ESSAYS IN INFORMATION ELICITATION AND MARKET DESIGN

BY

BLAKE RILEY

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Steven R. Williams, Chair
Professor Dan Bernhardt
Professor Roger Koenker
Professor Martin Perry

ABSTRACT

This dissertation consists of three essays in microeconomic theory. The first two focus on how to elicit information about the state of the world from strategic agents, either to make a decision or for its own sake. The third studies a model of decentralized two-sided matching markets.

In “Mechanisms for making accurate decisions in biased crowds”, I study decision rules for finding the true answer to a binary question using the opinions of biased agents. Taking majority rule as a baseline, I study peer-prediction decision rules, which ask agents to predict the opinions of others in addition to providing their own. Incorporating first-order beliefs into the decision rule has the potential to recognize the correct answer even when the majority is wrong. However, I show the majority rule is essentially the only deterministic, neutral, anonymous, and interim dominance solvable mechanism. I then characterize all randomized peer-prediction mechanisms with these properties, using this result to show majority rule is the optimal mechanism in this class. Finally, I consider a simple, non-incentive-compatible decision rule based on the median prediction that implements majority rule when all agents are strategic and improves on majority rule when an unknown subpopulation is honest.

In “Minimum truth serums with optional predictions”, I introduce a class of mechanisms for eliciting private correlated signals from a group of expected score maximizers without external verification or knowledge about the agents’ belief structure. Built on proper scoring rules, these *minimum truth serums* ask agents to report a signal and a prediction of the signals of others. If two agents with the same signal have the same expectations about the signals of others, the Bayesian incentive compatibility of these mechanisms follows with no further assumptions on the agents’ belief structure. With a slight modification, the mechanism is still feasible and incentive compatible when the prediction portion of the report is optional.

In “Uncoordinated two-sided matching markets”, I study a decentralized proposal model in joint work with Juan Fung. The study of two-sided matching markets is now a major subfield of market design, focused primarily on the variants of the deferred acceptance algorithm. As a centralized mechanism, deferred acceptance is guaranteed to return a stable match. However, there is little definite work on whether uncoordinated agents find a stable matching on their own and the consequences if not. We show that small to moderately large uncoordinated markets reach a stable match within n^2 proposals from each agent when the proposal strategy isn't completely naive. We also show that stopping the proposal process early before stabilizing results in a more egalitarian and higher welfare match, particular when the two sides of the market are unbalanced. This suggests uncoordinated markets wouldn't benefit from centralization unless there is an obvious failing like market unraveling.

To anyone curious.

ACKNOWLEDGMENTS

Many thanks to my advisor Steven Williams for his encouragement and the hours spent talking in his office about books, markets, math, and computer science. Further thanks to my committee members Dan Bernhardt for pushing me to grow, Roger Koenker for being a model scholar and making delightful literary references, and Martin Perry for his belief in my potential.

This journey would be incomplete without many friends: Dongwoo Kim, Rachel Shafer, and Jiaying Gu for their companionship during the core; Christine Jachetta, Richard Hislop, Sarah Miller, Andreas Hagemann, JW Schneider, David Quigley, Tyler Jachetta, and Jonathan Treffkorn for Thursdays (and once upon a time, Mondays) at the Blind Pig; Juan Fung, Tom Sahajdack, again Rachel Shafer, and others for our micro theory reading group; Sarge, Honey Bear, Teej, Point, Ruthless, Fox, Coyote, Ron Julio, Valhalla, and many many others for keeping me sane but weird. Particular thanks again to Juan Fung for his friendship and insight.

I appreciate my parents' time, attention, and love throughout the years. Finally, I am incredibly grateful for my wife Anjanette, her constant support, her discovered interest in economic principles, and her valiant efforts to understand the work presented here.

TABLE OF CONTENTS

CHAPTER 1

MECHANISMS FOR MAKING ACCURATE DECISIONS IN BIASED CROWDS

In October 2015, a team of inmates from Eastern New York Correctional Facility debated a three-time national champion team from Harvard. After hearing arguments, the panel of judges decided by majority vote in favor of the inmates[1]. Majority rule is a natural way to make group decisions like this one for multiple reasons. First, it is simple and transparent. Second, the procedure does not favor one side over the other or make distinctions between judges. Finally, it is strategically robust, making honest revelation a dominant strategy if agents want the final decision to match their personal opinion.

Although judges might vote to favor their opinion, the point of the competition is to decide which team is most skilled, not to aggregate the judges' preferences. The choice of winner should accurately reflect which team did better (according to some criteria) for the debate to be legitimate. However, the only way to identify which team is most skilled is through the judges' subjective assessments, and a judge's opinion could be correct or mistaken relative to the underlying truth. Since individual judges can be mistaken, the goal when aggregating opinions is to maximize the probability of choosing the most deserving team.

Information aggregation through voting is a long-studied question initiated by the Marquis de Condorcet in 1785 in his essay on majority decisions. The standard model following Condorcet assumes agents have noisy signals about the true state that are correct more often than not. An example would be debate judges who are 70% likely to vote for Harvard when the Harvard team is in fact better and 30% likely to vote for Harvard when the inmates are actually better. Under Condorcet's model, majority rule is more accurate than any given judge, smoothing out noise in opinions to identify the "wisdom of crowds." However,

[1] Leslie Brody, "Prison vs. Harvard in an Unlikely Debate," *The Wall Street Journal*, Oct. 8, 2015. <http://www.wsj.com/articles/an-unlikely-debate-prison-vs-harvard-1442616928>

aggregating noisy signals through majority rule can make matters worse when bias is present. Given the associations that come with Harvard undergraduates versus inmates convicted of violent crimes, a fair assessment is a lot to ask of a judge. Bias wouldn't be surprising and could go in either direction—discounting inmates because of their background or favoring them as underdogs.

Some degree of bias isn't fatal to the performance of majority rule. For instance, suppose each judge is 60% likely to favor Harvard in the state of the world where they are best and 90% likely to favor the inmates when they are best. This is a scenario where the judges are more impressed by a "good" team of inmates than a "good" team from Harvard. Nevertheless, when these opinions are aggregated, the group decision still favors the best team in each state of the world, with the only difference being Harvard wins by a smaller margin.

In contrast, suppose the bias is stronger and judges are 40% likely to correctly favor Harvard and 90% likely to correctly favor Eastern Correctional. The opinions are still correlated with the truth—comparatively more judges favor Harvard when they are best. Nonetheless, the Harvard supporters will be in the minority on average in each state. Majority rule will choose the inmates regardless of the truth.

Debate organizers concerned about potential bias could pick a decision procedure other than majority rule. However, doing so would require insight into the precise nature of the bias. For instance, a unanimity rule where the Harvard team wins only if all judges support them would counteract a bias towards Harvard, but would also exacerbate a bias towards the inmates. Debate organizers might not trust themselves to adjust the decision rule in the right direction, and the teams would be understandably upset at the asymmetric standard even if they did. A satifactory alternative to majority rule needs to be *neutral*, treating each option symmetrically.

Furthermore, groups like corporations, unions, or homeowners' associations need a single rule that can be applied consistently across different contexts. Achieving this can be difficult given that the degree of bias may change how we should interpret a particular level of support for one team over another. For instance, if judges are biased towards Harvard, Harvard may deserve to lose even with a two-thirds majority. A single rule responsive to different circumstances has to collect additional information from agents beyond their opinions. In particular, a decision rule could ask agents to predict the opinions of other group members. By comparing the actual level of support with the predicted level of support, a “peer-prediction” decision rule can potentially make more accurate decisions than majority rule without knowing the likelihoods of opinions in each state.

Consider the following example: three judges are independently and identically 40% likely to correctly favor the Harvard team and 90% likely to correctly favor the inmates. Each judge puts equal prior probability on either team being best and updates their beliefs after observing their own opinion using Bayes’ rule. Let the opinion of judge i be x_i and the best team be ω. Under majority rule, the Harvard team wins with probability

$$\begin{aligned}\Pr[\text{Majority for Harvard} \,|\, \omega = \text{H}] &= \Pr[\text{Three Harvard supporters} \,|\, \omega = \text{H}] \\ &\quad + \Pr[\text{Two Harvard supporters} \,|\, \omega = \text{H}] \\ &= \left(\frac{4}{10}\right)^3 + 3\left(\frac{4}{10}\right)^2 \frac{6}{10} = 0.35\end{aligned}$$

when they’re best, and the inmate team wins with probability 0.97 when they are best.

Rather than use majority rule, let’s say Harvard wins if the percentage in support of Harvard is greater than the average predicted support for the team. For example, if the average predicted support for Harvard is 70%, then Harvard would win with 80% support and lose with 60%, despite still being favored by the majority. Framing the rule in terms of Eastern Correctional would produce identical decisions, so this rule is neutral. Conditional on their opinion, a Harvard supporter expects another judge to support Harvard with probability

$$\begin{aligned}\Pr[x_j = \text{H} \,|\, x_i = \text{H}] &= \Pr[x_j = \text{H} \,|\, \omega = \text{H}] \cdot \Pr[\omega = \text{H} \,|\, x_i = \text{H}] \\ &\quad + \Pr[x_j = \text{H} \,|\, \omega = \text{EC}] \cdot \Pr[\omega = \text{EC} \,|\, x_i = \text{H}] \\ &= \frac{4}{10} \frac{\frac{4}{10}\frac{1}{2}}{\frac{4}{10}\frac{1}{2} + \frac{1}{10}\frac{1}{2}} + \frac{1}{10} \frac{\frac{1}{10}\frac{1}{2}}{\frac{4}{10}\frac{1}{2} + \frac{1}{10}\frac{1}{2}} = 0.34\end{aligned}$$

Similarly, an Eastern Correctional supporter expects others to support Harvard with probability 0.22. Assume all judges reveal their beliefs honestly. Since the average predictions for Harvard are 34% with three supporters, 30% with two, 26% with one, and 22% with zero, Harvard wins unless the judges unanimously support the inmates. Under this rule, the probability of Harvard winning conditional on being best is

$$\begin{aligned}\Pr[\text{Harvard support} > \text{predicted}\,|\,\omega = \text{H}] &= \Pr[100\%\text{ support} > 34\%\text{ predicted}\,|\,\omega = \text{H}]\\ &\quad + \Pr[66\%\text{ support} > 30\%\text{ predicted}\,|\,\omega = \text{H}]\\ &\quad + \Pr[33\%\text{ support} > 26\%\text{ predicted}\,|\,\omega = \text{H}]\\ &= \left(\frac{4}{10}\right)^3 + 3\left(\frac{4}{10}\right)^2\frac{6}{10} + 3\frac{4}{10}\left(\frac{6}{10}\right)^2 = 0.78\end{aligned}$$

and 0.73 for the inmates. The prior probability of this peer-prediction rule making the correct decision is then

$$\begin{aligned}&\Pr[\omega = \text{H}]\cdot\Pr[\text{Harvard support} > \text{predicted}\,|\,\omega = \text{H}]\\ &\quad + \Pr[\omega = \text{ENYCF}]\cdot\Pr[\text{ENYCF support} > \text{predicted}\,|\,\omega = \text{EC}]\\ &= \frac{1}{2}0.78 + \frac{1}{2}0.73 = 0.75\end{aligned}$$

compared to 0.66 for majority rule, producing more accurate decisions on average in addition to being fairer between states.

While promising, a problem remains with this particular peer-prediction rule: it's not incentive compatible if judges want the decision to match their own opinion. Incentive compatibility guarantees participants will report honestly even if they're willing to misreport. A judge would benefit from strategically claiming all others will have the opposite opinion. If a Harvard supporter predicts no other judges will favor Harvard, the average prediction will be lower, making it possible to secure a win with fewer supporters. Analogously, an inmate supporter maximizes the chances of an inmate win by predicting unanimous support of Harvard from the others. Under this strategy, the averaged predictions of how many other judges favor Harvard are:

Harvard support	Inmate support	Average predictions	% Harvard support
0	3	100%	0%
1	2	66%	33%
2	1	33%	66%
3	0	0%	100%

The percentage of Harvard supporters is greater than the average "prediction"(i.e. the condition for a Harvard win under this proposed peer-prediction rule) if and only if a majority favors Harvard. Because of strategic reporting, the outcome becomes identical to majority rule, and the potential benefits of using peer-prediction evaporate.

In this paper, I investigate whether incentive-compatible peer-prediction decision rules exist that are more accurate than majority rule. I require candidate peer-prediction rules to be *neutral*—symmetric between the two choices—and *anonymous*—symmetric between group members—like majority rule.

Different types of incentive compatibility constraints provide different guarantees for when a participant will truthfully reveal their information. Bayesian incentive compatibility is a standard requirement but is acknowledged to carry strong assumptions. A more robust alternative is dominant-strategy incentive compatibility. Majority rule, for instance, is dominant-strategy incentive compatible because voting for one team always makes it more likely they'll win. However, requring dominant-strategy incentive compatibility makes it impossible for the decision to incorporate predictions. Instead, I rely on an intermediate form of incentive compatibility based on iterated deletion of weakly interim-dominated strategies. A decision rule is *robustly implementable* if honest relevation survives this process of iterated deletion.

The paper proceeds as follows. Section 1.1 provides related literature. Section 3.2 describes the model and design objective. In section 1.3, I show predictions can play almost no role in deterministic, neutral, anonymous, and robustly implementable decision rules. As long as every agent thinks it's possible another agent holds the opposite opinion, the decision matches majority rule. Section 1.4 provides a characterization of randomized, neutral, anonymous, and robustly implementable decision rules in terms of a common functional form that varies only with the choice of two non-decreasing functions and two real numbers. In section 1.5, I numerically search for the optimal randomized mechanism using the analytical characterization of the previous section. Although randomized decision rules can non-trivially depend on agents' predictions, majority rule outperforms all rules that incorporate predictions. Despite the promise of peer-prediction rules for identifying the true state more frequently, these results show majority rule can't be beaten subject to incentive constraints.

However, since it is plausible some agents are willing to give sincere predictions, section 1.6 considers non-incentive-compatible rules that make more accurate decisions than majority rule when some agents are unconditionally honest and become equivalent to majority rule when all agents are strategic. For instance, one simple rule based on a weighted combination of the percentage in support and the median prediction makes 25-50% fewer mistakes than majority rule when half of the participants report honestly and half report strategically. Finally, section 3.4 concludes.

1.1 Related literature

Extensive work has been done to answer when groups can make correct decisions through voting procedures and when information can be elicited from strategic agents. Research on the accuracy of collective decisions dates to the Marquis de Condorcet's essay on majority rule. Condorcet's jury theorem now has many different forms (Grofman et al. 1983). In its standard version, it says majority rule is almost certain to choose the correct state as the number of agents voting grows large. Furthermore, simple majority rule is the optimal decision rule when each state has equal prior probability and agent's opinions are distributed identically and independently conditional on the state (Nitzan and Paroush 1982).

Across various extensions, the critical assumptions of the Condorcet jury theorem are that the average voter is more likely to favor the correct state than not and preferences do not change conditional on being the pivotal voter. Austen-Smith and Banks (1996) reconsider the second assumption, showing that sincere voting is typically not equilibrium behavior when agents have aligned preferences for the decision to match the true state. Following work on strategic voting has primarily focused on comparing particular voting rules, often reaching the conclusion that requiring unanimity is worse than simple majority or any supermajority (Feddersen and Pesendorfer 1998; Gerardi 2000; Duggan and Martinelli 2001). In this paper, I take a mechanism design approach to address violations of the first assumption while retaining the second.

A parallel line of research on eliciting information from strategic agents with differing preferences was initiated by Crawford and Sobel (1982). Many papers have considered elicitation from groups of experts, including Austen-Smith (1993); Feddersen and Pesendorfer (1997); Krishna and Morgan (2001); Battaglini (2004). Of particular relevance, Li et al. (2001); Wolinsky (2002); Glazer and Rubinstein (2004); Gerardi et al. (2009); Chwe (2010) take a mechanism design approach. Each of these considers implementation in Bayes-Nash equilibrium in contrast to my approach based on interim dominance-solvability and a lack of common knowledge about preferences or the information structure.

Peer-prediction mechanisms have been studied in the context of eliciting correlated private signals from groups of payment maximizers without preferences over the conclusions drawn from the collected information. Prelec (2004)'s *Bayesian truth serum* elicits signals in Bayes-Nash equilibrium even when the principal has no knowledge of the common prior or signal likelihoods, though the result holds only for a sufficiently large number of participants that depends on the unknown prior. Witkowski and Parkes (2012a) construct a variant of Prelec's mechanism that is incentive compatible for finite participants in the case of binary questions. Zhang and Chen (2014) and Riley (2014) provide detail-free mechanisms that are Bayesian incentive compatible for finite participants and any number of signals with arbitrary correlation structure. To my knowledge, this is the first paper to consider a peer-prediction mechanism without transfers.

The Bayesian truth serum scores also function as an anonymous and neutral decision rule that asymptotically chooses the correct state when agents are Bayesians with conditionally IID signals and a common prior, even in the presence of statistical bias (Prelec et al. 2014). However, this decision rule is not incentive compatible if agents have preferences over the result. Since the mechanism chooses the answer with the highest average score and scores can be unboundedly negative, a single agent can unilaterally force one answer off the table even if all others are honest. Although I consider non-incentive-compatible decision rules in this paper, my mechanisms dampen the influence of strategic behavior.

1.2 Model

A group of n agents face a decision between two choices A and B. The state $\omega \in \{A, B\}$ denotes the "correct" decision according to some standard, such as the most skilled of two competitors, the action that will maximize profits, or the true answer to a question. Where convenient, let the states have values $A = 1$ and $B = 0$. From the mechanism designer's perspective, the two states have equal prior probability.

Each individual i has an *opinion* $x_i \in \{a, b\}$ about the state and a *prediction* $p_i \in (0, 1)$ about the proportion of other agents who hold opinion a. In a slight abuse of notation, let x_i also be an indicator variable with values $x_i = 1$ if i holds the a opinion and $x_i = 0$ if i holds the b opinion. Let $n_a = \sum_i x_i$ be the number of participants stating opinion a, $n_b = n - n_a$ be the number of participants stating opinion b, and $\bar{x} = n_a / n$ be the proportion of respondents with opinion a. Let $\bar{x}_{-i} = \sum_{j \neq i} x_j / (n-1)$ be the proportion of agents other than i with opinion a.

Opinions are distributed independently conditional on the state with likelihoods $q_A = \Pr(x_i = a \,|\, \omega = A)$ and $q_B = \Pr(x_i = a \,|\, \omega = B)$. The likelihoods satisfy $q_A > q_B$, so opinions are positively correlated with the corresponding state but are otherwise unknown to the mechanism designer.

