pathak, and Chris Walters for helpful comments. Support from the National Science Foundation is gratefully acknowledged.*

Duke University. Email: xy.wang@duke.edu.

1 Introduction

Results from the economic study of market design have been used quite successfully in a variety of "real-world" markets in which price mechanisms fail or are not available. *Stability* was established early on as a natural equilibrium concept–in a world where agents belong to one of two groups, and gain by working together across groups, a one-to-one matching is stable if it satisfies two properties: first, each matched agent must prefer being with its partner to remaining single, and second, no two agents who are unmatched in the assignment can match with each other instead and both become better off. Gale and Shapley [5] proposed the simple but powerful deferred acceptance algorithm (DAA), which arrives at a stable matching given the ordinal preferences reported by each agent, and is used to faciliate a number of real-life two-sided matching markets, including the assignment of students to charter schools, and the assignment of medical students to residencies. Since it implements a stable matching, we call the DAA a stable mechanism. According to the DAA, one side is chosen to "propose". For example, suppose that in a standard marriage problem, the men are chosen to propose to the women. Then, each man proposes to his first choice. If a woman receives a single proposal, she holds it. If a woman receives multiple proposals, she holds her most preferred one and rejects the others. Men who were rejected then propose to their next favorite choice, and so on. A man never re-proposes to a woman who rejected him. The stable matching selected is called the man-optimal stable matching; when women propose, that stable matching is woman-optimal. The widespread usage of the DAA is due not only to its simplicity, but also to the fact that it is observed to work quite well in practice. Kojima and Pathak [7] remark, "In real-world applications, empirical studies have shown that stable mechanisms often succeed, whereas unstable ones often fail." Roth [13] provides a large body of evidence.

Yet the theoretical manipulability of stable mechanisms is well-known. For example, Roth [10] shows that there is no stable mechanism where truthfully reporting preferences is a dominant strategy for every agent. In addition, Roth and Sotomayor [15] show that, when any stable mechanism is applied to a marriage market in which preferences are strict and there is more than one stable matching, then at least one agent can misreport her preferences and become matched to a partner more desirable than the one she would have been assigned had she reported the truth, assuming all the other agents are telling the truth. Thus, in situations where there are multiple stable matchings with respect to the set of genuine preferences, at least one person always has an incentive to lie. Moreover, loosely speaking, there are often multiple stable matchings–the core is always nonempty, and the stable matching arrived at by the DAA when men propose generally differs from the matching arrived at when the women propose. Importantly, there are some limits to manipulability: Dubins and Freedman [4] and Roth [10] showed that under the man-optimal DAA, it is a dominant strategy for every man to report preferences truthfully, and similarly for the woman-optimal DAA and women. However, women still have incentives to misreport under the man-optimal DAA, and vice versa.

What might reconcile this apparent paradox of manipulation in theory but not in practice? The aim of this paper is to offer one explanation not yet explored in the literature for why agents do not

seem to take advantage of opportunities to misreport profitably. The intuition is that an agent's utility from a partnership may depend not only on her own views about her partner, but also on her partner's views of her. For example, in the student-school matching problem, a student considers the public profile of characteristics for each of the schools: student-teacher ratio, extracurriculars, financial aid, facilities, and so on. Based on these public profiles, the students rank the schools.

However, the students may very well care about how those schools rank them. If the student thinks a school is great because it has small classes and well-known teachers, but discovers that the school doesn't think very highly of the student, the student might re-think ranking that school highly. After all, if she ends up at the school, she may not receive any attention or resources, making it not such a great option after all.

This interdependence of utility is a realistic feature of many other matching scenarios, including employment and dating/marriage. Working at one's dream firm is not enjoyable if one feels overlooked or undervalued; dating a dream person is not pleasant if that person is dissatisfied. Hence, an agent may care about her desires as well as her desirability in a given match.

I formalize this notion of interdependent utility and explore stable matchings in this framework. Suppose that each market participant has a publicly observable profile of characteristics. The participants care about these characteristics, because once matched, a partnership can produce valued output, and production depends positively on the effort exerted by each partner. The amount of effort a partner exerts depends on her happiness with the partnership. After observing the public profiles of potential partners, an agent is able to determine how satisfied she would be in a relationship with each of the potential partners, and therefore how much effort she would exert in each relationship–she exerts more effort in relationships with people who have characteristics she likes. Thus, she forms "first-round" ordinal preferences over her potential partners. However, she doesn't know how much effort a potential partner would exert in the relationship without knowing something about how satisfied that person is with being her partner. An indicator of the level of satisfaction is that partner's ranking of her. Hence, in this formulation, an agent's "final" utility from a partnership depends on her "first-round" ranking of her partner as well as her partner's "first-round" ranking of her, because the output the partnership ultimately produces depends on both partners' contributions, and an agent's contribution is increasing in her satisfaction.

I show that, for sets of preferences where, in the standard model, there are multiple stable matchings and at least one person has an incentive to misreport, in this framework, there is either a unique stable matching and no agent has an incentive to lie given that everybody else is telling the truth, or there are multiple stable matchings, but still no agent has an incentive to lie given that everybody else is telling the truth. (By "the standard model", I mean the traditional one-to-one, two-sided matching model where each agent submits ordinal preferences, and the utility an agent i gets from being matched with j is higher the more highly she ranks j.) The intuition for this is the following: in the standard model, an agent might misreport her preferences (e.g. declare truncated preferences) in the hopes of getting partnered with someone she has ranked more highly than the person she would be assigned if she reported her preferences truthfully. However, the reason she

isn't assigned to the higher-ranked person if she reports truthfully is because that person isn't as enthusiastic about being partnered with her. In the standard, "independent utility" model, she doesn't care if her partner is enthusiastic about her or not–she just wants to end up with someone as high up her list as possible. But in the interdependent utility model of this paper, our agent cares how enthusiastic her partner is about her. She doesn't find it worthwhile to lie about preferences in order to be partnered with someone whom she ranks very highly, but who ranks her very poorly.