The prediction $p_i = \mathrm{E}_i[\bar{x}_{-i} \,|\, x_i]$ summarizes agent i's subjective beliefs about the opinions of others, and will be treated as a random variable distributed independently conditional on x_i from the perspective of the mechanism designer. Although I view agents symmetrically, the agents themselves can have arbitrary beliefs consistent with their predictions, seeing correlations between individuals or thinking particular agents are more likely to hold a position. For example, $p_i = 0.5$ is consistent with believing all other agents are equally and independently likely to hold either opinion, with others being perfectly correlated and equally likely to hold each opinion, or with half of the agents holding one opinion with certainty and half holding the other with certainty. I make no assumptions about higher-order beliefs.

Since agents can see correlations or distinctions between others, predictions aren't required to be consistent with Bayesian updating based on my specification. However, for predictions to retain some connection to the underlying state, I assume agents treat their opinions as IID signals on average, holding a "prior prediction" between the two likelihoods that is then updated upward upon observing $x_i = a$ or downward for $x_i = b$ plus some noise. In particular, I model predictions as normally distributed on a logistic scale:

$$\begin{aligned} &\ln\left(\tfrac{p_i}{1-p_i}\right) \sim \mathrm{Normal}\left(\mu_{x_i}, \sigma^2\right) \quad \text{s.t.} \qquad (1.1) \\ &\mu_a = \mu + \gamma, \quad \mu_b = \mu - \gamma \\ &\mu = \alpha \ln\left(\tfrac{q_A}{1-q_A}\right) + (1-\alpha) \ln\left(\tfrac{q_B}{1-q_B}\right) \end{aligned}$$

for some parameters $\alpha \in [0,1]$ and $\gamma \in \mathbb{R}_{++}$, which can be interpreted as the prior belief that $\omega = A$ and the amount of evidence participants consider their own opinion to be, respectively. The distribution of agent predictions comes into play when numerically evaluating the accuracy of mechanisms, so the choice of distribution can change the level of performance but doesn't substantively affect results.

1.2.1 Peer-prediction decision rules

A peer-prediction decision rule T for n agents takes opinions and predictions as inputs to produce a choice between the two states. Decision rules can be deterministic or randomized. A deterministic decision rule has output $T(x,p) \in \{A, B, \varnothing\}$, where $\varnothing$ is a "null choice" that can be used in situations with exact ties. A randomized decision rule has output $T(x,p) \in [0,1]$ denoting the probability A is chosen.

I focus on neutral and anonymous decision rules, retaining the properties of majority rule that no bias is built in towards either state or the opinion of any individual:

Definition 1.1 (Neutrality)**.** *A mechanism $T(x,p)$ is neutral if relabeling states A and B results in the complement of T, i.e. $T(x,p) = 1 - T(1-x, 1-p)$ for all x and p.*

Definition 1.2 (Anonymity)**.** *A mechanism $T(x,p)$ is anonymous if relabeling agents does not change T, i.e. $T(x,p) = T(\sigma(x), \sigma(p))$ for all permutations σ.*

The mechanism designer's objective is to maximize the probability the decision matches the true state:

$$\max_T \Pr[T(x,p) = \omega] \tag{1.2}$$

or equivalently

$$\min_T \mathrm{E}[|\omega - T(x,p)|]. \tag{1.3}$$

Each agent prefers the decision to match their own opinion. In particular, an agent with $x_i = a$ chooses a report (x_i', p_i') to solve

$$\max_{(x_i', p_i')} \Pr[T((x_i', x_{-i}'), (p_i', p_{-i}')) = A] \tag{1.4}$$

based on their conjecture about the reports (x'_{-i}, p'_{-i}) of others. Agents with $x_i = b$ then minimize the above objective.

1.2.2 Robustly implementable mechanisms

Mechanism design involves finding a procedure for collecting messages from agents and aggregating the reports into the desired outcome for each type profile while respecting the incentives of each participant. In general, a mechanism $\mathscr{M} = (M, g)$ consists of a space of message profiles M and an outcome function $g : M \to A$, where A is the set of possible outcomes. A mechanism implements T when the outcome of the induced game under some solution concept matches T.

Peer-prediction mechanisms have a message space where agents report an opinion and a probability distribution over the opinions of others. A peer-prediction mechanism can be seen as a "semi-direct" mechanism, asking agents to report a portion of their type rather than their full type, including a hierarchy of higher-order beliefs. The Bayesian truth serum (Prelec 2004) is a leading example of a peer-prediction mechanism. This mechanism has truth-telling as a Bayes-Nash equilibrium for sufficiently large groups of payment maximizers with an unknown common prior. The average difference in group scores can distinguish the true answer asymptotically, even with in the presence of false consensus (Prelec and Seung 2007). However, existing peer-prediction mechanisms assume agents care only about payments, not about influence. If agents have preferences over the aggregate score used to estimate the state, the Bayesian truth serum becomes highly manipulable.

Additionally, existing peer-prediction mechanisms depend on agents sharing a common prior [2]. While common priors are often singled-out as unrealistic, a possibly more concerning feature is that agents receive a single signal with agreed-upon conditional likelihoods. Realistically, each expert has seen evidence of various levels of strength that he may or may not have updated on properly, which points toward some form of robust implementation beyond Bayes-Nash equilibrium.

Standard notions of robust implementation include implementation in dominant-strategy or ex-post Nash equilibrium. However, any mechanism that

[2] Unless priors and posteriors can be elicited separately, as in Witkowski and Parkes (2012b).

makes one strategy a best response regardless of the types of others can't be sensitive to predictions in equilibrium. Two agents with the same opinion and different predictions have the same preferences ex-post, so the same outcome will be assigned when the two behave identically. Independence from higher-order beliefs is usually seen as a benefit, but comes at the cost of ruling out peer-prediction mechanisms before we even begin. Instead, I'll consider a peer-prediction mechanism to be robustly implementable if honest reporting is the dominance solvable outcome of the mechanism, surviving iterated deletion of weakly interim dominated strategies. Throughout the paper, I will rely on only two stages of strategy deletion.

Definition 1.3 (Weak interim dominance)**.** *A strategy m_i weakly interim dominates m_i' for an agent of type (x_i, p_i) if*

$$\int\int u_i(x_i, g(m_i, m_{-i}))\, d\phi(m_{-i} \,|\, x_{-i}, p_{-i})\, d\pi(x_{-i}, p_{-i}) \geq \tag{1.5}$$

$$\int\int u_i(x_i, g(m_i', m_{-i}))\, d\phi(m_{-i} \,|\, x_{-i}, p_{-i})\, d\pi(x_{-i}, p_{-i}) \tag{1.6}$$

for all beliefs π (a distribution over type profiles of others) and ϕ (a distribution over strategy profiles conditional on type profiles) such that $\mathrm{E}_\pi[\bar{x}_{-i}] = p_i$ to be consistent with i's type, with strict inequality for some beliefs.

Definition 1.4 (Dominance solvability)**.** *Given a mechanism $\mathscr{M} = (M, g)$, let $D_i^{\mathscr{M}}(x_i, p_i)$ be the set of strategies m_i that survive iterated deletion of all weakly interim dominated strategies at each stage for agent i of type (x_i, p_i). A mechanism is interim dominance solvable if $g(m) = g(m')$ for all profiles with $m_i, m_i' \in D_i^{\mathscr{M}}(x_i, p_i)$.*

Definition 1.5 (Robust implementation)**.** *A mechanism $\mathscr{M} = (M, g)$ robustly implements a peer-prediction mechanism T if the unique dominance solvable outcome when agents have types (x, p) is $T(x, p)$.*

Definition 1.6 (Robust incentive compatibility)**.** *A peer-prediction mechanism $T(x, p)$ is robustly incentive compatible if honesty is an interim best response for all conjectures about others' types consistent with the agent's prediction:*

$$\begin{aligned}\int T((a, x_{-i}), (p_i, p_{-i}))\, d\pi(x_{-i}, p_{-i}) &\geq \int T((x_i', x_{-i}), (p_i', p_{-i}))\, d\pi(x_{-i}, p_{-i}) \\ &\geq \int T((b, x_{-i}), (p_i, p_{-i}))\, d\pi(x_{-i}, p_{-i})\end{aligned} \tag{1.7}$$

for all x_i', p_i, p_i', and beliefs π such that

$$\mathrm{E}_\pi[\bar{x}_{-i}] = \int \frac{\#\{x_j = a \mid j \neq i\}}{n-1}\, d\pi(x_{-i}, p_{-i}) = p_i.$$

The following proposition provides a version of the revelation principle for this setting:

Proposition 1.1. *A mechanism $\mathscr{M} = (M, g)$ can robustly implement T only if T is robustly incentive compatible.*

Proof of Proposition 1.1 (Robust incentive compatibility). Suppose mechanism $\mathscr{M} = (M, g)$ robustly implements T, assigning outcome $g(m) = T(x, p)$ for each strategy profile $m \in \prod_i D_i^{\mathscr{M}}(x_i, p_i)$. Hence, given any $m_i \in D_i^{\mathscr{M}}(a, p_i)$ and $m_i' \in D_j^{\mathscr{M}}(x_j', p_j')$, we must have

$$\begin{aligned}
\int T((a, x_{-i}), (p_i, p_{-i}))\, d\pi(x_{-i}, p_{-i}) &= \int\int g(m_i, m_{-i})\, d\phi(m_{-i} \mid x_{-i}, p_{-i})\, d\pi(x_{-i}, p_{-i}) \\
&\geq \int\int g(m_i', m_{-i})\, d\phi(m_{-i} \mid x_{-i}, p_{-i})\, d\pi(x_{-i}, p_{-i}) \\
&= \int T((x_i', x_{-i}), (p_i', p_{-i}))\, d\pi(x_{-i}, p_{-i})
\end{aligned}$$

for all beliefs π (a distribution over type profiles of others) and ϕ (a distribution over strategy profiles conditional on type profiles) such that

$$\begin{gathered}
\mathrm{E}_\pi[\bar{x}_{-i}] = p_i \quad \text{and} \\
\Pr_\phi[m_{-i} \mid x_{-i}, p_{-i}] > 0 \implies m_{-i} \in \prod_{j \neq -i} D^{\mathscr{M}}(x_j, p_j)
\end{gathered}$$

since m_i either weakly dominates m_i' or is equivalent to it when agent i is type (a, p_i) and other agents play their dominance solvable strategies. This follows similarly for types (x_i', p_i') and (b, p_i), yielding the condition of robust incentive compatibility in line 1.7. □

Although the revelation principle provides some justification for restricting attention to incentive-compatible mechanisms, I will explore non-incentive-compatible decision rules that implement majority rule when all agents are strategic and outperform majority rule when some agents are honest later in the paper.

1.3 Deterministic decision rules

Consider decision rules which deterministically output a single state for any given profile. Even if first-order beliefs are included in reports, these turn out to play no functional role in the mechanism since the output is too coarse to respond to predictions. A robustly implementable, neutral, and anonymous decision rule can deviate from majority rule only when some agent mistakenly claims the realized profile was impossible:

Proposition 1.2. *If $T : \{a,b\}^n \times [0,1]^n \to \{A,B,\varnothing\}$ is a neutral, anonymous, and robustly implementable decision rule with $T(x) = \varnothing$ only if $\bar{x} = \frac{1}{2}$, then it agrees with majority rule on all profiles with interior predictions $p \in (0,1)^n$.*

The proof proceeds by showing profiles where agents correctly predict a bare majority must agree with majority rule and then expanding the set of profiles in agreement via incentive compatibility.

For an example of a deterministic decision rule where predictions do matter, consider a rule for three agents that maps all type profiles to the majority opinion except for

$$\begin{aligned}
&T((a,0),(a,0),(a,0)) = B,\\
&T((a,0),(a,0),(b,p_3)) = B \quad \forall p_3 \in (0,1],\\
&T((b,1),(b,1),(b,1)) = A, \text{ and}\\
&T((b,1),(b,1),(a,p_3)) = A \quad \forall p_3 \in [0,1),
\end{aligned}$$

as well as similar profiles for anonymity. This rule is neutral, anonymous, and robustly incentive compatible, so agreement with majority rule isn't required to extend to all profiles with extreme beliefs.

If decisions between the two states are randomized, the probability of choosing the A state in a neutral, anonymous, and robustly implementable mechanism is characterized in the next section.

1.4 Randomized decision rules

Randomized decision rules map report profiles into probabilities. Unlike deterministic rules, this set of decision rules can non-trivially incorporate predictions since there is more fine-grained control over the output.

As shown in the following theorem, all neutral, anonymous, and robustly implementable randomized rules for given n have a specific functional form that differ only by reference types $\phi_1, \phi_2 \in [\frac{1}{2}, 1]$ and nondecreasing functions τ and ξ. In this characterization, T can be decomposed into a *base score* (line 1.8) that depends solely on the proportion of agents endorsing a. The base score is adjusted by the mean *prediction scores* (line 1.9) of each agent, signed according to their opinion. The base score provides sufficient incentive for reports with a false opinion to be interim dominated. Conditioning on each player always wanting to honestly reveal their true opinions, agents will want to give their true prediction as long as their marginal influence is a proper scoring rule for the proportion of a endorsements. The parameters ϕ_1 and ϕ_2 describe prediction types where incentive constraints bind exactly. The function ξ weights prediction scores, controlling the magnitudes of rewards and punishments for prediction accuracy in each region of the unit interval.

This representation embeds the design constraints into the functional form, reducing the optimal mechanism design problem to a mildly-constrained search across ϕ_1, ϕ_2, τ, and ξ.

Proposition 1.3. *A neutral and anonymous peer-prediction randomized decision rule T is robustly implementable for n participants only if T can be represented as*

$$T(x,p) = \frac{1}{2} + \operatorname{sign}\left(\tfrac{n_a}{n} - \tfrac{1}{2}\right)\left(\tau\left(\left|\tfrac{n_a}{n} - \tfrac{1}{2}\right|\right) + \mathbb{1}(n\ odd)\frac{\delta\left(\frac{n-1}{2}\right)}{2n} + \frac{1}{n}\sum_{m=\lceil n/2\rceil}^{\max\{n_a,n_b\}-1} \delta(m)\right) \tag{1.8}$$

$$+\frac{1}{n}\sum_{i:x_i=a} R_\xi\left(p_i, \tfrac{n_a-1}{n-1}\right) - \frac{1}{n}\sum_{i\, x_i=b} R_\xi\left(1-p_i, 1-\tfrac{n_a}{n-1}\right) \tag{1.9}$$

$$\begin{aligned} s.t.\quad & \delta(m) = \max\left\{-R_\xi\left(\phi_1, \tfrac{m}{n-1}\right) - R_\xi\left(1, 1-\tfrac{m}{n-1}\right),\ -R_\xi\left(1-\phi_2, 1-\tfrac{m}{n-1}\right)\right\} \\ & R_\xi(p_i, \bar{x}) = \int_0^{p_i} (\bar{x} - t)\, d\xi(t) \end{aligned}$$

for $\phi_1, \phi_2 \in [\frac{1}{2}, 1]$ and non-decreasing functions $\xi : [0,1] \to \mathbb{R}_+$ and $\tau : [0, \frac{1}{2}] \to \mathbb{R}_+$. This representation is sufficient for robust implementation if τ is strictly increasing and the maximal output satisfies $T((a,1),\ldots,(a,1)) \leq 1$.

The requirement for sufficiency that τ be strictly increasing ensures incentives are strict, while the requirement that $T((a,1),\ldots,(a,1)) \leq 1$ ensures the output of T is always a proper probability contained in the unit interval.

1.5 Determining the optimal randomized mechanism

Using the representation stated in the previous section, I now investigate the optimal randomized decision rule. The mechanism design problem is to solve

$$\begin{aligned}&\min_{T} \mathrm{E}[|\omega - T(x,p)|]\\ &\quad \text{s.t. } T \text{ is neutral, anonymous, and robustly incentive compatible,}\end{aligned}$$

which is equivalent to

$$\begin{aligned}\min_{\phi_1,\phi_2,\tau,\xi} \;& \mathrm{E}[|\omega - T(x,p)|] =\\ &\Pr(\omega = A)\,\mathrm{E}[1 - T(x,p)\,|\,\omega = A] + \Pr(\omega = B)\,\mathrm{E}[T(x,p)\,|\,\omega = B]\\ \text{s.t. } & \phi_1,\phi_2 \in [\tfrac{1}{2},1] \text{ and } \xi : [0,1] \to \mathbb{R}_+, \tau : [0,\tfrac{1}{2}] \to \mathbb{R}_+ \text{ are non-decreasing.}\end{aligned}$$

Unfortunately, this problem isn't amenable to typical first-order solution methods. Corner solutions are likely since the objective function is linear in T and T is affine in τ and possibly ξ. When the objective is locally affine, first-order conditions at the boundary become trivial.

Since the conditional opinion likelihoods aren't known to the mechanism operator, a prior distribution over likelihoods must be specified. Some natural distributions of likelihoods include:

1. Uniform over all likelihood pairs (q_A, q_B) with positive correlation, satisfying $q_A > q_B$
2. Those concentrated around the diagonal or in a band offset from the diagonal
3. Uniform over all unbiased likelihood pairs, satisfying $q_B \leq 0.5 \leq q_A$
4. Uniform over all biased likelihood pairs, satisfying $q_B < q_A \leq 0.5$ or $0.5 \leq q_B < q_A$

The first and second possibility can be interpreted as each agent independently knowing the true state with probability λ and otherwise having opinion a with probability $(1-\lambda)\gamma$ and opinion b with probability $(1-\lambda)(1-\gamma)$, with both probabilities unknown. The first corresponds to a uniform prior over both λ and γ.

The second corresponds to a normal distribution (restricted to the unit interval) over λ and a uniform distribution over γ, allowing for more precise information about the expertise of participants.

As noted earlier, I model predictions as normally distributed on a logistic scale for parameters $\alpha \in [0,1]$ and $\gamma \in \mathbb{R}_{++}$ corresponding to a prior prediction and an degree of adjustment, respectively. In particular, I assume $\alpha, \gamma \sim \text{Unif}[0,1]$. Then, taking expectations across parameters $\theta = (q_A, q_B, \alpha, \gamma) \in [0,1]^4$, the likelihood of types in a given state is

$$\begin{aligned} \mathrm{E}_\theta[\Pr(x, p \mid \omega, \theta)] &= \int_\theta q_\omega^{n_a}(1-q_\omega)^{n-n_a}\left(\prod_{i=1}^{n} f_{x_i}(p_i \mid \theta)\right) g(q_A, q_B)\, d\theta \qquad (1.10)\\ \text{s.t.} \quad f_{x_i}(p_i \mid \theta) &= \frac{1}{p_i(1-p_i)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{\sigma^2}\left(\mu_{x_i}(\theta) - \ln\left(\frac{p_i}{1-p_i}\right)\right)^2\right) \\ \mu_{x_i}(\theta) &= \alpha \ln\left(\frac{q_A}{1-q_A}\right) + (1-\alpha)\ln\left(\frac{q_B}{1-q_B}\right) + (2\mathbb{1}(x_i = a) - 1)\gamma. \end{aligned}$$

I set $\sigma^2 = 1$ to produce a realistic amount of dispersion without the distribution bunching around 0 and 1, which tends to occur when the variance grows larger. Figure 1.1 shows typical prediction distributions.

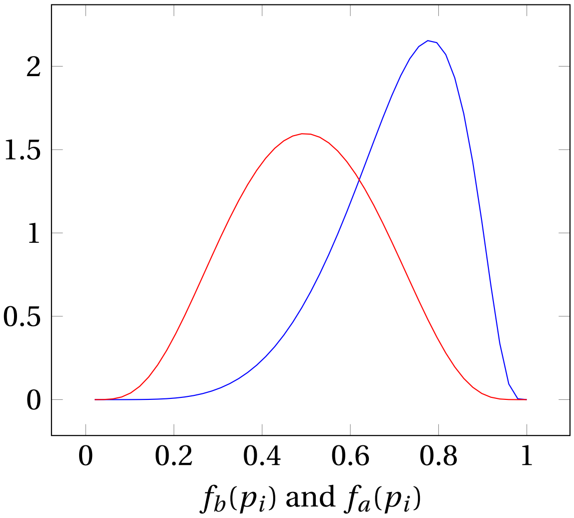

Figure 1.1: Prediction densities for agents with opinions a and b when $q_A = 0.8$, $q_B = 0.4$, $\sigma^2 = 1$, $\alpha = 0.5$, and $\gamma = 0.5$.

1.5.1 Representing decision rules numerically

As shown in Proposition 1.3, optimization over the class of robustly implementable mechanisms involves a search over three components: the scoring rule weighting function $\xi(t)$, the reference types ϕ_1, ϕ_2, and the extra base score $\tau(t)$. For a numerical solution, I approximate this infinite-dimensional problem with a finite-dimensional representation.

In full generality, the weighting function ξ used to parameterize the scoring rule $R_\xi(p_i, \bar{x}_{-i})$ can be any non-decreasing function with a domain of $[0,1]$. I approximate a general ξ by decomposing it into a continuously differentiable function and a step function, producing a scoring rule

$$R_\xi(p_i, \bar{x}_i) = \int_0^{p_i} (\bar{x}_{-i} - t)\xi'(t)\,dt + \sum_{k=1}^{K_\xi} \lambda_k \mathbb{1}(p_i \geq t_k)(\bar{x}_{-i} - t_k). \tag{1.11}$$

On the discrete portion, ξ has K_ξ atoms at points $t_k \in [0,1]$ with weights $\lambda_k \in \mathbb{R}_+$. On the continuously differentiable portion, I assume ξ' is piecewise linear with $H_\xi - 1$ segments at regular intervals, giving the integral a manageable closed form. The contribution to the total score on an interval $[t_1, t_2]$ where $\xi'(t)$ is linear is

$$\begin{gathered}\int_{t_1}^{t_2} (\bar{x}_{-i} - t)\left(\tfrac{\xi'(t_2)-\xi'(t_1)}{t_2-t_1}(t - t_1) + \xi'(t_1)\right) dt = \\ \frac{t_2 - t_1}{6}\big(3(\bar{x}_{-i} - t_1 - t_2)(\xi'(t_1) + \xi'(t_2)) - t_1\xi'(t_1) - t_2\xi'(t_2)\big).\end{gathered} \tag{1.12}$$

The score $R_\xi(p_i, \bar{x}_{-i})$ is the sum of this amount on each linear segment inside $[0, p_i]$, so ξ' can be parameterized by H_ξ values $\xi'_h \in \mathbb{R}_+$ at $0, 1/(H_\xi - 1), 2/(H_\xi - 1), \ldots, (H_\xi - 2)/(H_\xi - 1), 1$.

I also represent $\tau\left(\left|\frac{n_a}{n} - \frac{1}{2}\right|\right)$ using a continuous τ' with $H_\tau - 1$ linear segments and K_τ weighted atoms. Between densities parameters, atom locations, and atom weights for ξ and τ and the two reference types $\phi_1, \phi_2 \in [\frac{1}{2}, 1]$, the total parameter space is $H_\xi + 2K_\xi + H_\tau + 2K_\tau + 2$ dimensional.

1.5.2 Optimization methods

Using this finite-dimensional approximation, the scoring rule R_ξ is linear in the vectors of density values and atoms weights. The estimator $T(x, p)$ is then convex in these parameters when $\bar{x} > \frac{1}{2}$ and concave when $\bar{x} < \frac{1}{2}$ due to the changing sign on the maximum taken in $\delta(m)$. Since the estimator is neutral, we are always free to reassign labels to make a the majority opinion and $\bar{x} \geq 1/2$ so that the overall objective is convex in these parameters. The domain for each of these parameters is the entire positive real line, but since the objective diverges as any parameter diverges, the minimizer will be in some bounded interval.

The estimator is less well-behaved in terms of the reference types and atom position. A scoring rule is quasiconcave with ϕ as the prediction, and an atomic scoring rule $R(p_i, \bar{x}_{-i}) = \lambda_k \mathbb{1}(p_i \geq t_k)(\bar{x}_{-i} - t_k)$ is quasiconvex in t_k, but this won't necessarily aggregate up into quasiconvexity of the estimator or objective. The estimator is also discontinuous in these parameters, though the discontinuities will be smoothed out in expectation in the objective. Consequently, a global optimization procedure may be necessary to thoroughly search the parameter space.

The optimization problem is unconstrained aside from bounds on each parameter and possibly a constraint that the output is contained in the unit interval. For the output to be inside the unit interval, it is sufficient that the outcome when agents are unanimous and know they are unanimous satisfies $T((a, \ldots, a), (1, \ldots, 1)) \leq 1$. Values outside the unit interval are nonsensical for randomized decision rules. Values outside the unit interval are still undesirable for an estimator but might be acceptable if they occur only for nearly unanimous inputs, which we expect to be rare. After all, there is little reason to conduct a survey if an answer is obvious and everyone thinks it's obvious.

Optimization is done through the *Multi-level Single-linkage* global optimization algorithm, a multistart method that uses a clustering heuristic to avoid repeatedly returning to the same local minima on each local optimization. For local optimizations, I used Rowan (1990)'s *Subplex* algorithm, a variant of the Nelder-Mead simplex method done through a sequence of subspaces.

1.5.3 Optimal randomized decision rules don't use predictions

Unlike deterministic decision rules, randomized decision rules are able to incorporate predictions while remaining robustly implementable. However, randomized output typically hurts when maximizing the probability of a correct decision or minimizing the absolute deviation, so it's unclear whether the potential benefit is worth the cost.

Optimization over the class of robustly implementable peer-prediction decision rules returns a mechanism that depends only on opinions, using a τ with a single step and zeroing out ξ. This finding holds varying n and the prior on opinion likelihoods. Note the optimal randomized mechanism isn't necessarily majority rule. For some priors on opinion likelihoods (such as a uniform prior over all biased likelihood pairs), the optimal mechanism chooses A when $\bar{x}$ is sufficiently high, B when $\bar{x}$ is sufficiently low, and randomizes between them with equal probability when $\bar{x}$ is in an interval around $\frac{1}{2}$. If majority rule is the optimal randomized mechanism that uses only opinions, then it is also optimal in the class of peer-prediction mechanisms.

1.6 Simple peer-prediction rules with some sincere agents

The preceding results show majority rule is either the only robustly implementable decision rule or the only one worth considering, modified at most by randomizing in some interval around $\bar{x} = \frac{1}{2}$. Nevertheless, like most incentive-compatible direct mechanisms, the direct mechanism for majority rule takes strategic behavior for granted. Unlike an allocation setting, it is plausible that some agents are willing to unconditionally tell the truth and don't have preferences over the outcome. A non-incentive-compatible decision rule could implement majority rule when all agents are strategic and outperform it whenever some agents are sincere.

If all agents are Bayesians who think opinions are IID based on underlying likelihoods, all predictions will be inside the interval $[q_B, q_A]$. Without knowing the likelihoods themselves, a third party could easily conclude the state is likely to be A if the proportion of a opinions is higher than most predictions.

Although I allowed agent predictions as more dispersed, a similar identification of the state is possible in this setting. I model predictions as satisfying

$$\ln\left(\tfrac{q_B}{1-q_B}\right) < \mathrm{E}\left[\ln\left(\tfrac{p_i}{1-p_i}\right) \mid x_i = a\right] \quad \text{and} \quad \mathrm{E}\left[\ln\left(\tfrac{p_i}{1-p_i}\right) \mid x_i = a\right] < \ln\left(\tfrac{q_A}{1-q_A}\right)$$

which implies

$$q_B < \text{median}(p_i \mid x_i = a) \quad \text{and} \quad \text{median}(p_i \mid x_i = b) < q_A.$$

All else equal, we expect the state is more likely to be A when $\bar{x}$ is higher and when the proportion of a opinions is higher than the median group predictions, so one simple decision rule takes a linear combination of these magnitudes [3]:

$$\begin{aligned} T(x,p) = \mathbb{1}\big(\lambda_1(\bar{x} - \tfrac{1}{2}) \\ &+ \lambda_2(\bar{x} - \text{median}(p_i|x_i = a)) \\ &+ \lambda_3(\bar{x} - \text{median}(p_i|x_i = b)) > 0\big). \end{aligned}$$

For neutrality, we must have $\lambda_2 = \lambda_3$. For some partial incentive compatibility, the expression should have $\lambda_1 + \lambda_2 + \lambda_3 > 0$ to be increasing in $\bar{x}$. Under these constraints, the decision rule above is equivalent to

$$T(x,p) = \mathbb{1}\Big(\bar{x} + \tfrac{\lambda}{2}(1 - \text{median}(p_i|x_i = a) - \text{median}(p_i|x_i = b)) > \tfrac{1}{2}\Big).$$

This decision rule has majority rule as the unique dominance solvable outcome. Assuming $\lambda > 0$, all reports for an agent with $x_i = a$ are weakly dominated by either $(a,0)$ or $(b,0)$, depending on which group median the agent has the most influence over. Once all reports with interior predictions are eliminated, reporting one's true opinion becomes the unique weakly dominant strategy for each agent. The group medians cancel out, leaving only a comparison of $\bar{x}$ to $\frac{1}{2}$.

When all agents are sincere, this decision rule does quite well. For instance, when $n = 50$ and there is a uniform prior over opinion likelihoods, this rule for $\lambda = 0.9$ misclassifies the state approximately 13.5% of the time compared to 25% of time for majority rule.

The median is well-known as a robust location estimator, able to withstand up to 50% of the inputs being adversarially altered before becoming invalid. Suppose each agent is strategic with identical probability ρ and sincere otherwise. Since there are two weakly dominant strategies, it's not obvious what an agent will do when it expects only some agent to be strategic. If all strate-

[3] To avoid taking the median of an empty group, assume the output matches the unanimous opinion if all agents agree.

gic agents report their true opinion and a prediction $p_i \in \{0,1\}$, then the group medians quickly degrade to the extremes when $\rho > \frac{1}{2}$, reducing the decision to majority rule.

In contrast, consider the following even simpler decision rule:

$$T(x,p) = \mathbb{1}\left(\bar{x} + \lambda\left(\tfrac{1}{2} - \text{median}(p)\right) > \tfrac{1}{2}\right)$$

Call this the *median prediction rule*. Notice the decision is simply majority rule when $\lambda = 0$. When $\lambda > 0$, the unique weakly dominant strategy for the median prediction rule is for an agent with $x_i = a$ to report $(a, 0)$ and an agent with $x_i = b$ to report $(b, 1)$. Again assuming that agents have an IID chance of being strategic, the median of all predictions isn't influenced by strategic behavior until $\rho > 1 - \frac{\bar{x}}{2}$ since the inputs are being manipulated by two opposing groups of agents rather than a single-minded adversary.

Figures 1.2 and 1.3 depicts how the percentage of misclassified states for $n = 15$ and $n = 100$ respectively varies for different weights λ in the decision rule. This is shown for varying percentages of strategic agents. The optimal weight λ depends on the number of strategic agents, starting around $\lambda = 0.7$–0.8 for completely honest agents and increasing as agents become more strategic. The plots show the median prediction rule being more accurate on average for every choice of λ (except $\lambda < 1/n$, which is too small to change the decision from majority rule).

Figure 1.4 depicts the percentage of misclassified states as the percentage of strategic agents varies, with $\lambda \simeq 0.8$ optimized for $\rho = 0$ and $\lambda \simeq 0.95$ optimized for $\rho = 0.5$. As agents become more strategic and ρ increases to one, the median prediction rule agrees with majority rule more and more frequently.

While there is little reason to think the median prediction rule is optimal, it is simple and robust. Adding predictions to the group decision reduces the errors of majority rule due to bias and, at worst, becomes equivalent to majority rule when agents act strategically. Majority rule is still a useful means of aggregating preferences, but whenever the underlying goal is to aggregate information and it's conceivable that the majority can make the wrong choice, the median prediction rule is a strong alternative.

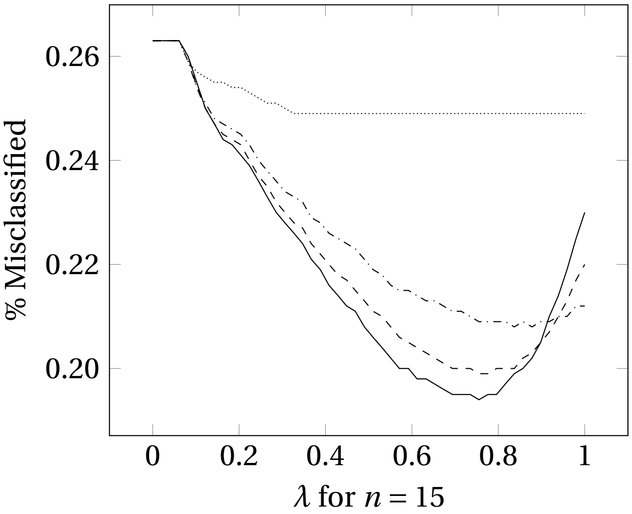

Figure 1.2: Effect of varying weight λ in the median prediction rule for percentage of strategic agents ρ in $\{0.0, 0.25, 0.5, 0.9\}$ in solid to dotted lines, respectively.

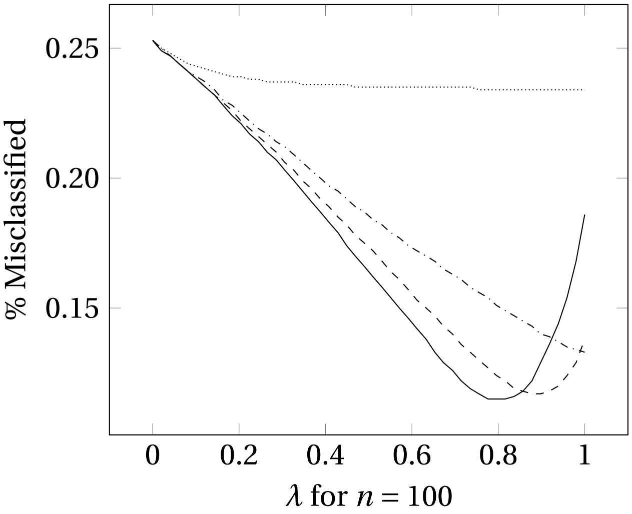

Figure 1.3: Effect of varying weight λ in the median prediction rule for percentage of strategic agents ρ in $\{0.0, 0.25, 0.5, 0.9\}$ in solid to dotted lines, respectively.

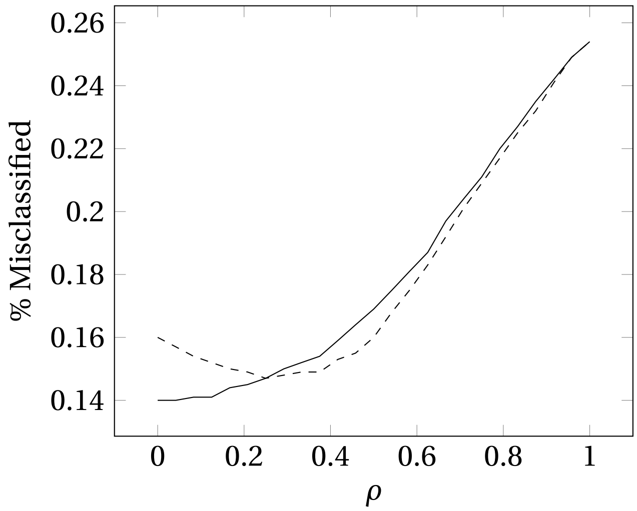

Figure 1.4: Percentage of incorrect decisions by the median prediction rule for $n = 50$ as the percentage of strategic agents ρ varies with $\lambda = 0.8$ solid and $\lambda = 0.95$ dashed.

1.7 Conclusion

My model takes a broad view of potential sources of bias, capturing two sources usually considered in isolation: preference-based bias and statistical bias. Agents with a preference-based bias have some stake in the conclusions drawn from their information. Agents are willing to distort or garble information themselves in order to influence the final results. This form of bias has been thoroughly investigated in the literature under the assumption of commonly known information and structure in order to facilitate analysis under Bayes-Nash equilibrium.

Alternatively, statistical bias is a systematic tendency for participants to hold a particular opinion besides the true state of the world, even if agents are sincere or have a common interest. Under this notion, opinions are biased in the sense that their likelihoods aren't symmetric across states of the world. In a mild form, 80% of the population might hold one opinion when it's correct, while only 60% of the population hold the opposite opinion when it's correct. In an extreme form, an opinion might be held by 90% of the population when it's correct and 75% of the population when it's incorrect, putting it in the majority regardless of the true state.

These two categories of bias are logically separate but are not easily separated in practice. Psychological notions of cognitive bias—defined as systematic deviations from some standard of judgment—can often be interpreted as a preference, and common usage conflates the two. The success of peer-prediction mechanisms can be seen as exploiting the false consensus effect identified in social and cognitive psychology (Marks and Miller 1987). Debate exists whether the false consensus effect is a cognitive bias or the rational consequence of updating beliefs about others conditional on one's own attributes (Dawes 1989), but either is compatible with my model.