Put another way, in the standard model, multiple stable matchings are associated with across-group disagreement. If, for each school, a school's most preferred student also ranks that school as most preferred, then there will be a unique stable matching. Multiplicity arises if, for instance, the school's most preferred student ranks a different school as her favorite. Interdependent utility "smooths" this across-group disagreement by reducing the distance, so to speak, between two agents' feelings about each other. In the standard model, an agent i in one group could despise agent j in the other group (rank him last), while agent j could adore agent i (rank her first). Hence, the stable match that the DAA selects when one group proposes may differ drastically from the stable match that the DAA selects when the other group proposes. Under interdependent utility, agent i's disapproval of j is tempered by the knowledge that j is devoted to i, and j's adoration of i is tempered by the knowledge that he is i's last choice. This diminished disparity in preferences across groups pushes the market participants towards a sort of agreement on a unique stable matching. In the limiting case where each agent equally weights her desires and her desirability, there is in fact a unique stable matching in almost every instance of ordinal preferences. Moreover, this unique stable matching is either the man-optimal or the woman-optimal matching in the standard framework. In addition, I identify a set of ordinal preferences which I call "perfectly antagonistic"–given these preferences, even under interdependent utility, there is not a unique stable matching (disagreement is irreconcilable, in some sense), but nevertheless it continues to be the case no agent has an incentive to misreport preferences.

Having shown that, in contrast with the standard model, the stable match is generally unique in a framework with interdependent utility where desires and desirability are equally weighted, and hence agents do not have an incentive to misreport their preferences, I then construct an appropriately-modified DAA, based off the sum of the ranks in each potential parternship between agent i in group one and agent j in group two, which selects the unique stable match. To be a bit more specific, a matrix is constructed, where the $(i,j)^{th}$ entry is the sum of the i^{th} member of group one's ranking of the j^{th} member of group two, and the j^{th} member of group two's ranking of the i^{th} member of group one. Since the stable match is unique, it doesn't matter which side proposes. Suppose WLOG that group one proposes. Then, each member of group one proposes to the member of group two with whom she has the minimal rank-sum. Ties can be broken randomly, or by using the initial ordinal rankings of the proposer. If a member of group two receives a single proposal, she holds it; if she receives multiple proposals, she keeps the one with minimal rank-sum and rejects the others. Any member of group one who was rejected then proposes to the member of group two with whom she has the next smallest rank-sum, and so on. No member ever re-reproposes

to someone who already rejected her.[1]

Other explanations exist in the literature for why misreporting does not seem to occur under stable mechanisms. One branch studies the computational complexity of calculating the optimal lie. Some researchers have used methods from computer science to show that the problem of profitable manipulation may be NP hard (Pini et al. [9]). On a related note, Roth and Rothblum [14] argue that part of the difficulty of constructing the optimal lie may stem from incomplete information. Perhaps one needs to know the preferences of one's fellow group members, or the preferences of the members of the other group. In fact, Roth and Rothblum [14] show that, even when information is highly incomplete, reporting truncated preferences (that is, declaring one's least preferred options to be unacceptable when they are actually acceptable, just heavily disliked) is still profitable. Truncation strategies are also not particularly computationally complex.

Another set of papers argues that there is little to gain from manipulation in large markets. Kojima and Pathak [7] show that in many-to-one matching markets, such as students to schools, as the number of participants goes to infinity but the length of preference lists stays fixed, the fraction of participants with an incentive to misreport their preferences, given that everyone else is telling the truth, goes to 0. More specifically, under the student-optimal stable mechanism in the schools-students matching problem, we know that it is a dominant strategy for students to report their preferences truthfully, but that schools have an incentive to misreport their quotas. Kojima and Pathak [7] show that the benefit to schools from manipulating their quotas diminishes as the market grows large.

The key distinction between the existing explanations for the absence of manipulation of stable mechanisms in practice and the explanation advanced by this paper is that in existing frameworks, market participants *do* have an incentive to misreport their preferences, only it is difficult to figure out how, or the gain from it is small. By contrast, the results of this paper suggest that if an agent simply cares not only about being matched with a partner she likes, but also about being matched with a partner who likes her, then in fact there will be a unique stable matching and no agent will have an incentive to lie if the others are telling the truth.

The rest of the paper proceeds as follows. In the next section, I work through a simple example to demonstrate how an agent may gain in the standard framework by misreporting her preferences. I then set up the model of interdependent utility, and present the results. I revisit the example to demonstrate that agents no longer have the incentive to lie in a framework with interdependent utility. Finally, I conclude.

2 An Example

Consider a standard marriage market with four women and four men. Suppose nobody prefers being single to being matched, and ordinal preferences are:

[1] The model, the algorithm (including how to deal with declaring unacceptable agents), and the results will be presented much more precisely in the coming sections. This is merely meant to serve as an overview of the main idea of the paper.

$$
\begin{aligned}
M1 &: W2 > W3 > W1 > W4 \\
M2 &: W1 > W2 > W4 > W3 \\
M3 &: W4 > W3 > W2 > W1 \\
M4 &: W1 > W3 > W2 > W4
\end{aligned}
$$

$$
\begin{aligned}
W1 &: M1 > M3 > M2 > M4 \\
W2 &: M1 > M3 > M4 > M2 \\
W3 &: M2 > M1 > M3 > M4 \\
W4 &: M4 > M2 > M3 > M1
\end{aligned}
$$

Then, in the first round of the man-proposing DAA, the men propose as follows:

$$
\begin{aligned}
M1 &\rightarrow W2 \\
M2 &\rightarrow W1 \\
M3 &\rightarrow W4 \\
M4 &\rightarrow W1
\end{aligned}
$$

Since $W1$ receives two proposals, she holds the one she prefers, which is $M2$'s proposal. Thus, $M4$ must propose to his next favorite woman, $W3$, who has no competing proposals. The DAA ends at this point, and the man-optimal stable matching is:

$$
\mu_M = \begin{bmatrix} M1 & M2 & M3 & M4 \\ W2 & W1 & W4 & W3 \end{bmatrix}
$$

The woman-proposing DAA runs in a similar fashion, except the women propose. This produces the woman-optimal stable matching:

$$
\mu_W = \begin{bmatrix} M1 & M2 & M3 & M4 \\ W2 & W3 & W1 & W4 \end{bmatrix}
$$

Note that $\mu_M \neq \mu_W$, so there are multiple stable matchings. We know from Dubins and Freedman [4] and Roth [10] that given these preferences, under any stable mechanism, at least one person has an incentive to misreport preferences given the others are telling the truth.