I assume agents' preferences over conclusions do not change conditional on the reports of others, in contrast to the strategic voting literature where preferences can change dramatically after updating on others' information. Reality is somewhere in between, with people updating on the information of others at a discount relative to their own information (Yaniv and Kleinberger 2000). In the canonical example of a jury voting whether to convict a defendant, it's very plausible an agent would revise their opinion upon learning others are unanimous since a juror (ideally) doesn't have a personal connection to the question of guilt. Experimental work by Guarnaschelli et al. (2000) roughly supports the strategic voting model of Feddersen and Pesendorfer (1998), though the experiment sensibly asked participants to make the bland decision of which jar they drew colored balls from. In a more emotionally-charged situation like a committee decision wrapped up in office politics, I expect agents to stick to their opinions regardless of how others might report.

Possible future directions include experimental work testing the accuracy of these mechanisms, expanding the scope of the model beyond binary questions, and characterizing the equivalence class of peer-prediction decision rules that implement majority rule in a way amenable to optimizing accuracy in partially-strategic "equilibrium."

1.8 Computational details

The numerical results were computed in *Julia, v0.4.0* using the MLSL and SBPLX optimization algorithms of *NLopt.jl* and the h-adaptive integration algorithm of *Cubature.jl*, both implemented by Steven G. Johnson.

CHAPTER 2

MINIMUM TRUTH SERUMS WITH OPTIONAL PREDICTIONS

A central concern of mechanism design is how to collect private information from individuals while acknowledging their incentives. The prototypical mechanism design problem in economics involves eliciting preferences in order to allocate goods. However, individuals can have relevant private information besides preferences. For example, consider university officials interested in the prevalence of drinking among freshmen. Students might be inclined to misreport when asked about their drinking habits even if the survey is purely informational. Unlike the prototypical problem, honesty is important for its own sake rather than as a means to acheiving efficiency or maximum revenue, and a well-designed mechanism could provide positive rewards for honesty.

The Bayesian truth serum, introduced by Prelec (2004), was one of the first mechanisms for eliciting private information from a group of agents to any subjective, hypothetical, or unverifiable question. The mechanism operates by asking agents for an *information report* from a finite set of answers—corresponding to the agent's private signal—and a *prediction report* of the distribution of signals of other agents—corresponding to the agent's first-order posterior beliefs—and assigning scores to agents based on the collective reports. The original truth serum has many desirable properties, such as being

- *detail free*, requiring zero knowledge of the agents' common prior to determine scores,
- *interim individually rational*, giving each agent a non-negative expected payment conditional on their private information, and
- *collusion resistant*, with the truth-telling equilibrium being interim Pareto-dominant among all Bayes-Nash equilibria.

The mechanism can also be *ex-post budget balanced*, with total payments summing to zero for every possible realization of reports, though at the cost of individual rationality.

Many potential concerns about the robustness of the Bayesian truth serum remain. First, incentive compatibility holds only for a sufficiently large population, with the necessary size depending on the agents' unknown prior. Second, the mechanism could require arbitrarily large payments from the agents. Third, the assumption of common priors might be overly strong. Finally, requiring a prediction report in addition to an information report could make the mechanism too complex and unwieldy for some agents. Since Prelec's original work, multiple papers have sought to alleviate these concerns.

Contemporaneously with Prelec, the *peer-prediction mechanism* of Miller et al. (2005) depends only on an information report but is not detail-free, assuming precise knowledge of the posterior beliefs of agents for each signal. Jurca and Faltings (2011) investigate detail-free mechanisms that depend only on an information report, showing incentive compatibility is not possible in general, and develop instead a notion of *helpful reporting*. Jurca and Faltings (2007) and Carvalho and Larson (2012) discuss various forms of collusion resistance in peer-prediction mechanism.

Addressing Prelec directly, Witkowski and Parkes (2012a) introduce the *robust Bayesian truth serum* for the special case of binary signals, which is incentive compatible for $n \geq 3$ agents and has bounded payments. Under the additional assumption that agents' beliefs could be elicited both before and after receipt of their private signal[1], Witkowski and Parkes (2012b) provide a mechanism that is incentive compatible for $n \geq 2$ agents while eliminating the common prior assumption, again for binary signals. In both papers, Witkowski and Parkes favor ex-post individual rationality over ex-post budget-balance, a reasonable choice since the two properties are incompatible in all non-trivial mechanisms for this setting.

The approaches of Radanovic and Faltings (2013) and Zhang and Chen (2014) are the most similar to this paper. Each introduce mechanisms that are incentive compatible for small groups agents for any finite number of signals, extending the binary truth serum of Witkowski and Parkes (2012a). Furthermore, they consider correlated signals in general, not conditionally independent signals as in Prelec or Witkowski and Parkes.

Radanovic and Faltings prove no mechanism can be incentive compatible for

[1] For instance, before and after receiving an item purchased online.

all belief structures if it uses only information reports or is *decomposable* (i.e. additively separable in the information and prediction reports). They give sufficient conditions for each respective type of mechanism to be incentive compatible, which can be roughly interpreted as each signal being sufficiently positively correlated with itself across agents. Radanovic and Faltings then introduce mechanisms that meet these necessary conditions for $n \geq 2$ agents.

The mechanism of Zhang and Chen on the other hand puts no constraints on the correlation between signals, requiring instead a second-order stochastic relevance condition in addition to common priors for $n \geq 3$ agents. They achieve this by operating the mechanism sequentially, first collecting signals and then collecting predictions after passing one signal report to each agent.

In this paper, I introduce a class of non-decomposable truth serums where assuming *common predictions* is sufficient for weak incentive compatibility. The common predictions assumption is similar in spirit to common priors, but is neither weaker nor stronger. Under additional mild assumptions—stochastic relevance of signals, full support of agent beliefs, and the use of a strictly proper scoring rule—the mechanism is strictly incentive compatible when the number of agents is one more than the number of possible signals. I make no assumptions about the degree or direction of correlation between signals. Furthermore, with minor modifications, the mechanism is still feasible if prediction reports are optional. Agents will have a strict incentive to make a prediction even if others might omit theirs. I also address some potential concerns about the robustness of the mechanism, like how to eliminate Pareto-dominating uninformative equilibria and how to weaken the assumption of risk neutrality.

2.1 Model

The respondent pool contains n rational, risk-neutral, and self-interested agents. Let a typical agent have index i. Each agent receives a *signal* x_i drawn from a shared finite set T. The vector $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_n) \in T^n$ is the *signal profile* of the participants. Signals are private, with each agent knowing their own with certainty and having probabilistic beliefs about the signals of others. Signals can represent an attribute of the agent or observations about an external state of the world, depending on the question of interest to the mechanism operator. For instance, college freshmen could be asked to report the number of drinks of alcohol they've had in the past week from the set $T = \{\textit{0 drinks, 1-2 drinks, 3-6 drinks, 7+ drinks}\}$. Alternatively, crowdsourced reviewers could be asked to evaluate a writing sample according to some task guidlines, reporting a rating from the set $T = \{\textit{1, 2, 3, 4, 5}\}$. In this last example, a signal of $x_i = \textit{4}$ reflects i's judgement that the writing sample fits into the *4* category given their observations, although i could still be uncertain about what rating the writing "truly" deserves.

Signals are correlated across agents. Conditional on their private signal, each agent has a posterior *prediction* $p_i = p(x_i) = \mathrm{E}_i[\bar{x}_{-i} \,|\, x_i] \in \Delta^T$ of the distribution of others' signals, where $\bar{x}^t_{-i} = \sum_{j \neq i} \mathbb{1}(x_j = t)/(n-1)$ is the sample proportion of the agents except for i receiving signal t and $\bar{x}_{-i} = (\ldots, \bar{x}^t_{-i}, \ldots) \in \Delta^T$ is the vector of sample proportions. I will use $p_i = p(x_i)$ to refer to i's actual prediction and $p(\hat{x}_i)$ to refer to i's prediction conditional on a hypothetical signal $\hat{x}_i$. The vector $\boldsymbol{p} = (p_1, \ldots, p_n) \in (\Delta^T)^n$ is the *prediction profile* of agents. Note that I consider any correlated signals similarly to Zhang and Chen (2014), generalizing the conditionally-independent signals of Prelec or Witkowski and Parkes.

To be meaningful, different signals should convey some distinct information, as expressed in the following assumption:

Definition 2.1 (Stochastic relevance)**.** *Signals are* stochastically relevant *if different signals induce different predictions:*

$$\forall i, j: \quad x_i \neq x_j \implies p(x_i) \neq p(x_j)$$

Furthermore, private signals are the only source of differences in beliefs:

Definition 2.2 (Common predictions)**.** *Agents have* common predictions *if agents with the same signal have the same posterior predictions:*

$$\forall i, j: \quad x_i = x_j \implies p(x_i) = p(x_j)$$

Common predictions is the converse of stochastic relevance, and together they imply a bijection between signals and predictions across agents.

For incentives to be strict, agents should think every profile of signals is possible:

Definition 2.3 (Full support)**.** *Agents' beliefs have full support if*

$$\forall i, \forall (\ldots, t_{i-1}, t_{i+1}, \ldots) \in T^{n-1}: \quad \Pr_i[x_{-i} = (\ldots, t_{i-1}, t_{i+1}, \ldots) \mid x_i] > 0$$

The variable $\delta_i = \min_{t \in T} |\{x_j = t \mid j \neq i\}|$ denotes the minimum number of agents except for i in each signal group. The agents will frequently condition on at least one other agent reporting each signal, so let $p_{i|\delta_i \geq 1} = \mathrm{E}[\bar{x}_{-i} \mid x_i \; and \, \delta_i \geq 1]$ and $p_{i|\delta_i = 0} = \mathrm{E}[\bar{x}_{-i} \mid x_i \; and \, \delta_i = 0]$. Then, the agent's prediction p_i satisfies

$$p_i = \Pr_i[\delta_i = 0 \mid x_i] \; p_{i|\delta_i=0} + \Pr_i[\delta_i \geq 1 \mid x_i] \; p_{i|\delta_i \geq 1}$$

The mechanism collects signals and predictions from all agents simultaneously, assigning agents a score $S_i(\boldsymbol{x}, \boldsymbol{p})$ based on the vectors of reports $\boldsymbol{x}$ and $\boldsymbol{p}$. Agents are risk-neutral and maximize their expected score from the mechanism conditional on their private information. The score can be interpreted as a payment from the mechanism to the agent—with negative scores being payments from the agent to the mechanism—or an abstract reputation score.

2.1.1 Common predictions vs common priors

The common predictions assumption is distinct from the common prior assumption used extensively in this literature. The two are still similar in spirit, both saying that differences in beliefs come only from differences in observations. Although I will not assume common priors in this paper, I define it here for comparison:

Definition 2.4 (Common priors)**.** *Agents have* common priors *if all assign the same prior probability to each signal profile:*

$$\forall i, j, \forall (t_1, \ldots, t_n) \in T^n: \quad \Pr_i[\boldsymbol{x} = (t_1, \ldots, t_n)] = \Pr_j[\boldsymbol{x} = (t_1, \ldots, t_n)]$$

Common priors is neither necessary nor sufficient for common predictions. A common prior where agents have different marginal distributions will not satisfy common predictions in general, though it could if some agents have zero probability of receiving some signals due to the ambiguity of conditioning on a zero probability event. Agents can have common predictions and uncommon priors if an agent disagrees with others about the prior marginal distribution of his own signal.

When agents have common predictions and common priors, beliefs will have some degree of symmetry across agents, though an exact characterization is beyond the scope of this paper. An *exchangeable* common prior—with $\Pr_i[\boldsymbol{x}] = \Pr_i[\sigma(\boldsymbol{x})]$ for every signal profile $\boldsymbol{x}$ and every permutation $\sigma(\cdot)$—is clearly sufficient for common predictions. For a simple example of a non-exchangeable common prior with common predictions, consider three agents with $T = \{a, b, c\}$ where a profile with three identical signals has probability $15/102$, three profiles satisfy $\Pr[(a, b, c)] = \Pr[(c, a, b)] = \Pr[(b, c, a)] = 12/102$, and the remaining 21 profiles have probability $1/102$ for each. An agent receiving $x_i = a$ has the prediction $p(a) = (1/2, 1/4, 1/4)$. However, $\Pr[x_2 = b \,|\, x_1 = a] = 42/102 \neq 9/102 = \Pr[x_3 = b \,|\, x_1 = a]$, so agents 2 and 3 are not identical from agent 1's perspective.

2.2 Proper scoring rules

Truth serums have their foundations in scoring rules. Proper scoring rules are incentive schemes for eliciting a rational, risk-neutral payment-maximizer's honest probabilities of some event. For an overview of the theory of scoring rules and their applications, see Gneiting and Raftery (2007). The event in question is usually assumed to be publicly observed or externally verified, such as the weather tomorrow or the winner of an election. In this paper, however, the agents will be scored on their predictions of the reported signals of other agents.

A scoring rule R assigns payments $R(t, \hat{p}_i)$ for a realized outcome t of a random variable x and a reported probability distribution $\hat{p}_i$. If the agent's expected score $E[R(x, \hat{p}_i)]$ is maximized when they report their honest subjective probabilities of each event, the scoring rule is *proper*. Scoring rules are *strictly proper* if the honest prediction is the unique maximizer of the expected score and *weakly proper* if other reports can also be maximizers.

Well-known examples of strictly proper scoring rules for discrete random variables include the logarithmic rule (Good 1952)

$$R_{\log}(t, p_i) = \ln(p_i^t)$$

and the quadratic scoring rule (Brier 1950)

$$R_{\text{quad}}(t, p_i) = 1 - \frac{1}{2} \sum_{s \in T} (\mathbb{1}(t = s) - p_i^s)^2$$

where p_i^t is the probability assigned to realization t.

If R_1 and R_2 are proper scoring rules, then the affine combination

$$R(t, p_i) = a_1 R_1(t, p_i) + a_2 R_2(t, p_i) + b$$

is also a proper scoring rule for all $a_1, a_2 \geq 0$ and $b \in \mathbb{R}$. Using this property, a scoring rule for the probability of a multinomial event can be easily extended to a scoring rule for the expected proportion of outcomes by averaging over scores for each individual event. For instance, the extended logarithmic scoring rule is

$$R_{\log}(\bar{x}, p_i) = \sum_{t \in T} \bar{x}^t \ln(p_i^t)$$

and the extended quadratic scoring rule is

$$R_{\text{quad}}(\bar{x}, p_i) = \frac{1}{2} + \sum_{t \in T} p_i^t \bar{x}^t - \frac{1}{2}(p_i^t)^2$$

where $\bar{x}^t$ is the proportion of t outcomes in the sample. Extended scoring rules are affine in their first argument and are proper when

$$\forall p_i, \hat{p}_i \in \Delta^T, \quad R(p_i, p_i) \geq R(p_i, \hat{p}_i)$$

All extended proper scoring rules for discrete events can be represented using some convex function $F : \Delta^T \to \mathbb{R}$ (Savage 1971) as

$$R(\bar{x}, p_i) = F(p_i) + \langle F'(p_i), \bar{x} - p_i \rangle$$

where $F'(p)$ is a subgradient of F (supporting F from below analogous to the gradient of a differentiable function) and $\langle \cdot, \cdot \rangle$ is the standard inner product. The function $F(p)$ can be interpreted as the score an agent with belief p expects to receive when reporting honestly. For instance, we have

$$R_{\log}(\bar{x}, p) = \sum_{t\in T} p^t \ln(p^t) + \big\langle (\ldots, \ln(p^t)+1, \ldots), \bar{x} - p \big\rangle, \text{ for } F(p) = \sum_{t\in T} p^t \ln(p^t)$$

and

$$R_{\text{quad}}(\bar{x}, p) = \frac{1}{2} + \frac{1}{2}\sum_{t\in T}(p^t)^2 + \langle p, \bar{x} - p\rangle, \text{ for } F(p) = \frac{1}{2} + \frac{1}{2}\sum_{t\in T}(p^t)^2.$$

2.3 Minimum truth serums

A truth serum is a detail-free mechanism for collecting private signals from a group of agents without external verification. The mechanism operates by soliciting from each agent i their signal x_i and prediction p_i of the distribution of signals of other agents. Given some proper scoring rule R, the agents' predictions will be scored as $R(\bar{x}_{-i}, p_i)$ based on the actual distribution of other signals $\bar{x}_{-i}$. Since R is proper, this score will be maximized in expectation when the agent reports p_i honestly. For the agent to honestly reveal their signal x_i as well, the final score will be bounded above by the score for the "average" prediction of others reporting the same signal. The predictions of others with the same signal will be aggregated so that if all the inputs are identical, then that same value is the output, as expressed in the following definition:

Definition 2.5 (Unanimous aggregator). *A function $g : \cup_{k\in\mathbb{N}}(\Delta^T)^k \to \Delta^T$ that maps profiles of predictions back into predictions is a unanimous or idempotent aggregator if $g(\{p_j, \ldots, p_j\}) = p_j$ for all $p_j \in \Delta^T$.*

Common averaging functions such as the arithmetic mean or the normalized geometric mean are unanimous. Alternatively, the aggregator could select one input according to some criterion, such as an element in the set with minimum Euclidean norm breaking ties lexicographically.

The precise definition of the minimum truth serum class is as follows:

Definition 2.6 (Class of minimum truth serums). *Given a proper scoring rule R and a unanimous aggregator g, a minimum truth serum is defined by the following procedure:*

1. *Collect the vectors of signals $\boldsymbol{x}$ and predictions $\mathbf{p}$ from agents simultaneously, with elements x_i and p_i being the reports of agent i.*

2. *Define the variables*

$$\delta_i = \min_{t \in T} |\{x_j = t \,|\, j \neq i\}|$$

for each agent denoting the minimum number of other agents that can be found in each signal group.

3. *If* $\delta_i \geq 1$*, compute the proxy prediction*

$$q_i(x_i) = g(\{p_j \in p_{-i} \,|\, x_j = x_i\})$$

4. *Assign scores to each agent as*

$$S_i(\boldsymbol{x}, \boldsymbol{p}) = \begin{cases} R(\bar{x}_{-i}, p_i) & \text{if } \delta_i = 0 \\ \min\left\{R(\bar{x}_{-i}, p_i), R(\bar{x}_{-i}, q_i(x_i))\right\} & \text{if } \delta_i \geq 1 \end{cases}$$

When there are no other agents besides i reporting x_i, the proxy prediction clearly isn't well defined. However, due to the definition of δ_i, the mechanism will sometimes avoid comparing i to the others with the same signal even when proxy prediction can be computed. This is so that i cannot purposefully report a rare signal to avoid the proxy upper bound. Since δ_i depends only on the reports of others, i cannot directly manipulate it.

Definition 2.7 (Bayesian incentive compatibility)**.** *A truth serum is Bayesian incentive compatible if, for all* i*,*

$$\mathrm{E}[S_i((x_i, x_{-i}), (p_i, p_{-i})) \,|\, x_i, p_i] \geq \mathrm{E}[S_i((\hat{x}_i, x_{-i}), (\hat{p}_i, p_{-i})) \,|\, x_i, p_i]$$

for all $x_i, \hat{x}_i, p_i, \hat{p}_i$ *and is strictly incentive compatible if the inequality is always strict when* $x_i \neq \hat{x}_i$ *or* $p_i \neq \hat{p}_i$*.*

Proposition 2.1. *All minimum truth serums are Bayesian incentive compatible if agents have common predictions. If* $n \geq |T| + 1$*,* R *is strictly proper, beliefs have full support, and signals are stochastically relevant, then the truth serum is strictly incentive compatible.*

Proof. Suppose all agents except i report their signals and predictions honestly. Then, the expected score of agent i when giving report $(\hat{x}_i, \hat{p}_i)$ (with all probabilities conditional on x_i) satisfies

$$\begin{aligned}
&\mathrm{E}[S_i((\hat{x}_i, x_{-i}), (\hat{p}_i, p_{-i}))] = \\
&\qquad \Pr_i[\delta_i = 0] \sum_{x_{-i}:\delta_i=0} R(\bar{x}_{-i}, \hat{p}_i) \Pr_i[x_{-i} \,|\, \delta_i = 0] \\
&\qquad\quad + \Pr_i[\delta_i \geq 1] \sum_{x_{-i}:\delta_i \geq 1} \min\left\{R(\bar{x}_{-i}, p_i), R(\bar{x}_{-i}, q_i(\hat{x}_i)\right\} \Pr_i[x_{-i} \,|\, \delta_i \geq 1] \\
&\qquad \leq \Pr_i[\delta_i = 0] \sum_{x_{-i}:\delta_i=0} R(\bar{x}_{-i}, \hat{p}_i) \Pr_i[x_{-i} \,|\, \delta_i = 0] \\
&\qquad\quad + \Pr_i[\delta_i \geq 1] \sum_{x_{-i}:\delta_i \geq 1} R(\bar{x}_{-i}, \hat{p}_i) \Pr_i[x_{-i} \,|\, \delta_i \geq 1] \\
&\qquad = \sum_{x_{-i}} R(\bar{x}_{-i}, \hat{p}_i) \Pr_i[x_{-i}] \\
&\qquad = R(p_i, \hat{p}_i) \leq R(p_i, p_i)
\end{aligned}$$

where the last line follows from p_i being i's expectation of $\bar{x}_{-i}$ and R being affine and maximized at $\hat{p}_i = p_i$ as a proper scoring rule. Since $p_j = p_i$ when $x_j = x_i$ by common predictions, we have $q_i(x_i) = g(\{p_j \in p_{-i} \,|\, x_j = x_i\}) = p_i$ because g is unanimous. Hence, the expected score when reporting truthfully achieves the upper bound above, and the honest report is a best response for i. Therefore, the minimum truth serum is (weakly) Bayesian incentive compatible. Notice that if $n < |T| + 1$, then δ_i is always zero and the agent will be indifferent between all information reports since there aren't enough other agents to fill each category.

Now assume $n \geq |T| + 1$, R is strictly proper, beliefs have full support, and signals are stochastically relevant. Beliefs having full support and $n \geq |T| + 1$ imply $\Pr_i[\delta_i \geq 1] > 0$, meaning there is some chance agent i will face the proxy upper bound. There are two cases to consider besides the honest report. First, any report with $\hat{p}_i \neq p_i$ will result in a strictly lower score since R is strictly proper. Second, reporting a dishonest signal $\hat{x}_i \neq x_i$ and an honest prediction p_i will lead to a strictly lower score since i will occasionally be matched with $q_i(\hat{x}_i) = p_j \neq p_i$ (by stochastic relevance), and we must occasionally have $R(\bar{x}_{-i}, q_i(\hat{x}_i)) < R(\bar{x}_{-i}, p_i)$ since R is strictly proper. Therefore, honesty is the unique best response under these assumptions. □

Although I've considered the number of agents n to be fixed, this could be a random variable from i's perspective. In this case, the assumption that $n \geq |T| + 1$ can be replaced with $P_i[n \geq |T| + 1] > 0$ for strict incentive compatiblity.

This mechanism is one of two known truth serums that can be used when a signal is negatively correlated with itself across agents[2]. Negative correlation can occur if there are a limited number of certain signals, where an observation by one agent "blocks" another agent's observation.

Consider the following example: A group of eleven birdwatching enthusiasts will attempt to sight the Lesser Jubjub, with each stationing themselves in a different area of a valley. Since the enthusiasts are prone to boasting in the absence of incentives for honesty, the president of their society will ask each whether they saw the bird (corresponding to a set of answers $T = \{Yes, No\}$) and give payments according to the minimum truth serum using the log scoring rule and some aggregator. Since the Lesser Jubjub is known to maintain a very small territory, if one watcher catches sight of it, she'll conclude it's less likely that others have seen it. In particular, assume beliefs are:

	p_i^{Yes}	p_i^{No}
$x_i = Yes$	.05	.95
$x_i = No$	.10	.90

Suppose the honestly reported signals are $x_1 = Yes$ and $x_i = No$ for all other i. Then, $\delta_1 = 0$ since all other ten reported *No*, and $\delta_i = 1$ for $i > 1$ since each answer was given by someone besides i. Finally, the score of agent 1 is $S_1(x, p) = R(\bar{x}_{-1}, p_1) = R((0.0, 1.0), (0.05, 0.95)) = \ln(0.95) \simeq -0.051$, and the scores of all other agents are

$$\begin{aligned} S_i(x,p) &= \min\{R(\bar{x}_{-i}, p_i), R(\bar{x}_{-i}, g(x_i))\} \\ &= \min\{R((0.1, 0.9), (0.1, 0.9)), R((0.1, 0.9), (0.1, 0.9))\} \\ &= 0.1 \ln(0.1) + 0.9 \ln(0.9) \simeq -0.325. \end{aligned}$$

2.4 Optional prediction reports

Asking agents to provide an information report is straightforward since everyone has experience with surveys and ratings. However, requiring all agents to submit a prediction report could be a practical barrier to deploying a truth serum, especially for a moderately large set of answers. Impossibility results

[2] The other being the knowledge-free peer prediction mechanism of Zhang and Chen (2014).

have been given (Radanovic and Faltings 2013) about mechanisms that depend only on information reports, which I skirt by including predictions, but making them optional. Because the minimum truth serum allows for the possibility of an agent being the only one in their signal group, the mechanism can be easily modified to operate when some predictions are missing. Rather than defining scores conditional on all signals being reported by some other agent, scores will be conditional on all signals being given by an agent who also made a prediction.

Definition 2.8 (Minimum truth serum with optional predictions)**.** *Given a bounded proper scoring rule R and a unanimous aggregator g, a minimum truth serum with optional predictions is defined by the following procedure:*

1. *Collect the vectors of signals $\boldsymbol{x}$ and predictions $\boldsymbol{p}$ from agents simultaneously, where each agent has the option of selecting $p_i = \varnothing$.*
2. *Define the variables*

$$\delta_i = \min_{t \in T} |\{x_j = t \mid j \neq i \text{ and } p_j \neq \varnothing\}|$$

for each agent denoting the minimum number of other agents with predictions that can be found in each signal group.

3. *If $\delta_i \geq 1$, compute the proxy prediction*

$$q_i(x_i) = g(\{p_j \in p_{-i} \mid x_j = x_i \text{ and } p_j \neq \varnothing\})$$

4. *Assign scores to each agent with $p_i \neq \varnothing$ as*

$$S_i(\boldsymbol{x},\boldsymbol{p}) = \begin{cases} R(\bar{x}_{-i}, p_i) & \text{if } \delta_i = 0 \\ \min\left\{R(\bar{x}_{-i}, p_i), R(\bar{x}_{-i}, q_i(x_i))\right\} & \text{if } \delta_i \geq 1 \end{cases}$$

5. *Assign scores to each agent with $p_i = \varnothing$ as*

$$S_i(\boldsymbol{x},\boldsymbol{p}) = \begin{cases} \min_{q \in \Delta^T}\{R(\bar{x}_{-i}, q)\} & \text{if } \delta_i = 0 \\ R(\bar{x}_{-i}, q_i(x_i)) & \text{if } \delta_i \geq 1 \end{cases}$$

giving the agent the minimum possible score according to R when $\delta_i = 0$.

Proposition 2.2. *All minimum truth serums with optional predictions are Bayesian incentive compatible if agents have common predictions. If $n \geq |T|+1$, R is strictly proper, beliefs have full support, and signals are stochastically relevant, then the truth serum is strictly incentive compatible.*

Proof. In addition to the argument of the previous proof, we need to establish that each agent prefers giving a prediction to reporting $\hat{p}_i = \varnothing$. If all other agents report truthfully—but possibly with some giving the null prediction—the expected score for reporting $(\hat{x}_i, \varnothing)$ conditional on x_i is

$$\begin{aligned}
\mathrm{E}[S_i((\hat{x}_i, x_{-i}), (\varnothing, p_{-i}))] = \Pr_i[\delta_i = 0] &\sum_{x_{-i}:\delta_i=0} \min_{q\in\Delta^T}\{R(\bar{x}_{-i}, q)\} \Pr_i[x_{-i} \,|\, \delta_i = 0] \\
&+ \Pr_i[\delta_i \geq 1] \sum_{x_{-i}:\delta_i\geq 1} R(\bar{x}_{-i}, p(\hat{x}_i)) \Pr_i[x_{-i} \,|\, \delta_i \geq 1] \\
&\leq \Pr_i[\delta_i = 0] \sum_{x_{-i}:\delta_i=0} R(\bar{x}_{-i}, p(\hat{x}_i)) \Pr_i[x_{-i} \,|\, \delta_i = 0] \\
&+ \Pr_i[\delta_i \geq 1] \sum_{x_{-i}:\delta_i\geq 1} R(\bar{x}_{-i}, p(\hat{x}_i)) \Pr_i[x_{-i} \,|\, \delta_i \geq 1] \\
&= \sum_{x_{-i}} R(\bar{x}_{-i}, p(\hat{x}_i)) \Pr_i[x_{-i}] = R(p_i, p(\hat{x}_i)) \leq R(p_i, p_i) \\
&= \mathrm{E}\, S_i(x_i, p_i)
\end{aligned}$$

since $q_i(\hat{x}_i) = p(\hat{x}_i)$ by common predictions, so the agent is never better off when omitting the prediction. When beliefs have full support, the first inequality will be strict and providing a full, honest report is a strict best response. □

Notice that the proof concludes something slightly stronger than Bayesian incentive compatibility, showing a full and honest report is an interim best response even if others fail to best respond and omit their prediction.

While making a full report is preferable, the point of optional predictions is that we expect agents to occasionally omit them in practice. Let's consider the incentives of an agent conditional on reporting $\hat{p}_i = \varnothing$. Ideally, such an agent would still prefer reporting their true signal, but this won't always be the case. The reported signal only affects payoffs when $\delta_i \geq 1$, so the agent should choose their report $\hat{x}_i$ conditional on this event to maximize the expected score $R(p_{i|\delta_i\geq 1}, q_i(\hat{x}_i))$. Since $p_i \neq p_{i|\delta_i\geq 1}$ in general, this opens up the possibility that the agent would want to misreport their signal if another prediction $p(\hat{x}_i)$ approximates $p_{i|\delta_i\geq 1}$ better than $p_i = p(x_i)$ does under R[3].

[3] The notion of how close one probability vector is to another under a proper scoring rule

For instance, suppose the posterior predictions are $p^a(a) = \mathrm{E}[\bar{x}^a_{-i} \,|\, a] = 0.99$ and $p^a(b) = 0.45$ given two possible signals a and b, with $p^a_{i|\delta_i \geq 1}(a) = \mathrm{E}[\bar{x}^a_{-i} \,|\, a \; and \; \delta_i \geq 1] = 0.5$. While an agent with $x_i = a$ thinks B signals are rare, agents with the b signal are relatively better predictors when they are present. Since $R(p_{i|\delta_i \geq 1}(a), p(b)) = R((.5,.5),(.45,.55)) > R((.5,.5),(.99,.01)) = R(p_{i|\delta_i \geq 1}(a), p(a))$ for typical scoring rules, an agent with $x_i = A$ would prefer report $(b, \varnothing)$ over $(a, \varnothing)$.

Despite this possibility, these “failures” of incentive compatibility out of equilibrium are unconcerning. If an agent is sophisticated enough to notice an improvement from misreporting their signal conditional on omitting their prediction, they would be sophisticated enough to best respond by giving a full report. Furthermore, these gains can exist only when $p_{i|\delta_i \geq 1}$ is sufficiently different from p_i for a fixed set of posterior predictions. Since $p_{i|\delta_i \geq 1} \to p_i$ as $\Pr_i[\delta_i \geq 1 \,|\, x_i] \to 1$, this possibility goes away completely for large enough n if agents become increasingly certain that every signal will be given by at least one person along with a prediction.

2.5 Further considerations for robust mechanisms

2.5.1 Individual rationality and budget balance

In the honest Bayes-Nash equilibrium, each agent receives a score of $R(\bar{x}_{-i}, p_i)$. Depending on the scoring rule R used, these scores could be negative, positive, or sum to any amount. If agents have the option to sit out from the mechanism, then participation shouldn’t leave them worse off. Whether agents can’t be worse off in every case or on average leads to the following two notions of individual rationality:

Definition 2.9. *A truth serum is* ex-post individually rational (EPIR) *if the realized score satisfies* $S_i(\boldsymbol{x}, \boldsymbol{p}) \geq 0$ *for all* i *and for all profiles of reports* $\boldsymbol{x}$ *and* $\boldsymbol{p}$.

Definition 2.10. *A truth serum is* interim individually rational (IIR) *if the expected score over the types of others conditional on* x_i *satisfies* $\mathrm{E}[S_i((x_i, x_{-i}), (p_i, p_{-i})) \,|\, x_i] \geq 0$ *for all* i.

R can be formalized as its corresponding *Bregman divergence* $D_R(p,q) = R(p,p) - R(p,q)$. For example, the Bregman divergence of the quadratic scoring rule is squared Euclidean distance, and the Bregman divergence of the log scoring rule is the Kullback-Leibler divergence.

Similarly, we can consider two forms of budget-balance depending on whether the total scores for all agents sum to zero for all realizations or on average:

Definition 2.11. *A truth serum is* ex-post budget balanced *(*EPBB*) if the total scores satisfy* $\sum_i S_i(\boldsymbol{x}, \boldsymbol{p}) = 0$ *for all profiles of reports* $\boldsymbol{x}$ *and* $\boldsymbol{p}$.

Definition 2.12. *A truth serum is* ex-ante budget balanced (EABB) *if the total scores satisfy* $\mathrm{E}[\sum_i S_i(\boldsymbol{x}, \boldsymbol{p})] = 0$ *in expectation over all profiles* $(\boldsymbol{x}, \boldsymbol{p})$

Since a primary advantage of the truth serum is that it is detail-free, the notion of ex-ante budget balance isn't applicable. Unfortunately, ex-post budget balance can be a strong condition. EPBB and EPIR are incompatible for any non-trivial mechanism in this setting. IIR will also tend to be violated under EPBB mechanisms. When agents have common priors, EPBB and IIR will be mutually be satisfied only if the iterim expected scores are exactly zero for each type since there are no gains from interaction. Individual rationality also becomes a more complex goal if agents have some unknown costs of participation, reflecting either effort to acquire a signal (Dasgupta and Ghosh 2013; Witkowski et al. 2013) or concern over privacy (Ghosh and Roth 2015).

2.5.2 Collusion resistance and balanced truth serums

As the minimum truth serum is defined, honesty is a Bayes-Nash equilibrium, but it is not the only one. In fact, untruthful equilibria might Pareto-dominate honest revelation. Notice that any strategy profile where the agents draw $\hat{x}_i$ according to a common fixed distribution $\hat{p}$ and report $(\hat{x}_i, \hat{p})$ is a Bayes-Nash equilibrium. Since the expected score is $\mathrm{E}\, S_i(\hat{x}_i, \hat{p}) = R(\hat{p}, \hat{p}) = F(\hat{p})$ for some convex F by the Savage representation, the total scores will be maximized when the agents can coordinate on a degenerate distribution and report some signal t^* with probability one. One benefit of ex-post budget balance is that it guarantees agents cannot coordinate to increase their total score, and hence provides a means of reducing the tempation of any untruthful equilibria.

Define the balanced minimum truth serum as follows:

Definition 2.13 (Balanced minimum truth serum)**.** *Given a proper scoring rule* R *and a unanimous aggregator* g*, the balanced minimum truth serum scores* *(with or without optional predictions) are*

$$S_i^b(\boldsymbol{x},\boldsymbol{p}) = \frac{1}{n-1}\sum_{j\neq i}\big(S_i(x_{-j},p_{-j}) - S_j(x_{-i},p_{-i})\big)$$

Since each term $S_i(x_{-j},p_{-j})$ is individually incentive compatible and the terms $S_j(x_{-i},p_{-i})$ don't depend on i's report, these scores are still Bayesian incentive compatible with the caveat that we now require $\Pr_i[n \geq |T|+2] > 0$ for strict incentive compatibility. Ex-post budget balance follows from each term $S_i(x_{-j},p_{-j})$ appearing in the total scores exactly once with a positive sign—in the score of agent i—and exactly once with a negative sign—in the score of agent j. The uninformative coordination equilibria still exist, but no longer Pareto-dominate the honest equilibrium since the total scores are fixed at zero.

As noted earlier, these mechanisms will not be interim individually rational in general. For example, suppose $T = \{a,b\}$ and $n = 4$. Agents believe signals are conditionally independent based on two states, A and B, with $\Pr[A] = 0.1$, $\Pr[B] = 0.9$, $\Pr[x_j = a|A] = 0.9$, and $\Pr[x_j = b|B] = 0.1$. Under the minimum truth serum with the log scoring rule, an agent with $x_i = a$ expects $S_i(x_{-j},p_{-j})$ to be -0.693 and $S_j(x_{-i},p_{-i})$ to be -0.593, leading to the expected balanced score to be approximately -0.1. Thus, agents with $x_i = a$ would prefer a score of zero from not participating if possible.

If the scoring rule R is bounded by the interval $[M_1, M_2]$, then the balanced truth serum scores are always greater than $M_1 - M_2$. Subtracting this negative constant from each agent's balanced score guarantees ex-post individual rationality while making total scores always sum to $n(M_2 - M_1)$.