For example, suppose the man-proposing DAA is being run, and $W4$ "misreports" her preferences by truncating them. That is, $W4$ reports that the only man acceptable to her is $M4$ (but her true preferences are as described above).

Then, in the first round of the man-proposing DAA, the men propose as follows:

$$\begin{aligned} M1 &\rightarrow W2 \\ M2 &\rightarrow W1 \\ M3 &\rightarrow W4 \\ M4 &\rightarrow W1 \end{aligned}$$

$W1$ again receives multiple proposals from $M2$ and $M4$, and holds the one she prefers, $M2$'s, but in addition $W4$ rejects $M3$'s proposal, because she has declared him to be unacceptable. Hence, $M4$ and $M3$ have to propose again. $M4$ proposes to his next favorite woman, $W3$, while $M3$ proposes to $W3$ as well.

$$\begin{aligned} M1 &\rightarrow W2 \\ M2 &\rightarrow W1 \\ M3 &\rightarrow W3 \\ M4 &\rightarrow W3 \end{aligned}$$

$W3$ holds the proposal she prefers, $M3$'s, so $M4$ proposes to $W2$.

$$\begin{aligned} M1 &\rightarrow W2 \\ M2 &\rightarrow W1 \\ M3 &\rightarrow W3 \\ M4 &\rightarrow W2 \end{aligned}$$

But $W2$ already has a proposal from $M1$, who is her top choice, so she rejects $M4$, who finally proposes to $W4$, and she accepts. This leads to the matching:

$$\begin{bmatrix} M1 & M2 & M3 & M4 \\ W2 & W1 & W3 & W4 \end{bmatrix}$$

So, by lying and truncating her preferences, $W4$ ends up with her top choice of partner, $M4$, under the man-proposing DAA, whereas she would have ended up with $M3$ if she had told the truth. In this model, an agent is strictly happier to be paired with a partner she ranks more highly.

Now, let's turn to the model of this paper.

3 The Model

The economy is composed of two disjoint groups of rational, utility-maximizing agents, who have utility functions $u(x) = x$. Denote the groups by M and W, where $|M| = |W| \in \{2, 3, ..., N\}$, $N < \infty$. Suppose that members of M and W have to build partnerships one-to-one across groups in order to produce some kind of valuable output. For example, these two groups could be men and women, or students and schools, or employees and employers.

Suppose that each agent i has a public profile of characteristics, $X_i \in \mathbb{R}^k$, $k \in \mathbb{N}$, which is known to herself and observable to others. For example, for a school, this public profile might consist of the average student-teacher ratio, the number of classrooms, the extracurriculars available, the facilities, teacher quality, and so forth. For a student, this public profile might consist of past grades and past exam scores.

Suppose that the output produced by a matched partnership (i, j) depends on the characteristics of the partners, X_i and X_j, in the following way. Any matched partnership (i, j) has the possibility of capturing a pie of size y. However, once matched, i and j must contribute to the relationship in order to capture some fraction of the pie. The level of an agent's contribution depends on her compatibility and satisfaction with her partner. Since a contribution is personally costly, an agent will contribute highly only if she is satisfied with the relationship. The joint contributions of the agents determine the output produced; output is increasing in the level of contribution.

In particular, an agent $i \in M$ matched with an agent $j \in W$ chooses a contribution:

$$c_i^j : X_i \times X_j \to \mathbb{R}$$

(There is no restriction or assumption on the form of this mapping, or on the profiles of characteristics.)

So, $\{c_i^1, ..., c_i^N\}$ is the vector whose j^{th} element describes the contribution i would make in a relationship with the j^{th} member of W. Let $\rho(c_i^j)$ denote the ranking of the size of the contribution c_i^j relative to the other elements in the vector, from largest to smallest. For example, if c_i^j is the largest number in the vector, then $\rho(c_i^j) = 1$, indicating that i perceives j to be her most compatible partner, and is most willing to make contributions in a relationship with him. Assume that no elements of the vector are the same, so that each i has strict preferences over possible partners j, based on her evaluation of the compatibility of their characteristics.

Suppose that the fraction of the pie of size y captured by a partnership (i, j) is $\theta(i, j)$, where

$$\theta(i, j) = f(\rho(c_i^j)) + f(\rho(c_j^i))$$

and

$$f(x) = \frac{1}{2}\frac{N - x}{N - 1}$$

So, the output produced by a partnership (i, j) is:

$$\theta(i,j)Y = \frac{N - \left(\frac{\rho(c_i^j)+\rho(c_j^i)}{2}\right)}{N-1}Y$$

where $\rho(\cdot) \in \{1, ..., N\}$ since it is a ranking.

For instance, suppose that a student i has the characteristic that she wants to become a mathematician, and so wishes to match with a school which prioritizes the sciences and excels at math competitions. Suppose school j prioritizes the sciences, while school k prioritizes theater. Then, student i recognizes she is much more compatible with school j than with school k–this matters because she knows that at school j, she will work very hard and become involved with the math competitions, while at the same time receiving the support and the resources that she needs from school j, which is excited to have her as a student. By contrast, if she were to attend school k, the student knows she would not exert very much effort in theater, and would likely also be ignored by the school, as they prefer to devote their resources to those with talent in theater.

Hence, student i ranks school j above school k–not only does i prefer j to k because she prefers science to theater, she also knows that school j will give her time and resources. Hence, both i and j will contribute highly to a relationship with each other, and this will lead to high productivity. In particular, if i and j rank each other first, that is, $c_i^j = c_j^i = 1$, then $f(c_i^j) = f(c_j^i) = \frac{1}{2}$, so that $\theta(i,j) = 1$, and a partnership between student i and school j captures the entire pie y.