2.5.3 Risk aversion and probabilistic rewards

Another point of concern for robustness is the assumption of risk-neutral agents. If agents are risk-averse as we might typically expect, Bayesian incentive compatiblity or interim individual rationality could be violated. When the Bernoulli utility function $u_i(\cdot)$ of an agent is known, assigned scores can be transformed to $u_i^{-1}(S_i(\boldsymbol{x},\boldsymbol{p}))$ to counteract the agent's risk preference. When the agent's risk preference is unknown, another trick is available: payment in lottery shares. An expected utility maximizer can be risk-averse across varying rewards, but will always have risk-neutral preferences over a varying probability of a fixed reward. Reinterpreting scores as the probability of winning a prize means risk neutrality holds without loss of generality. Now, ex-post individual rationality has a dual role of ensuring scores are proper probabilities. Hossain and Okui (2013) and Schlag and van der Weele (2009) explore this trick theoretically and experimentally in the case of eliciting judgements from a single agent.

There are two ways of incorporating probablistic rewards into a truth serum. First, the mechanism could give each agent the opportunity to win their own prize. Using a scoring rule R bounded in the interval $[0, M]$, the mechanism assigns scores $S_i(\boldsymbol{x}, \boldsymbol{p})$ and draws thresholds $K_i \sim \text{Unif}[0, M]$. Agent i wins their prize if and only if $K_i < S_i(\boldsymbol{x}, \boldsymbol{p})$. Alternatively, the mechanism could award a single prize, splitting shares in the prize lottery according to scores. This entails using a balanced truth serum adjusted so that total scores sum to one and scores are always non-negative:

Definition 2.14 (Lottery minimum truth serum)**.** *Given a proper scoring rule R bounded between $[0,1]$ and a unanimous aggregator g, the lottery minimum truth serum scores (with or without optional predictions) are*

$$S_i^{\ell}(\boldsymbol{x}, \boldsymbol{p}) = \frac{1}{n} + \frac{1}{n(n-1)} \sum_{j \neq i} \left(S_i(x_{-j}, p_{-j}) - S_j(x_{-i}, p_{-i}) \right),$$

denoting the probability that agent i wins the prize.

2.6 Conclusion

This paper gives a new class of detail-free mechanisms for eliciting correlated signals. The only restrictive assumption for incentive compatibility is that all agents with the same signal have the same posterior expectations. The obvious next question is whether the common predictions assumption can be weakened, possibly at the cost of different conditions on belief structures. The idea underlying this paper is that under common predictions, another agent with the same signal can act as a perfect proxy for an agent's belief. This suggests incentive compatibility should be feasible when agents think others with the same signal are better proxies on average, if not exactly.

In this paper, I've assumed agents have no direct preferences over how their reports are used, which is plausible for many surveys and ratings. In more general settings, the incentives for honesty provided by a truth serum could be used to counteract other incentives for dishonesty. One direction to explore is when it's possible to layer truth serum transfers on top of another mechanism to provide incentive compatibility. This broad idea has an established history in mechanism design. For instance, Crémer and McLean (1988) employ a similar transfer scheme to prove their full surplus extraction result for correlated types. Though presented more as a paradox than a practical result, their mechanism assumes precise knowledge of the agents' common prior. The question of when detail-free transfers like the minimum truth serum could fill an analogous role in general implementation problems remains open.

CHAPTER 3

UNCOORDINATED TWO-SIDED MATCHING MARKETS

(Joint with Juan Fung)

3.1 Introduction

Suppose you are tasked with pairing a group of men and a group of women together when each person has preferences over potential partners. Can you accomplish this task such that, once your matching is in place, no man or woman can obtain a better partner on their own? In a seminal paper, Gale and Shapley (1962) model this situation as a marriage market and present an elegant solution: the deferred acceptance algorithm. Gale and Shapley (1962) use their algorithm to prove that such *stable* matchings exist in two-sided markets. While Gale and Shapley (1962) did not aim to provide guidance on applied market design, their algorithm has come to play a key role in the design of centralized markets.[1] The question we address in this paper is: how well can agents do without centralization?

Gale and Shapley (1962)'s algorithm has been independently discovered by practitioners many times. As Roth (1984) showed, the National Resident Matching Program (NRMP) for new doctors had been using a version of deferred acceptance since 1951. Several other instances of the algorithm in the field were later documented, including the job market for clinical psychologists and dorm room assignment at MIT (Roth 2008). Such observations renewed interest in Gale and Shapley (1962)'s stylized model of matching, with stability as the principal objective and deferred acceptance as the foundation for practical market design. Eventually, deferred acceptance became fundamental to the intentional reorganization of existing markets as centralized clearinghouses,

[1] Gale and Shapley (1962) do state their hope "that some of the ideas introduced here might usefully be applied to certain phases of the admissions problem."

most notably in the redesign of the assignment mechanisms for Boston Public Schools (Abdulkadiroğlu et al. 2005b) and New York City High Schools (Abdulkadiroğlu et al. 2005a).

The deferred acceptance algorithm is practical in applications since it can find a stable matching in polynomial time. With n men and m women in a marriage market, deferred acceptance computes a stable matching in at most $n \cdot m$ steps. Its speed and simplicity make it a natural candidate for centralized design, but there is nothing inherently centralized about deferred acceptance. The standard interpretation of the algorithm as one side of the market making successive proposals has a decentralized flavor to it since proposing agents don't need information beyond their own preferences. However, agents have to be *coordinated* in two ways to properly execute deferred acceptance. First, only one side of the market can make proposals. Second, proposals must be grouped into rounds where agents propose at most once (though the order of proposals within rounds can be arbitrary). We argue that coordination is the distinguishing feature between centralized and decentralized markets.

In particular, we consider a class of decentralized matching markets introduced by Roth and Vande Vate (1990). Their idea, roughly, is to let agents take turns making proposals to a more preferred partner. Two agents that prefer each other to their current respective partners are said to form a *blocking pair*. Roth and Vande Vate (1990) show that, starting from an arbitrary matching, such random proposal processes eventually converge to a stable matching if each blocking pair has positive probability of being selected.

While these results suggest agents can attain stability without a centralized authority, Ackermann et al. (2011) show some markets almost certainly require exponentially many proposals to reach stability when blocking pairs are rematched uniformly at random. One is tempted to conclude that, in practice, decentralized markets cannot be ensured of reaching a stable matching. However, the process described above portrays market participants as naive. In particular, agents will accept any proposal when they are single. We introduce slightly more sophisticated behavior into the process and show that the process converges to stability in polynomial time. Moreover, since centralization by way of deferred acceptance requires at most a quadratic number rounds, we compare welfare under our process after a comparable number of rounds to the outcome of deferred acceptance.

In section 3.2, we introduce the matching environment and the model for proposal processes. Computational results are presented in section 3.3. In section 3.3.1, we explore convergence to stable matching in terms of number of proposals, while in section 3.3.2, we compare welfare of our proposal processes *before* convergence to welfare under stable matching.

3.2 Model

3.2.1 Environment

We consider a standard one-to-one matching environment with strict preferences. In particular, a *matching market* of size n is a triple $\theta = (M, W, \succ)$ consisting of a set of n men M, a set of $n + m$ women W, and preferences $\succ = \{\succ_i\}_{i \in M \cup W}$ for each agent i over potential partners on the other side of the market.

Preferences are strict linear orders over the set of potential partners for that agent and the option to remain unmatched, denoted by $\emptyset$. Agent i finds agent j *acceptable* if $j \succ_i \emptyset$. We will consider environments θ where every man finds every woman acceptable and vice-versa. Let Θ_n denote the set of all such markets of size n. An environment θ is *balanced* for some n if $m = 0$, and otherwise it is *unbalanced.*

A *matching* is a function $\mu : M \cup W \to M \cup W \cup \{\emptyset\}$ describing how agents are paired. Let $\mathscr{X}(\theta)$ denote the set of all matchings for θ. The *partner* of agent i under matching μ is $\mu(i)$. The agent is *single* if $\mu(i) = \emptyset$. Valid matchings satisfy the following properties:

1. If $\mu(i) \neq \emptyset$, then $\mu(\mu(i)) = i$, i.e. if an agent i has a partner, $\mu(i)$, then $\mu(i)$'s partner is i.

2. $\mu(M) \subseteq W \cup \{\emptyset\}$ and $\mu(W) \subseteq M \cup \{\emptyset\}$, i.e. all agents are either matched to an agent from the other side or single.

A matching μ is *stable* in θ if there is no pair of agents m, w such that $m \succ_w \mu(w)$ and $w \succ_m \mu(w)$ and no agent i such that $\emptyset \succ_i \mu(i)$. Since we consider only environments where all agents are mutually acceptable, the second requirement that all agents prefer their partners to being single is satisfied in every matching. Let $\mathscr{S}(\theta)$ denote the set of stable matchings for market θ.

3.2.2 Proposal processes as models of matching

Given a matching market θ, agents begin at some initial matching μ_0 with distribution $f_0(\mu) = Pr(\mu_0 = \mu)$. We will primarily consider μ_0 as the empty matching where $\mu_0(i) = \emptyset$ for all $i \in M \cup W$. Matchings evolve according to a sequence of proposals between agents, with one proposal being made at discrete time steps, $t = 1, 2, \ldots$.

A *proposal process*, P, describes who makes a proposal at each step, who the proposer proposes to, and the conditions for whether a proposal is accepted. Given matching μ_t at step t, if agent i proposes to agent j and the proposal is accepted, then μ_t is updated as follows:

1. i and j are paired together: $\mu_{t+1}(i) = j$ and $\mu_{t+1}(j) = i$,
2. i and j's previous partners (if any) are now single: $\mu_{t+1}(\mu_t(i)) = \mu_{t+1}(\mu_t(j)) = \emptyset$,
3. and $\mu_{t+1}(k) = \mu_t(k)$ for all other agents k.

If the proposal is not accepted, then no change occurs, and $\mu_{t+1} = \mu_t$.

3.2.3 Deferred acceptance as a proposal process

In general, a proposal process may be random, inducing a random walk over the set of matchings, although deterministic proposal processes can also be considered. For instance, two versions of the deferred acceptance algorithm—with either men proposing or women proposing—are important examples of deterministic proposal processes.

Deferred acceptance can also be implemented semi-randomly by picking a fixed permutation of the proposing side and having the agents cycle through proposals according to this schedule. Supposing men are proposing, each man knows every other man has acted exactly once since his last proposal. This fixed schedule provides a monotonicity guarantee that underlie deferred acceptance. If an agent is unsure whether or not someone else has acted since their last proposal, the same monotonicity guarantee is no longer present and proposers are unable to work down their preference lists in the same way. Someone who once rejected me might now accept me, so the set of possible partners to consider making proposals doesn't shrink. Of course, it wouldn't be worthwhile to only always propose to my first choice, so a lack of knowledge about when others have acted necessitates some randomness in how proposals are made. The continued need to explore all possible partners particularly holds if both men and women make proposals.

3.2.4 Uncoordinated proposal processes

To represent agents being unable to fully coordinate or observe the actions of others, we consider proposal processes where at each point in time, one agent is selected at random to act. The probability that a man makes a proposal might differ from the probability a woman makes a proposal, but within each side, all agents have an equal chance of acting.

With no knowledge about the actions of others, how should an agent choose who to propose to? A simple answer is to randomly propose to someone better than that agent's current match. A woman who has repeatedly rejected a man might still potentially accept him now, so the man might plausibly think it's worth another shot. If the man can keep track of all the partners of the women he is proposing to, the best he can do is never propose to a women who is with the same partner as a time when she rejected him.

Definition 3.1 (Random better (best) reply)**.** *Begin at random matching μ_0 with some probability $p(\mu_0)$. Given μ_{t-1} at time t:*

1. *Pick a proposing agent $i_t \in M \cup W$ at random*
2. ***Better (best) reply**: Agent i_t proposes to an agent (the best agent) j_t such that*

   $$j_t \succ_{i_t} \mu_{t-1}(i_t)$$

   choosing uniformly at random if multiple such j_t exist.
3. *Agent j_t accepts if $i_t \succ_{j_t} \mu_{t-1}(j_t)$*
   - (a) *If j_t accepts, update μ_t*
   - (b) *Else, set $\mu_t = \mu_{t-1}$*

   Set $t = t + 1$, and return to step 1.

Ackermann et al. (2011) show that these processes may take exponentially many rounds to reach stability. On some level, naive behavior by both proposers and responders causes unnecessary delay. We now introduce two proposal processes that model plausible behavior for both proposers and responders without requiring a high level of sophistication.

To address naive behavior by responders, we introduce aspirations as a way for responders to keep from settling for inferior matches early on in the process. An agent's aspiration level is simply the minimum partner rank the agent is willing to accept. A responder will accept a proposal subject to their current aspiration level, which evolves over time.

Let $\rho(i, j)$ denote agent i's *ranking* of agent j wrt $\succ_i$:

$$\rho(i, j) \equiv |\{k : k \succeq_i j\}|$$

We say that agent i has *aspiration level* $\alpha_t(i)$ at μ_t if i accepts a proposal from any agent $j \in \{j : \rho(i, j) \leq \alpha_t(i)\}$ and rejects a proposal otherwise. Aspiration levels are set and updated as follows:

Definition 3.2 (Updating aspiration levels)**.** *Given an environment θ and a proposal process P with aspiration adjustment $a_n \in \mathbb{R}_+$:*

- *At $t = 0$, initial aspiration is $\alpha_0(j) = 1, \forall j$*
- *At $t > 0$, if j receives a proposal from i, then*

$$\alpha_t(j) = \begin{cases} \rho(j, i), & \mu_t(j) = i \\ \alpha_t(j) + a_n, & \mu_t(j) = \emptyset \end{cases}$$

otherwise, $\alpha_t(j) = \alpha_{t-1}(j)$

In other words, a responder j who accepts a proposal from a proposer i sets their aspiration to their current partner's rank. If instead j rejects i's proposal, then j must adjust $\alpha_t(j)$ to become less picky if j remains single, or else j's aspiration remains set to their current partner, $\mu_t(j)$.

To address naive behavior by proposers, we allow agents to learn who is better than them. In particular, if a man i is rejected by a woman j when she is partnered with some other man k, then man i should not waste another proposal on woman j if she remains partnered with k. That is, man i learns by revealed preference that $k \succ_j i$. Let $\lambda(i)$ denote i's *unattainable set.* This is the set of all pairs of agents in which the woman's partner is preferred to man i,

$$\lambda(i) = \{(j,k) \in W \times M : k \succ_j i\},$$

with a symmetric definition for a woman i. Since agent preferences are private information, this set is empty at $t = 0$ and updated by agent i whenever he is rejected in favor of another man. We thus define learning as follows:

Definition 3.3 (Learning)**.** *Given an environment θ, a proposal process P admits learning if each agent i's unattainable set $\lambda_t(i)$ is updated at $t > 0$ as*

$$\lambda_t(i) = \begin{cases} \lambda_{t-1}(i) \cup (j, \mu_t(j)), & j \textit{ rejects } i \textit{ at } t \\ \lambda_{t-1}(i), & \textit{else} \end{cases}$$

with $\lambda_0(i) = \emptyset, \forall i \in M \cup W$.

An agent j is *attainable at μ_t* for agent i if

$$(j, \mu_t(j)) \not\in \lambda_t(i),$$

that is, if j is not partnered with an agent $\mu_t(j)$ for whom i has been rejected before. It is natural, for example, for a man making a proposal to propose to an attainable woman, rather than proposing to a woman who will surely reject him.

We are now ready to define our first proposal process.

Definition 3.4 (Random best attainable)**.** *Begin at random matching μ_0 with some probability $p(\mu_0)$. Given μ_{t-1} at time t:*

1. *Pick a proposing agent $i_t \in M \cup W$ at random*
2. ***Best attainable proposal**: Agent i_t proposes to the best agent j_t such that*

$$j_t \succ_{i_t} \mu_{t-1}(i_t) \textit{ and } (j_t, \mu_{t-1}(j_t)) \not\in \lambda_t(i_t)$$

   (a) *If no such j_t exists, set $\mu_t = \mu_{t-1}$, set $t = t+1$, and return to step 1.*

   (b) *Else, go to step 3.*

3. *Agent j_t accepts if $\rho(j_t, i_t) \succ_{j_t} \alpha_{t-1}(j_t)$*

(a) *If j_t accepts, update α_t, μ_t*

(b) *Else, update α_t and set $\mu_t = \mu_{t-1}$*

Set $t = t+1$, and return to step 1.

Together, learning and aspirations guide pairs of men and women into "good" matches early on, so that only small adjustments are necessary in order to get to a stable match. In contrast, better and best reply dynamics are characterized by an inordinate amount of matching, breaking up, and re-matching, with adjustments toward stability fairly random.

Learning allows proposers to use their proposals more wisely. In principle, proposing agents can learn the preferences of agents on the other side, given sufficient proposals. Beyond learning all of the men preferred to himself, a man can learn which other men do not represent "competition." Thus, learning can facilitate earlier pairwise matching.

On the other side, aspirations ensure responders do not settle into inferior matches too quickly. The danger from doing so, e.g., in the best reply dynamics, is that a lot of time is spent re-matching.

One potential problem with this process is that a man may cycle through a long list of attainable women before actually securing a match. Indeed, once a man learns which other men are preferred he has no proposal to make if the women remain matched to such men.

An alternative is for a proposing man to pursue *single* women first. If a man does not have a best attainable woman available, but there are unmatched women waiting around, then it seems reasonable to go after a single woman rather than losing the opportunity to propose. If it is desirable for agents to secure early matches, then it is natural that men propose to single women first. Of course, if no single woman is available, the best a man can do is to pursue his best attainable woman, if one is available.

We are now ready to define our second proposal process.

Definition 3.5 (Random singles first)**.** *Modify the random best attainable proposal process as follows:*

2. ***Best singles proposal**: Agent i_t proposes to the best agent j_t such that*

$$\mu_{t-1}(j_t) = \emptyset$$

(a) ***Best attainable proposal**: If no best single j_t exists, i_t proposes to the best attainable j_t*

(b) *If no such j_t exists, set $\mu_t = \mu_{t-1}$, set $t = t + 1$, and return to step 1.*

(c) *Else, go to step 3.*

The random singles first process is even more biased toward early matching than the random best attainable process. It represents a more risk-averse approach, in the sense that being matched is more important than finding the best match. Combined with learning and aspirations, the process should settle more quickly into "good" matches so that the path toward stability is less volatile in terms of re-matching.

3.3 Computational results

In this section, we present computational results on convergence to stability and welfare for our two proposal processes, in both balanced and unbalanced markets.

In what follows, the aspiration adjustment for an environment θ of size n is set to decrease with the inverse of n as:

$$a_n \equiv \frac{10}{n},$$

where $n = 10$ is the smallest market we consider.

3.3.1 Convergence to stability in uncoordinated matching markets

The most natural way to simulate random sampling of a matching market $\theta \in \Theta_n$ is to sample a preference list for each agent. Since we only consider markets in which every agent prefers being matched to being unmatched, this amounts to sampling a permutation of a $\{1, 2, \ldots, n\}$. Sampling is carried out uniformly at random, independently for each agent. Sufficient random sampling in this way should capture most typical instances, but convergence of a proposal process is best characterized by convergence in the worst cases.[2]

[2] A natural question is what effect correlated preferences would have on finding a stable matching. It turns out that correlation makes finding a stable matching much easier, because

How much random sampling is needed to credibly capture the worst cases is unclear in general. In the balanced case, however, we actually know what the worst case instances look like. Consider the class of markets θ, represented as a weighted graph, in Table 3.1. A market can be represented as a weighted graph as follows: let $\omega(m,w) \in \{1,\dots,n\}$ denote the weight of edge (m,w). Then

$$m \succ_w m' \Leftrightarrow \omega(m,w) < \omega(m',w)$$
$$w \succ_m w' \Leftrightarrow \omega(m,w) > \omega(m,w')$$

Thus, the instances represented in Table 3.1 are such that a woman's favorite man ranks that woman *last* in his preferences, while a woman's least favorite man ranks her *first*; a woman's second favorite man ranks that woman second to last, and so on.

In such markets, any matching such that every woman is matched to their k^{th} ranked man is stable. To see this, suppose $k = 2$. Then the weight for each pair (m,w) must be 2, in which case a man m can only improve his partner's rank by matching with a woman for whom m is *worse* than her current partner. Thus, there are no blocking pairs. Ensuring that each woman is matched to a man just so is what prevents random better and best reply processes from finding stability.

	m_1	m_2	m_3	$\dots$	m_{n-2}	m_{n-1}	m_n
w_1	1	2	3	$\dots$	$n-2$	$n-1$	n
w_2	n	1	2	3	$\dots$	$n-2$	$n-1$
w_3	$n-1$	n	1	2	3	$\dots$	$n-2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
w_{n-1}	3	4	5	$\dots$	n	1	2
w_n	2	3	4	$\dots$	$n-1$	n	1

Table 3.1: An instance of hard preferences in balanced markets, reproduced from Ackermann et al. (2011).

The class of preferences shown in Table 3.1 are presented by Ackermann et al. (2011) as instances in which random better and best reply proposal processes can take $2^{\Omega(n)}$ steps to converge on a stable matching. Ackermann et al. (2011) do not claim that this class uniquely represents the worst case scenario. However, Hoffman et al. (2013)'s characterization of the time to reach a given stable

in some sense the correlation coordinates agent behavior.

matching in terms of the size and depth of its jealousy graph shows that such instances indeed are the most problematic for random better (and, by extension, best) reply processes.

The number of proposals needed to reach a stable matching, as a multiple of n^3, is shown in Figures 3.1-3.5. We consider the random best attainable process first.

In random balanced markets, the best attainable process appears to consistently find a stable matching at a small fraction of n^3 steps, especially as n grows large. This is illustrated in Figure 3.1 for 200 simulations from $n = 10, \ldots, 500$. However, as shown in Figure 3.2, the hard instances prove challenging. Over 150 simulations for $n = 10, \ldots, 300$, the number of proposals needed to find a stable matching is growing in n. Note that while the larger instances appear to converge at a reasonable multiple of n^3, we cap the simulations at $25n^3$. In other words, the random best attainable process appears to be exploding with n.

We now turn to the random singles first process. Figure 3.3 shows results of 500 simulations of random balanced environments for $n = 10, \ldots, 500$. Once again, the rate of convergence is a fraction of n^3, which is unsurprising as this process is an improvement over the best attainable process. Moreover, the addition of another agent on one side of the market has a negligible impact on convergence, as seen in Figure 3.4. With an additional woman in the market, the singles first proposal process still finds a stable matching at a rate less than n^3, based again on 500 simulations for $n = 10, \ldots, 500$.

In Figure 3.5, we plot convergence for hard instances of a balanced market, for $n = 10, \ldots, 500$, based on 300 simulations. Note that the time to reach stability is still bounded by a small multiple of n^3 below approximately $n = 400$. Beyond this amount, potentially exponential growth starts to take over. Still, this is stark improvement over the naive random better and best reply processes where exponential growth is immediately apparent. This is encouraging if we hope moderately large uncoordinated markets reach stability.

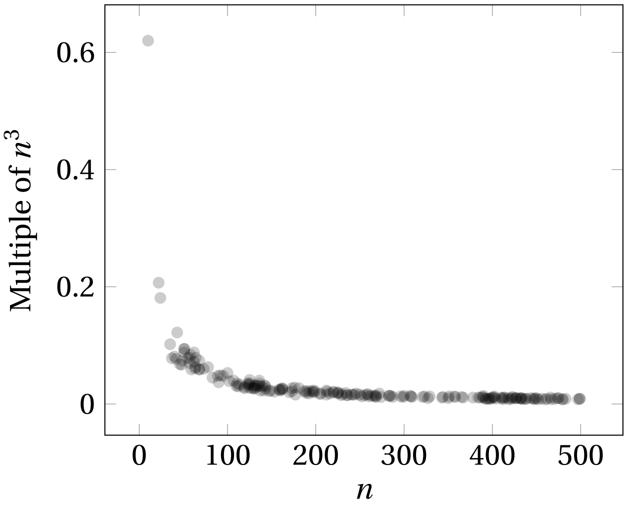

Figure 3.1: Number of proposals to reach stability in Random Best Attainable process in balanced random environment

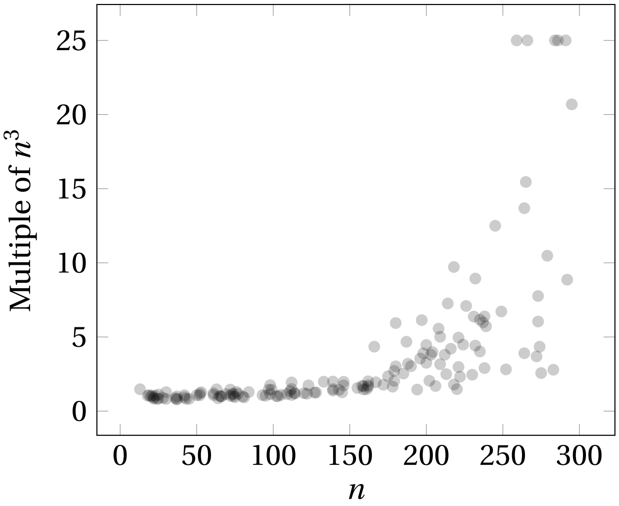

Figure 3.2: Number of proposals to reach stability in Random Best Attainable process in hard environment

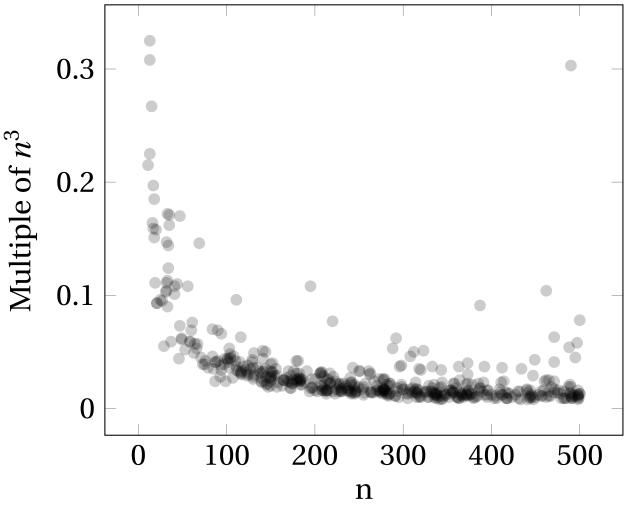

Figure 3.3: Number of proposals to reach stability in Random Singles First process in balanced random environment

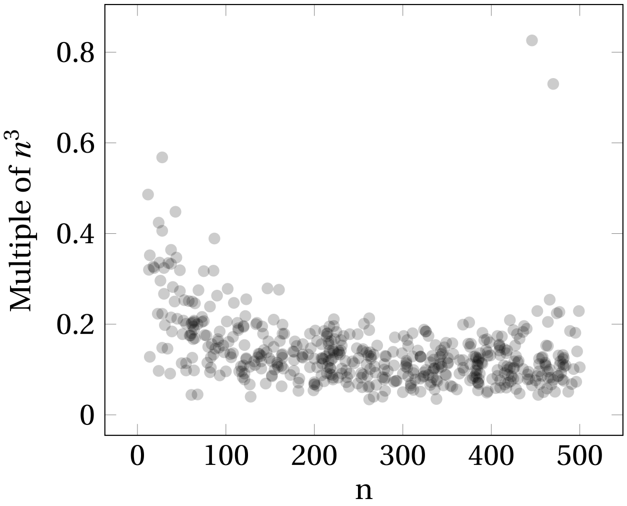

Figure 3.4: Number of proposals to reach stability in Random Singles First process in almost balanced random environment

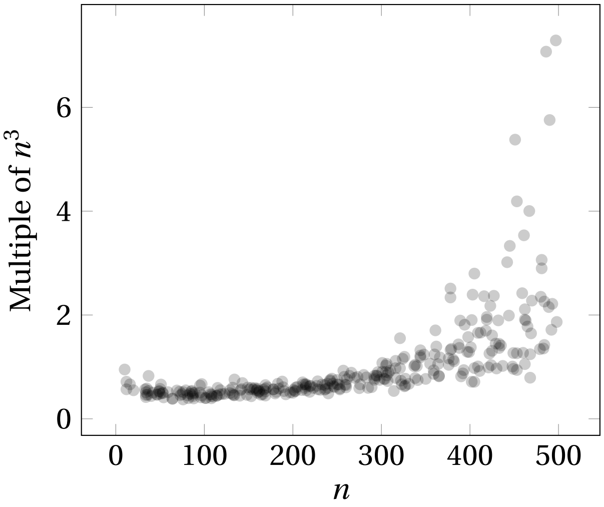

Figure 3.5: Number of proposals to reach stability in Random Singles First process in hard environment, decreasing aspiration adjustment exponentially

3.3.2 Welfare of uncoordinated markets prior to reaching stability

While a polynomial number of proposals is much more feasible than an exponential number of proposals, it can still be unrealistic for growth at rates faster than $O(n^2)$. Since the deferred acceptance algorithm can require $O(n)$ proposals from each agent to find a stable match, we should expect a randomized proposal process to take at least as long. If we interpret a proposal as an indication of interest rather than a formal proposal, a person in a market with $n = 500$ people on each side could plausibly make $1,000$ or $2,000$ proposals, corresponding to $O(n^2)$ proposals total. On the other hand, $250,000$ or $500,000$ proposals per agent—corresponding to $O(n^3)$ proposals total—pushes the bounds of believability even with a loose interpretation what counts as a proposal. Is there a downside to agents stopping their search early before reaching stability? In this section, we investigate the welfare of matches found by random proposal processes after a small multiple of n^2 proposals.

The biggest potential cost of uncoordinated matching is that too many agents are single at a given point in time. The Random Singles First process attempts to address this problem in a greedy fashion by having agents break up an existing couple only when unavoidable. Figures 3.7 and 3.6 show the proportion of single agents at multiples of n^2 for balanced and almost balanced markets respectively at $n = 50$. Nearly every agent is matched with some partner within the first n^2 proposals. Figure 3.8 of the proportion of single agents in an almost balanced market of $n = 1000$ shows this isn't simply a feature of small markets. Once again, most agents are partnered within the first n^2 proposals and, after the initial matching, about 5% of agents are single at any one time. Since most agents are matched, it is now reasonable to focus on the welfare of matched agents.

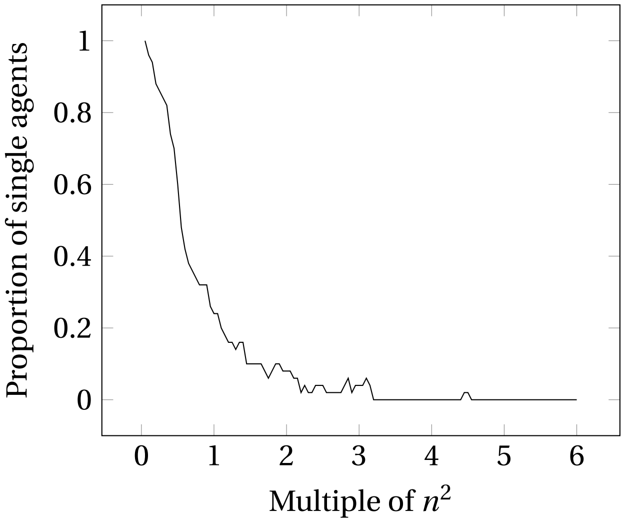

Figure 3.6: Proportion single after multiple of proposal in almost balanced environment for $n = 50$.

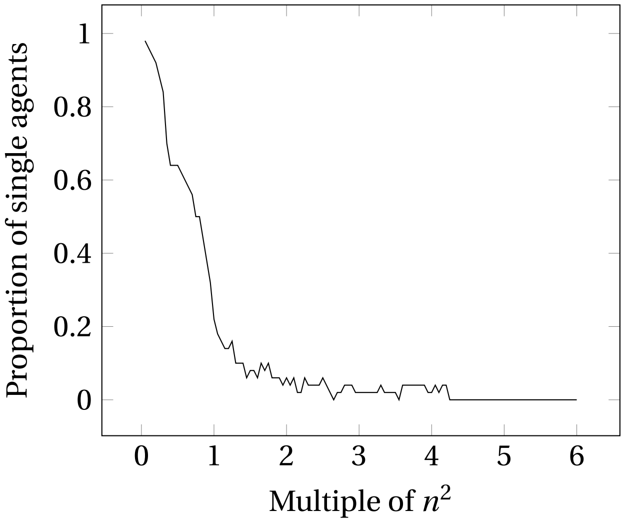

Figure 3.7: Proportion of single agents after multiple of proposal in balanced environment for $n = 50$.

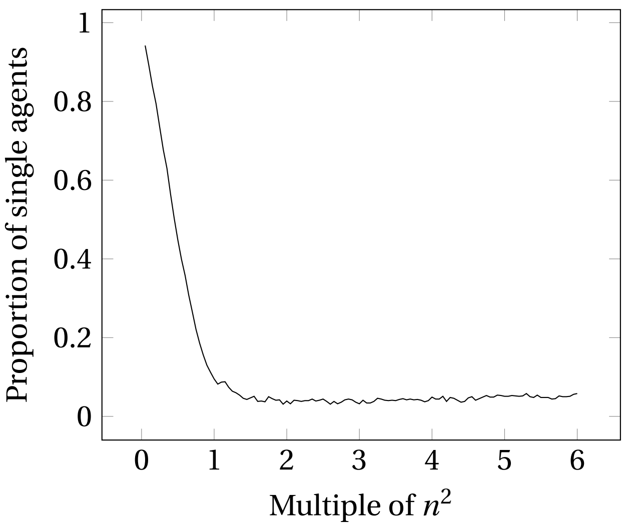

Figure 3.8: Proportion of single agents after multiple of proposal in almost balanced environment for $n = 1000$.

A natural measure of welfare in environments characterized by rank-order preferences is the average rank of agents at a given match. Consider an environment θ_n with n men and $n+m$ women. Let $\rho_t(i) \equiv \rho(i, \mu_t(i)) = |\{j : j \succeq_i \mu_t(i)\}|$ denote the *rank* of i's partner at matching μ_t. Then the *average rank* of men's partners at μ_t is $R_t^M \equiv \frac{1}{n_t} \sum_{i \in M} \rho_t(i)$, where $n_t = |\{i \in M : \mu_t(i) \in W|$ is the number of men matched to women under μ_t. If μ_{MOSM} is the man-optimal stable matching, the average rank of men at μ_{MOSM} is R^M_{MOSM}.

For a given environment θ, we can easily compute the optimal stable matches μ_{MOSM} and μ_{WOSM} in order to compare average rank at the current matching of a proposal process. In particular, we will focus on the average improvement in rank at μ_t relative to woman-optimal stable matching:

$$Q_t^M \equiv \frac{1}{n_t} \sum_{i \in M} \left(\rho_{\text{WOSM}}(i) - \rho_t(i)\right) \qquad \text{(Men)}$$

$$Q_t^W \equiv \frac{1}{n_t} \sum_{i \in W} \left(\rho_{\text{WOSM}}(i) - \rho_t(i)\right) \qquad \text{(Women)}$$

$$Q_t^T \equiv \left(Q_t^M + Q_t^W\right)/2 \qquad \text{(Total)}$$

The literature has developed a good idea of what the average ranks of optimal matches look like in terms of market size n when preferences are drawn uniformly at random. For balanced θ with n men and n women, Pittel (1989) has shown that $R^M_{\text{MOSM}} \xrightarrow{p} \log n$ and $R^M_{\text{WOSM}} \xrightarrow{p} \frac{n}{\log n}$. Ashlagi et al. (ming) explore unbalanced markets and find that the difference in welfare between the man-optimal and woman-optimal matches collapses rapidly with the addition of even a single person. In a market with n men and $n+1$ women, the average rank of the men's partners is $\log n$ and the average rank of the women's partners is $\frac{m}{\log n}$ with high probability in every stable match. In other words, being on the short side of the market provides the same advantage as being on the proposing side in a balanced market. Because the addition of a single person drastically changes the set of stable matches, we should view balanced markets as a special case. Without loss of generality, we assume men are on the short side when considering unbalanced markets.

In 500 simulated almost balanced markets with n men and $n+1$ women, all but one match found by the Random Singles First process after $5n^2$ proposals had a better average rank among all matched agents relative to the MOSM. In the sole simulation that didn't have strictly better average welfare, the process found the unique stable match within the given number of proposals. Relative to the WOSM, 485 simulations had strictly better average ranks, 12 were equal to the WOSM, and 3 were worse. Every time the total average rank was better than the WOSM average rank, men did relatively worse and women did relatively better.

As shown in figure 3.9, the short side of the market still has a better average rank than the long side in almost every instance despite doing worse than in any stable match. The short side is unable to fully use its advantage when matching is uncoordinated. Figure 3.10 shows how the average rank of men compares to the WOSM as n changes. Figure 3.11 shows the average rank across both men and women relative to the WOSM. Not only is the uncoordinated match is more egalitarian between the two sides relative to stability, it results in an overall better average ranks.

The situation is even more striking when the imbalance between sides grows. We now consider markets with n men on the short side and $1.5n$ women on the long side of the market. Figure 3.12 shows the short side of the market is roughly logarithmically worse in the uncoordinated match relative to the WOSM. With 500 men and 750 women, this means the short side is matched with their 2nd or 3rd rank partner rather than their 1st or 2nd rank partner on average. In contrast, a person on the the long side of the market is paired with their 140th best partner rather than their 220th best partner.