Note that in this model, an agent always prefers being matched to somebody rather than remaining single. However, when I discuss a mechanism to implement stable matches in this setting, I will assume that the "central computer" which is running the mechanism does not know that nobody prefers being single, and so agents are able to misreport partners as being unacceptable if they wish.

A match function μ assigns distinct members of M to distinct members of W, where each person is assigned at most one partner. A matching is an equilibrium if it is *stable*–that is, it must satisfy two properties:

1. Individual rationality: every agent prefers being matched with her partner to remaining single (this is satisfied for every agent in this model)

2. No blocking: no two agents who are unmatched under μ can make themselves both better off by matching with each other instead of obeying the assignment μ

This is the set up of the model. The natural questions to ask now are: does at least one stable matching always exist? When is there a unique stable matching? Can we construct an algorithm to select a stable matching? What are the strategic properties of this algorithm? These questions are addressed in the next section.

4 Results

Because this is a model of transferable utility (all agents are risk-neutral with utility functions $u(x) = x$), we know that an equilibrium matching maximizes aggregate output[2]. That is, we know that each partnership produces a joint output. An equilibrium matching is one in which any switching of partners results in a decrease in the sum of output across all pairs.

Recall that the joint output of a partnership (i, j) is:

$$\theta(i,j)Y = \frac{N - \left(\frac{\rho(c_i^j)+\rho(c_j^i)}{2}\right)}{N-1}Y$$

where $\rho(c_i^j)$ is i's ranking of j based on i's evaluation of the compatibility between i's characteristics and j's public profile, X_j, and $\rho(c_j^i)$ is j's ranking of i based on j's evaluation of the compatibility between j's characteristics and i's public profile, X_i.

Thus, it is clear that the matching which maximizes the sum of output across all pairs is also the matching which minimizes the sum of within-partnership rank-sums across all pairs. That is, if i ranks j m^{th} $(\rho(c_i^j) = m)$, and j ranks i n^{th} $(\rho(c_j^i) = n)$, then the within-partnership rank-sum is $(m + n)$.

This suggests the following approach. The model can be mapped into an alternative model where the utility of an agent i from being matched with an agent j directly depends on i's ranking of j and j's ranking of i (where the rankings are based on each agent's evaluation of the public profile of characteristics of potential partners). In particular, let $u_{ij} = -(\rho_i^j + \rho_j^i)$. A profitable block in this model would be if an agent i could instead match with an agent k such that the sum of i's rank of k and k's rank of i is *smaller* than the sum of agent i's rank of j and j's rank of i. Then, we look for the matching which minimizes the total sum of ranks across all matched pairs. This matching will be the same matching that maximizes aggregate output in the underlying model. The intuition is that in the underlying model, an agent i wants to work with an agent j whom she views as compatible, because she knows she will contribute more to a relationship in which she feels satisfied, but she wants agent j to view her as compatible, too–if j doesn't have a high opinion of her, than j will not contribute to the relationship, and the output produced will be low. Hence, i cares about j's ranking of her as well as her ranking of j.

So, this model can be analyzed by studying the alternative model where each agent forms ordinal preferences over potential partners based off their characteristics, but additionally cares about how those potential partners rank her, so that her utility from a partnership depends equally and directly on her characteristics-based ranking of her partner and on her partner's characteristics-based ranking of her.

To think about existence of equilibrium, and other properties of the equilibrium, it will be helpful to develop several definitions.

[2] Becker [1] and Shapley and Shubik [16] both noted this.

Definition 1 *The* ***M matrix*** *is the* $N \times N$ *matrix where the* $1 \times N$ *row vector* $M(i,:)$ *represents man* i*'s rankings of women* $\{W1, ..., WN\}$*. That is, the* (i,j) *element of the* M *matrix is the rank that man* i *assigns to woman* j*. The* ***W matrix*** *is the* $N \times N$ *matrix where the* $N \times 1$ *column vector* $W(:,j)$ *represents woman* j*'s rankings of men* $\{M1, ..., MN\}$*. That is, the* (i,j) *element of the* W *matrix is the rank that woman* j *assigns to man* i*.*

Note that if no man finds any woman unacceptable, each row of M will be some permutation of $[1...N]$. Similarly, if no woman finds any man unacceptable, each column of W will be some permutation of $[1...N]$.

While it is the case that in this model, an agent always prefers being matched to somebody rather than remaining single, I suppose that the entity which runs the stable mechanism and to which the ordinal preferences are reported does not know this. Hence an agent is able to misreport a potential partner as being unacceptable by ranking her $2N$.

Definition 2 *The* ***ranks matrix*** *is the* $N \times N$ *matrix* $\mathbf{ranks} = M + W$*.*

In the previous example, the M matrix and the W matrix would be as follows:

$$M: \begin{bmatrix} 3 & 1 & 2 & 4 \\ 1 & 2 & 4 & 3 \\ 4 & 3 & 2 & 1 \\ 1 & 3 & 2 & 4 \end{bmatrix}, \; W: \begin{bmatrix} 1 & 1 & 2 & 4 \\ 3 & 4 & 1 & 2 \\ 2 & 2 & 3 & 3 \\ 4 & 3 & 4 & 1 \end{bmatrix}$$

Since members of M are along the rows and members of W are along the columns, this tells us that $W2$ is $M3$'s 3^{rd} choice, while $M3$ is $W2$'s 2^{nd} choice.

So, the ranks matrix is:

$$ranks: \begin{bmatrix} 4 & 2 & 4 & 8 \\ 4 & 6 & 5 & 5 \\ 6 & 5 & 5 & 4 \\ 5 & 6 & 6 & 5 \end{bmatrix}$$

The first result establishes both existence of a stable matching (nonemptiness of the core), and in so doing demonstrates how the DAA can be modified to find a stable matching for this interdependent utility framework.

(A member of M will be referred to as a "man", and a member of W as a "woman", for ease of exposition.)

Proposition 3 *(Existence) A stable match with respect to any given set of strict preferences under interdependent utility always exists.*

Proof. The proof is by construction. Consider a modified deferred acceptance algorithm where, in the men-proposing version, men propose first to the woman with whom he would have lowest rank-sum, out of the pool of women he finds acceptable. (If a man finds no women acceptable, he makes no proposals and remains single.) If more than one woman yields the same minimal rank-sum, the man chooses the woman whom he also privately prefers ("selfish tiebreak"). If a woman receives multiple acceptable proposals, she "holds" the man with whom she would have minimal rank-sum, and rejects the other proposals. If two men yield the same minimal rank-sum, she picks the man she also privately prefers. If a woman receives one acceptable proposal, she "holds" that man. If a woman receives only unacceptable proposals, she rejects them. If a woman receives no proposals, she remains single.

In the next round, men who were rejected in the directly previous round propose again, to the woman with whom he would have lowest rank-sum, out of the pool of women who have not rejected him yet and whom he finds acceptable. Women who receive proposals re-evaluate all their options, including men they might have "held" at earlier stages, and keep only the man with whom their rank-sum is smallest (breaking ties selfishly).

This algorithm repeats until no man has been rejected in the directly previous round.

The match this algorithm produces must be stable. Consider a match generated by the algorithm, and suppose that man m has a lower rank-sum with woman w' than with woman w, who is his current partner. Since man m is matched to woman w, he must have found w acceptable (because otherwise he would not have proposed to her), and by transitivity (rationality of preferences) woman w' must also be acceptable to man m. Thus, man m must have proposed to woman w' before proposing to woman w. Since man m is not matched to w' at the end of the algorithm, w' must have rejected man m for a man with whom she has lower rank-sum, or with whom she has the same rank-sum but privately prefers. Thus, woman w' is matched, at the end of the algorithm, to a man she likes more than man m, since preferences are transitive and women cannot do worse as the algorithm progresses, because men propose to them and they hold acceptances. Thus, (m, w') cannot be a blocking pair. (The same logic applies if man m is single at the end of the algorithm, and there exists an acceptable woman w' for him. It must have been that he proposed to w', but that she rejected him in favor of a man she likes more. Thus, (m, w') cannot be a blocking pair.)

Now suppose that man m has the same rank-sum with woman w' than with woman w, who is his current partner, but m privately prefers w' to w. Since man m is matched to woman w, he must have found w acceptable (because otherwise he would not have proposed to her), and by transitivity (rationality of preferences) woman w' must also be acceptable to man m. Thus, man m must have proposed to woman w' before proposing to woman w, by the selfish tiebreak rule. Since man m is not matched to w' at the end of the algorithm, w' must have rejected man m for a man with whom she has lower rank-sum, or for a man with whom she has the same rank-sum but privately prefers. Thus, woman w' is matched, at the end of the algorithm, to a man she likes more than man m, since preferences are transitive and women cannot do worse as the algorithm progresses, because men propose to them and they hold acceptances. So, (m, w') cannot be a blocking pair.

Since there are no blocking pairs, the match must be stable. ■

This establishes that the core is nonempty in this model, and describes a method for finding a stable matching, given the reported ordinal preferences.

What more can we say about stable matchings in this setting? To think about this, we must partition the set of possible preference profiles.

Definition 4 *Given a $T \times T$ matrix A, a matrix B is the* ***vertical reflection*** *of A iff $b_{ij} = a_{i(T-j)}$ for every row i and column j, $i, j \in \{1, ..., T\}$.*

Definition 5 *A preference profile $\begin{bmatrix} M & W \end{bmatrix}$ is* ***perfectly antagonistic*** *iff M is symmetric ($M = M'$), W is symmetric ($W = W'$), and W is the vertical reflection of M. In other words, for every agent i, [agent i's m^{th} choice ranks agent i n^{th}] $\Rightarrow$ [agent i's n^{th} choice ranks agent i m^{th}], $m, n \in \{1, ..., N\}$, $m+n = (N+1)$. Note that in this case, every element of the ranks matrix is equal to $(N+1)$. A preference profile is* ***imperfectly antagonistic*** *iff it is not perfectly antagonistic.*

Definition 6 *A special case of imperfectly antagonistic preferences is* ***mutual x*** *preferences, where, given some $x \in \{1, ..., N\}$, every agent's x^{th} choice ranks him/her x^{th}. An agent's preferences are* ***mutual first*** *if his/her first choice ranks him/her first.*

(The case where every agent has mutual first preferences is not interesting, since there is no within-side or between-side conflict of any kind, and the optimal matching is obvious–everybody gets his or her first choice.)

There exist also preference profiles where a subset of agents have perfectly antagonistic preferences, over all agents *in that subset.*

Definition 7 *A preference profile is* ***essentially perfectly antagonistic*** *if players' preferences can be partitioned into mutual first preferences and perfectly antagonistic preferences. That is, $M_f \subset M$, $W_f \subset W$, $|M_f| = |M_w| = K \in \{1, 2, ..., N\}$, where M_f, W_f have mutual first preferences for each other, while $M \backslash M_f$ and $W \backslash W_f$, both of cardinality $(N - K)$, have perfectly antagonistic preferences for each other. A preference profile is* ***essentially imperfectly antagonistic*** *iff it is not essentially perfectly antagonistic.*

This partition of preference profiles is specific to the structure of interdependence in this model. Because the model maps into a framework where an agent's utility from a match depends directly and equally on each partner's rank of the other, and in particular the dependence is linear, we must treat specially sets of preference profiles where the rank-sum for all possible partnerships within a weak subset of agents in M and W is exactly the same.

For example, the classic case of perfect antagonism is the following:

$$\begin{aligned} M1 &: W1 > W2 \\ M2 &: W2 > W1 \end{aligned}$$

$$W1 \quad : \quad M2 > M1$$
$$W2 \quad : \quad M1 > M2$$

$M1$'s favorite woman is $W1$, but her favorite man is $M2$, whose favorite woman is $W2$, whose favorite man is $M1$. Thus, the ranks matrix is:

$$ranks : \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}$$

and both of the possible matches are stable.

Note that the antagonism is "perfect", because not only does everyone's first choice prefer someone else, but if the match were $(M1, W1)$ and $(M2, W2)$, then making $M1$ happier by partnering him with $W2$ decreases $W2$'s happiness by exactly the amount that $M1$'s happiness was increased. That is, no one's happiness can be increased via a change of partner without decreasing someone else's happiness by an exactly-offsetting amount.

The next result shows that there is a unique stable matching given essentially imperfectly antagonistic preference profiles, while there are multiple stable matchings given essentially perfectly antagonistic preference profiles.

Proposition 8 *(a) (Essentially imperfectly antagonistic preferences) The deferred acceptance algorithm generates the same match when men propose as when women propose. Further, the generated match is the unique stable match with respect to the preference profile, under interdependent utility.*

(b) (Essentially perfectly antagonistic preferences) There are multiple stable matchings: the K men and K women who mutually prefer each other first are always matched with each other, and the $(N-K)$ men and women with perfectly antagonistic preferences are matched with each other such that all men receive their m^{th} choice, and all women receive their n^{th} choice, where $m \in \{1, ..., N-K\}, n \in \{1, ..., N-K\}$, $m + n = N - K + 1$.

Proof. (a) Consider any given imperfectly antagonistic preference profile. Let a stable matching with respect to this preference profile be denoted μ. By definition of stability, we know that every agent is matched to an acceptable partner, and there are no blocking pairs: that is, there is no pair (i, j), $i \in M$, $j \in W$, such that $-(rank_i(\mu(i)) + rank_{\mu(i)}(i)) < -(rank_i(j) + rank_j(i))$ and $-(rank_j(\mu(j)) + rank_{\mu(j)}(j)) < -(rank_i(j) + rank_j(i))$. Now consider a different matching, obtained by switching the partners of two men, i and i', so that i is now matched with $\mu(i')$ and i' is matched with $\mu(i)$. But then it must be that *both* $(i, \mu(i'))$ and $(i', \mu(i))$ are worse off–if $(i, \mu(i'))$ became better off, they would have blocked the original match, and if $(i', \mu(i))$ became better off, they would have blocked the original match. But then this new match cannot be stable, since $(i, \mu(i))$ and $(i', \mu(i'))$ are blocking pairs. Since any possible matching can be achieved by switching

two partners in the original match, it follows the the original stable matching μ must be the unique stable match.

Since it was proven in the previous proposition that the deferred acceptance algorithm (regardless of which side proposes) generates a stable matching, it must be that the deferred acceptance algorithm generates the same match when men propose as when women propose.

(b) It is clear that in any stable matching, mutual firsts are matched to each other. Consider the remaining $(N-K)$ agents, who have perfectly antagonistic preferences over each other, and let their section of the M, W, and $ranks$ matrices be denoted M_T, W_T, and $ranks_T$, respectively. Then, by definition, $M_T = M'_T, W_T = W'_T$, M_T is the vertical reflection of W_T, and every element of $ranks_T$ is the same, and equal to $(N-K+1)$. Hence, every agent gets the same base utility $-u(x,y) = (N-K+1)$ from being matched with any partner. Moreover, in any possible matching, all men are matched to their m^{th} choice, while all women are matched to their $(N-K+1-m)^{th}$ choice. The selfish tiebreak rule means that man i and woman j can be a blocking pair iff man i ranks j more highly than his given partner, *and* woman j ranks man i more highly than her given partner. But, by definition of perfectly antagonistic preferences, every agent on one side's gain comes at their partner on the other side's loss. Thus, there can be no blocking pairs from matches where mutual first agents are matched to each other, and among the remaining perfectly antagonistic agents, each man is matched to his m^{th} choice, and each woman is matched to her $(N-K+1-m)^{th}$ choice. ■

Hence, many preference profiles which yield multiple stable matchings in the standard framework (some of which are better for M and some of which are better for W), and which therefore seem characterized by a seemingly irreconcilable level of conflict, may not be so irreconcilable after all, if we account for interdependence of utility. Except in cases of perfect antagonism (where an increase in anybody's happiness is exactly offset by a decrease in someone else's happiness), there is a unique stable matching, and therefore it is irrelevant which side proposes. Moreover, although there are multiple stable matchings in cases of perfect antagonism, market participants are indifferent over those matchings–it isn't the case that some of the matchings are better for one set of people, while others are better for another set of people.

Not only is the stable matching unique given essentially imperfectly antagonistic preference profiles, but it has a special property–it is either the man- or the woman-optimal stable matching from the standard framework.

Proposition 9 *The unique stable match with respect to a given (essentially) imperfectly antagonistic preference profile under interdependent utility is either the man- or the woman-optimal stable match with respect to preferences under standard utility.*

Proof. We know that, in the standard model, the woman-optimal match gives every woman her best achievable mate and the man his worst achievable mate; the man-optimal match gives every man his best achievable mate and every woman her worst achievable mate. "Achievable" here is

in the standard sense of "most preferred mate among the complete set of stable matches" (where stability is based purely on independent utility, so that a man and woman block iff they each rank each other higher than their original partners). Clearly, in the independent utility framework, every woman's weakly favorite match is the woman-optimal match, and every man's weakly favorite match is the man-optimal match, where these two matches generally differ. However, if agents have linearly interdependent utility, the woman-optimal match is the match where each woman is matched to her best "achievable" mate, and the man is matched to a woman *who likes him to the degree* that he is her *best* "achievable" mate. Similarly, in the man-optimal match, the man is matched to his best "achievable" mate, and the woman is matched to a man who *likes her to the degree* that she is his *best* "achievable" mate. Thus, all other matches that are not the man- or the woman-optimal match must do strictly worse under interdependent utility, since men and women not being matched to their best achievable mates implies that men and women are also not matched to someone who desires them to the degree that they are their best achievable mate. And, as long as preferences are not (essentially) perfectly antagonistic, either the man- or the woman-optimal must do strictly better. Therefore, the unique stable match with respect to a given imperfectly antagonistic preference profile under interdependent utility must be either the man- or woman-optimal stable match with respect to the same preferences, under independent utility. ■

Finally, what can we say about strategic properties of the stable mechanism? We know in the standard case that the core is nonempty, and that the set of stable matchings is generally not unique–in particular, the man-optimal stable matching differs from the woman-optimal stable matching. We know that under a stable mechanism, truthtelling is never a dominant strategy, and moreover, if there are multiple stable matchings, then at least one person has an incentive to misreport preferences, assuming the other are telling the truth.

By contrast, we know that if agents' utilities are interdependent, in the sense that they care about the contributions made by each partner to the relationship, where a partner's contributions are increasing in satisfaction with the relationship, then the core is nonempty, and the stable matching is generally unique. Multiple stable matchings arise only in cases of perfectly antagonistic preference profiles, and in that case, all agents are simply indifferent over those matchings.

The natural question to ask is, will at least one agent have an incentive to misreport preferences under a stable mechanism in this model, assuming all other participants are telling the truth? The answer turns out to be no.

Proposition 10 *In the basic model with interdependent utility and (essentially) imperfectly antagonistic preferences, as long as no agent is pivotally antagonistic, no individual agent has an incentive to misreport preferences if all other agents are telling the truth. Therefore, given a stable rule, it is always an equilibrium for an agent to state his or her true preferences.*

Proof. Suppose that the preference profile is imperfectly antagonistic. Suppose an agent misreports his preferences while all other agents report the truth. The principal uses a stable matching

mechanism, so the resulting match will be the *unique* stable match with respect to the *false* preferences, where the agent is matched to a different partner than he would have been had he reported truthfully (else, there would have been no reason to lie). But by Proposition 9(a), we know that this agent and the partner he would have had had he reported truthfully must be a blocking pair in the unique stable match with respect to the false preferences. That is, the agent is better off with his truthful partner than the partner he gets by lying. Hence, truthtelling is an equilibrium strategy.

Now, suppose the full preference profile is essentially perfectly antagonistic. Then all agents are indifferent over the multiple stable matchings, so no agent has an incentive to lie. ■

Hence, if agents' utilities are interdependent in the sense of this model, then truthful reporting of preferences is a best response under a stable mechanism such as a modified DAA, if everyone else is reporting truthfully. The across-group disagreement that resulted in multiple stable matchings in the standard framework is resolved by interdependence–an agent no longer cares only about being matched with someone as high up her list as possible. She also values her partner's opinion of her, since that is her only indication of the contribution her partner would make to a joint relationship, and if her partner is dissatisfied, the contribution, and subsequently the output, will be low.

Thus, our observation that stable mechanisms work well in practice despite the seemingly various ways to manipulate profitably in theory might be due to the fact that agents care about their desirability as well as their desires. If this is the case, then I've shown that there are no ways to gain by miresporting preferences under a stable mechanism, including reporting truncated preferences. So, one interpretation of these results is that unaccounted-for interdependence explains the smooth functioning of stable mechanisms.

Another interpretation is that, if we think there *is* interdependence of agents' utilities, (or we know there is interdependence based on empirical or experimental work), then we should implement the stable mechanism described in the existence proof. That is, we should ask both sides to report their ordinal preferences, and then run the modified DAA on the ranks matrix to find the unique stable matching (which side proposes is irrelevant). We can honestly assure market participants that truthtelling is an optimal strategy.

5 An Example: Followup

Recall the ordinal preferences from the first example:

$$\begin{aligned} M1 &: W2 > W3 > W1 > W4 \\ M2 &: W1 > W2 > W4 > W3 \\ M3 &: W4 > W3 > W2 > W1 \\ M4 &: W1 > W3 > W2 > W4 \end{aligned}$$

$$
\begin{aligned}
W1 &: M1 > M3 > M2 > M4 \\
W2 &: M1 > M3 > M4 > M2 \\
W3 &: M2 > M1 > M3 > M4 \\
W4 &: M4 > M2 > M3 > M1
\end{aligned}
$$

I showed that the man-optimal and the woman-optimal stable matching differed:

$$
\mu_M = \begin{bmatrix} M1 & M2 & M3 & M4 \\ W2 & W1 & W4 & W3 \end{bmatrix}
$$

$$
\mu_W = \begin{bmatrix} M1 & M2 & M3 & M4 \\ W2 & W3 & W1 & W4 \end{bmatrix}
$$

Thus, at least one person has an incentive to misreport preferences under a stable mechanism. For example, if the man-optimal DAA is the stable mechanism, then $W4$ can gain by reporting only $M4$ to be acceptable to her.

Now suppose that agents have interdependent utility as described by this model. Then the ranks matrix is:

$$
ranks: \begin{bmatrix} 4 & 2 & 4 & 8 \\ 4 & 6 & 5 & 5 \\ 6 & 5 & 5 & 4 \\ 5 & 6 & 6 & 5 \end{bmatrix}
$$

Suppose men propose. Then the first round is:

$$
\begin{aligned}
M1 &\rightarrow W2 \\
M2 &\rightarrow W1 \\
M3 &\rightarrow W4 \\
M4 &\rightarrow W1
\end{aligned}
$$

$W1$ holds $M2$'s proposal and rejects $M4$'s, so he proposes to his next choice, $W4$, and the next round is:

$$\begin{aligned} M1 &\rightarrow W2 \\ M2 &\rightarrow W1 \\ M3 &\rightarrow W4 \\ M4 &\rightarrow W4 \end{aligned}$$

$W4$ holds $M3$'s proposal and rejects $M4$'s, so he proposes to $W3$, and the stable matching is:

$$\begin{bmatrix} M1 & M2 & M3 & M4 \\ W2 & W1 & W4 & W3 \end{bmatrix}$$

And when women propose, the algorithm finds the same stable match. This is the unique stable matching.

Note that this match is the man-optimal stable match from the standard framework.

Now, let's explore strategic aspects. For example, does $W4$ still have an incentive to report $M4$ as being her only acceptable partner?

Suppose she ranks $M4$, 1, and the other three men, 8 (recall that the protocol for declaring someone unacceptable is to rank him $2N$).

Then the ranks matrix is:

$$ranks: \begin{bmatrix} 4 & 2 & 4 & 12 \\ 4 & 6 & 5 & 11 \\ 6 & 5 & 5 & 9 \\ 5 & 6 & 6 & 5 \end{bmatrix}$$

and the match found by the modified DAA is:

$$\begin{bmatrix} M1 & M2 & M3 & M4 \\ W2 & W1 & W3 & W4 \end{bmatrix}$$

But this lie is not profitable for $W4$, precisely because of interdependent utility–while $W4$ does rank $M4$ more highly than $M3$, her previous partner, $M4$ ranks her last. By contrast, $M3$ ranks her first, and she ranks him third. If $W4$ were to partner with $M4$, the output they would produce would be $\frac{1}{2}Y$, while if $W4$ partnered with $M3$, the output they would produce would be $\frac{2}{3}Y$. Hence, $W4$ actually prefers $M3$ as a partner over $M4$–$M3$ would contribute highly to their relationship, while $M4$ wouldn't contribute at all.

6 Conclusion

Dubins and Freedman [4] and Roth [10] showed that it's impossible to design a stable mechanism under which truthtelling is a dominant strategy for every agent, or even a best response for

every agent, assuming the others are telling the truth. Hence, if we wish to continue using stable mechanisms, research must focus on understanding why market participants do not misreport their preferences, even when they have opportunities to gain by doing so.

In this paper, I offer an explanation which departs from the existing work on computational complexity, incomplete information, and large markets. I show that the utility of makret participants may in fact be interdependent–an agent may care about a potential partner's ranking of her, in addition to her ranking of her partner. I show that when the interdependence is linear and desires and desirability are equally-weighted, there is either a unique stable matching, or multiple stable matchings across which all agents are indifferent. Thus, no agent has an incentive to lie under a stable mechanism, if the other participants are telling the truth.

There are several important caveats to these results, and some important directions for future research. In this paper, interdependence is modeled in a very specific way, and this leads to a partition of preference profiles which is specific to the structure of interdependence (that is, perfect antagonism is a concept which seems specific to this model). However, the intuition for uniqueness does not seem to be bound by the functional form. It would be of great interest to generalize the concept of interdependent utility.

Experimental work would benefit this research greatly. There is no stable mechanism under which truthtelling is a best response to others telling the truth in the standard model, but I've shown that there is such a truth-inducing stable mechanism if agents care sufficiently about their partners' perceptions. It would therefore be of interest to run the following simple one-to-one, two-sided matching experiment: in the first round, each participant in the experiment is shown the profiles of characteristics of all the members in the other group, and asked to submit their rankings of the members in the other group. Then, it is revealed to each participant i how she was ranked by each member of the other group. Agents would then be allowed to re-submit their preference lists. Are there differences between the original and the updated preference lists? Do the differences appear to have a special structure?

Currently, our way of dealing with the possibility of strategic manipulation is to push two-sided problems to be somewhat one-sided. In the NRMP, ony residents submit rankings over hospitals, while in student-school matching problems, only students are allowed to rank schools. (Hospitals and schools are still allowed to set quotas.) This somewhat supresses the two-sidedness of the problem, and induces residents and students to report truthfully. However, if our hope is for market design to expand and find new applications, then we need to have a better understanding of strategic manipulation, and better ways of dealing with it. This paper suggests that, contrary to existing explanations where agents have an incentive to misreport but cannot bear the cost of calculating the optimal lie, or do not have enough information, or gain little from it, it may be the case that agents do not have an incentive to misreport after all. Agents know that the partners they could get by lying are precisely those that they desire but who do not desire them, and they may recognize that such a partnership might not be so satisfying. This paper identifies one structure of interdependence in which the stable matching is unique, and constructs a mechanism to find that

stable matching. Given how powerful interdependence can be, a fruitful next step would be to try and experimentally identify the structure of interdependence in real-life matching problems, and then to incorporate these observations into further theoretical research.

References

[1] Becker, G.S. "A theory of marriage: Part I." Journal of Political Economy (81) 1973 813-846.

[2] Chakraborty, A., A. Citanna, and M. Ostrovsky. "Two-sided matching with interdependent values." Journal of Economic Theory (145) 2010 85-105.

[3] Chung, K.S. "Implementing stable matchings under incomplete information." PhD Dissertation, Chapter 2 1999.

[4] Dubins, L. and D. Freedman. "Machiavelli and the Gale-Shapley algorithm." American Mathematical Monthly (88) 1981 485-494.

[5] Gale, D. and L. Shapley. "College admissions and the stability of marriage." American Mathematical Monthly (69) 1962 9-15.

[6] Gale, D. and M. Sotomayor. "Ms. Machiavelli and the stable matching problem." American Mathematical Monthly (92) 1985 261-268.

[7] Kara, T. and T. Sonmez. "Nash implementation of matching rules." Journal of Economic Theory (68) 1996 425-439.

[8] Kojima, F. and P. Pathak. "Incentives and stability in large two-sided matching markets." American Economic Review (99) 2009 608-627.

[9] Mumcu, A. and I. Saglam. "One-to-one matching with interdependent preferences." MPRA Paper No. 1908 2007.

[10] Pini, M., F. Rossi, K. Venable, and T. Walsh. "Manipulation and gender neutrality in stable marriage procedures." AAMAS '09: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems 2009 655-672.

[11] Roth, A. "The economics of matching stability and incentives." Mathematics of Operations Research (7) 1982 617-628.

[12] Roth, A. "Misrepresentation and stability in the marriage problem." Journal of Economic Theory (34) 1984 383-387.

[13] Roth, A. "Matching with incomplete information." Games and Economic Behavior (1) 1989 191-209.

[14] Roth, A. "The economist as engineer: Game theory, experimentation, and computation as tools for design economics." Econometrica (70) 2002 1341-78.

[15] Roth, A. and U. Rothblum. "Truncation strategies in matching markets–in search of advice for participants." Econometrica (67) 1999 21-43.

[16] Roth, A. and M. Sotomayor. "Two-sided matching." Cambridge University Press, Cambridge, MA, 1990.

[17] Shapley, L. and M. Shubik. "The assignment game I: The Core." Intl. J. Game Theory (1) 1971 111-130.