These results suggest that decentralized matching can be better for agents

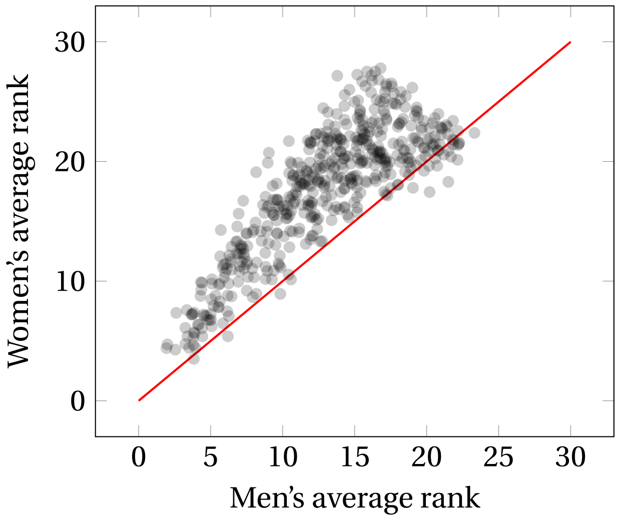

Figure 3.9: Average ranks for n men and $n+1$ women after $5n^2$ proposals.

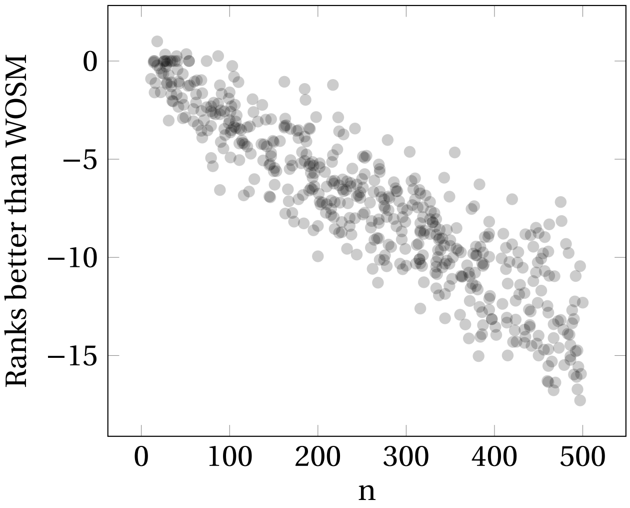

Figure 3.10: Men's average rank relative to the WOSM with n men and $n+1$ women after $5n^2$ proposals.

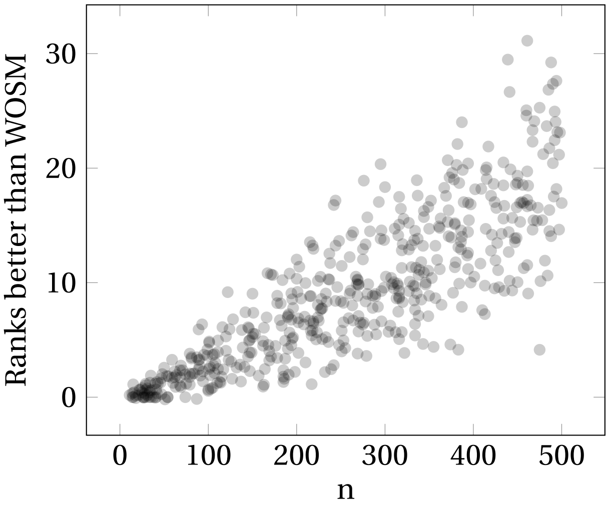

Figure 3.11: Total average rank relative to the WOSM with n men and $n+1$ women after $5n^2$ proposals.

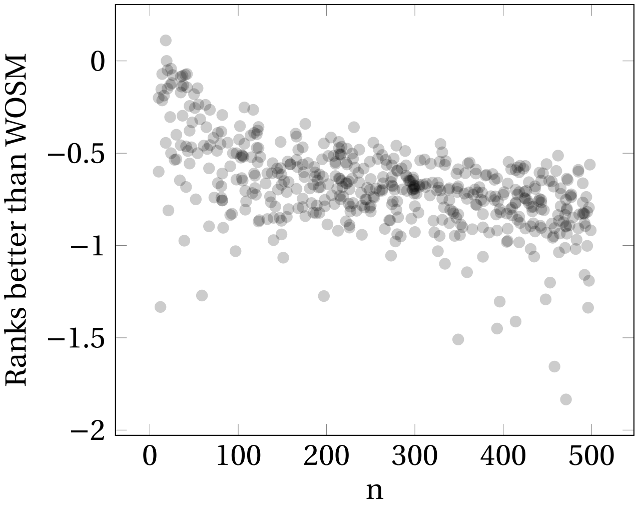

Figure 3.12: Men's average rank relative to the WOSM with n men and $1.5n$ women after $5n^2$ proposals.

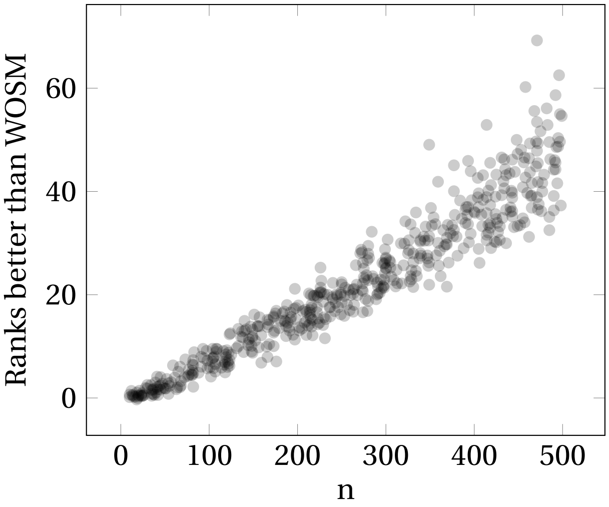

Figure 3.13: Total average rank relative to the WOSM with n men and $1.5n$ women after $5n^2$ proposals.

overall, particularly if an agent is unsure *a priori* whether they will be on the long or short side of the market. The evident hardness of finding stability in large coordinated markets ends up being an advantage rather than a failing. Unless a market is obviously failing due to instability or there is a reason all matches should happen quickly and simultaneously, this is a reason to avoid centralization.

3.4 Conclusion

We present two proposal processes that suggest uncoordinated two-sided matching markets perform well when agents aren't completely naive. While our results don't fully resolve the question of whether uncoordinated markets tend to reach stable matches, either answer turns out to be encouraging. Simulations show the random singles first proposal process reaches stability in small to medium sized markets within a small multiple of n^3 proposals. This holds even for the hardest-to-match set of preferences. On the other hand, if the process is cut-off before reaching stability, the resulting matches are more egalitarian and have better average welfare. Either way, our results suggest centralization has no advantage unless the market is unraveling or suffers another clear market failure.

Substantial work remains in the study of decentalized matching markets. The asymptotic behavior of uncoordinated matching for random preferences or more realistic correlated preferences remains an open question. Another open question is whether a more sophisticated proposal process reaches stability with high probability in polynomial time for all possible preferences.

APPENDIX

PROOFS

Proof of Proposition 1.2. Let T by any deterministic, neutral, anonymous, and robustly implementable decision rule. At least one such decision rule exists since majority rule satisfies these properties.

Suppose n is odd. I will establish the following facts about T in turn:

1. $T\big(\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n+1}{2}},(b,p_{\frac{n+3}{2}}),\ldots,(b,p_n)\big)=A, \quad \forall p_{\frac{n+3}{2}},\ldots,p_n\in[0,1]$
2. $T\big((a,1),\ldots,(a,1)\big)=A$
3. $T\big((a,p_1),(b,\frac{1}{n-1}),\ldots,(b,\frac{1}{n-1})\big)=B, \quad \forall p_1\in[0,1]$
4. $T\big((a,p_1),(a,1),\ldots,(a,1)\big)=A, \quad \forall p_1\in(0,1]$
5. $T\big(\underbrace{(a,p_1),\ldots,(a,p_{\frac{n+1}{2}})}_{\frac{n+1}{2}},(b,p_{\frac{n+3}{2}}),\ldots,(b,p_n)\big)=A, \quad \forall p_i\in(0,1)$
6. $T\big((a,p_1),\ldots,(a,p_m),(b,p_{m+1}),\ldots,(b,p_n)\big)=A, \quad \forall p_i\in(0,1),\forall m\geq\frac{n+1}{2}$

The first three facts say that majorities of various sizes map to the majority opinion when those supporters have correct beliefs. The fourth says that one member of a full majority can have an arbitrary prediction without disturbing the outcome. The fifth says that all members of a bare majority can have arbitrary interior beliefs without changing the outcome. Finally, the sixth is the conclusion of the theorem.

I prove the first fact by contradiction. Suppose there are some predictions $p'_{\frac{n+3}{2}},\ldots,p'_n$ such that

$$T\big(\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n+1}{2}},(b,p'_{\frac{n+3}{2}}),\ldots,(b,p'_n)\big)=B.$$

We must then also have

$$T\big((a,p_1),\underbrace{(a,\tfrac{1}{2}),\dots,(a,\tfrac{1}{2})}_{\frac{n-1}{2}},(b,p'_{\frac{n+3}{2}}),\dots,(b,p'_n)\big) = B, \quad \forall p_1 \in [0,1]$$

for agent one with type $(a,\frac{1}{2})$ to report truthfully, since the agent could be certain this profile will occur (consistent with the prediction of $p_1 = \frac{1}{2}$ that half of the other agents have opinion a) and thus can't expect to switch the outcome to A by reporting some other prediction. In particular,

$$T\big((a,\tfrac{n-3}{2(n-1)}),\underbrace{(a,\tfrac{1}{2}),\dots,(a,\tfrac{1}{2})}_{\frac{n-1}{2}},(b,p'_{\frac{n+3}{2}}),\dots,(b,p'_n)\big) = B.$$

Successively applying the same reasoning to all agents with opinion a yields

$$T\big(\underbrace{(a,\tfrac{n-3}{2(n-1)}),\dots,(a,\tfrac{n-3}{2(n-1))})}_{\frac{n+1}{2}},(b,p'_{\frac{n+3}{2}}),\dots,(b,p'_n)\big) = B.$$

For an agent with type $(b,\frac{1}{2})$ to report truthfully, we must have

$$T\big(\underbrace{(a,\tfrac{n-3}{2(n-1)}),\dots,(a,\tfrac{n-3}{2(n-1))})}_{\frac{n-1}{2}},(b,\tfrac{1}{2}),(b,p'_{\frac{n+3}{2}}),\dots,(b,p'_n)\big) = B$$

and then

$$T\big(\underbrace{(a,\tfrac{n-3}{2(n-1)}),\dots,(a,\tfrac{n-3}{2(n-1)})}_{\frac{n-1}{2}},\underbrace{(b,\tfrac{1}{2}),\dots,(b,\tfrac{1}{2})}_{\frac{n+1}{2}}\big) = B$$

by successively applying incentive compatibility for the remaining agents with opinion b. Applying neutrality and anonymity yields

$$T\big(\underbrace{(a,\tfrac{1}{2}),\dots,(a,\tfrac{1}{2})}_{\frac{n+1}{2}},\underbrace{(b,\tfrac{n+1}{2(n-1)}),\dots,(b,\tfrac{n+1}{2n-2})}_{\frac{n-1}{2}}\big) = A.$$

For the agents with type $(b,\frac{n+1}{2(n-1)})$ who think the previous profile is certain (consistent with their prediction) to report truthfully, the outcome cannot switch to B for any other prediction report. Changing the predictions of agents with opinion b in turn yields

$$T\big(\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n+1}{2}},(b,p'_{\frac{n+3}{2}}),\ldots,(b,p'_n)\big) = A,$$

which is the original profile assumed to map to B, resulting in a contradiction.

The second fact follows from the first. Changing an agent from opinion b and an arbitrary prediction to opinion a and an accurate prediction for that profile must leave the outcome unchanged at A for incentive compatibility. This can be repeated until all agents have opinion a. Notice that as the types of other agents change, what was once an accurate prediction might become inaccurate. Updating the prediction of an agent with opinion a to be accurate for the profile must also leave the outcome unchanged, so the prediction for each can be changed to $p_i = 1$, resulting in $T\big((a,1),\ldots,(a,1)\big) = A$.

The third fact follows from the first similarly to the second. All but one agent with opinion b can be replaced by an agent with opinion a and an accurate opinion for that profile. By anonymity, this yields

$$T\big((b,p_1),(a,\tfrac{n-2}{n-1}),\ldots,(a,\tfrac{n-2}{n-1})\big) = A, \quad \forall p_1 \in [0,1]$$

and finally by neutrality,

$$T\big((a,p_1),(b,\tfrac{1}{n-1}),\ldots,(b,\tfrac{1}{n-1})\big) = B, \quad \forall p_1 \in [0,1].$$

To establish the fourth fact, suppose agent one has type (a,p_1) with $p_1 \in (0,1]$ based on a belief that all other agents share type $(a,1)$ with probability p_1 and type $(b,\frac{1}{n-1})$ with probability $1-p_1$. If $T((a,p_1),(a,1),\ldots,(a,1)) = B$, then agent one expects the outcome from reporting truthfully to always be B by fact 3. If the agent misreported as type $(a,1)$, then the outcome would occasionally be A, producing a strictly better deviation. Therefore, we must have

$$T\big((a,p_1),(a,1),\ldots,(a,1)\big) = A, \quad \forall p_1 \in (0,1]$$

The fifth fact follows similarly to the fourth. Suppose agent one has type (a,p_1). If $p_1 \in (0,\frac{1}{2})$, consider an agent who is sure either fact 1 or 3 would apply if he reported $(a,\frac{1}{2})$. Since the choice of prediction won't change the outcome when fact 3 applies, the outcome for being honest must match the outcome when fact 1 would apply. Alternatively, if $p \in (\frac{1}{2}, 1)$, consider an agent who is certain either fact 1 or 4 would apply when reporting $(a, \frac{1}{2})$. Since this report always results in outcome A, the honest report must also result in A for the same profile of others. Taking these observations together with fact 1, we have

$$T\big((a,p_1),\underbrace{(a,\tfrac{1}{2})\ldots,(a,\tfrac{1}{2})}_{\frac{n-1}{2}},(b,p_{\frac{n+3}{2}}),\ldots,(b,p_n)\big) = A, \quad \forall p_1, p_{\frac{n+3}{2}}\ldots,p_n \in (0,1)$$

Repeating this reasoning for the remaining agents with opinion a yields

$$T\big(\underbrace{(a,p_1),\ldots,(a,p_{\frac{n+1}{2}})}_{\frac{n+1}{2}},(b,p_{\frac{n+3}{2}}),\ldots,(b,p_n)\big) = A, \quad \forall p_1,\ldots,p_n \in (0,1)$$

For the sixth fact, notice that replacing an agent with opinion b with an agent type $(a, \frac{n+1}{2(n-1)})$ in fact 5 must preserve the outcome of A. Applying the same argument as in the proof of fact 5 says any prediction $p_i \in (0,1)$, not just $\frac{n+1}{2(n-1)}$, must produce an outcome of A. This process can be repeated, adding further a supporters inductively. Therefore, any number of agents with opinion a and interior beliefs can be added, resulting in

$$T\big((a,p_1),\ldots,(a,p_m),(b,p_{m+1}),\ldots,(b,p_n)\big) = A, \quad \forall p_1,\ldots,p_n \in (0,1), \forall m \geq \frac{n+1}{2},$$

which concludes the proof that any neutral, anonymous, and robustly incentive compatible decision rule must be equivalent to majority rule when agents have interior predictions for odd n.

For even n, the first step is to establish that

$$T\big(\underbrace{(a,\tfrac{n}{2n-2}),\ldots,(a,\tfrac{n}{2n-2})}_{\frac{n}{2}+1},(b,p_{\frac{n}{2}+2}),\ldots,(b,p_n)\big) = A, \quad \forall p_{\frac{n}{2}+2},\ldots,p_n \in [0,1]$$

analogously to the first fact when n is odd. From this, the remaining facts follow, concluding with agreement with majority rule for all interior predictions when a majority exists. Furthermore,

$$T\big(\underbrace{(a,\tfrac{n-2}{2n-2}),\ldots,(a,\tfrac{n-2}{2n-2})}_{\frac{n}{2}},\underbrace{(b,\tfrac{n}{2n-2}),\ldots,(b,\tfrac{n}{2n-2})}_{\frac{n}{2}}\big) = \varnothing$$

for neutrality because this profile with correct predictions is complementary to itself. Changing the prediction of an agent with opinion a can't switch the outcome to A without giving an agent in this profile an incentive to misreport. The outcome also cannot switch to B without making a report of $(a, \frac{n-2}{2n-2})$ dominate an honest report of (a, p_i) for an agent that puts positive probability on this profile since the prediction doesn't matter for any non-balanced profile. By induction, all profiles with $\bar{x} = \frac{1}{2}$ and interior predictions must have $T(x, p) = \varnothing$ in agreement with majority rule.

□

Necessity of Proposition 1.3. Suppose agent i believes p_{-i} is fixed conditional on x_{-i}, reducing beliefs over the types of others to $\pi(x_{-i})$. Incentive compatibility implies

$$\sum_{x_{-i}} \pi(x_{-i})\, T((a, x_{-i}), (p_i, p_{-i})) \geq \sum_{x_{-i}} \pi(x_{-i})\, T((a, x_{-i}), (p'_i, p_{-i}))$$

for all p_i, p'_i, p_{-i}, and π such that $\mathrm{E}_\pi[\bar{x}_{-i}] = p_i$, so that agent i does not want to misreport her prediction p_i. Hence, T is a proper scoring rule for the mean of x_{-i} from the perspective of agent i holding $x_i = a$ fixed. By the McCarthy-Savage representation of proper scoring rules, T must be representable from the perspective of agent i as

$$T((a, x_{-i}), p) = \kappa_i(x, p_{-i}) + G_i(p_i; p_{-i}) + (\bar{x}_{-i} - p_i) G'_i(p_i; p_{-i}) \tag{1}$$

using some G_i convex in p_i, where G'_i is a subderivative in p_i. Without loss of generality, we can suppose $G_i(0; p_{-i}) = 0$ and $G'_i(0; p_{-i}) = 0$ by folding $G_i(0; p_{-i})$ and $\bar{x}_{-i} G'_i(0; p_{-i})$ into κ_i if necessary. Since $G'_i(p_i; p_{-i})$ must be non-decreasing as a subderivative of a convex function, it has bounded variation on $[0, 1]$ and we are free to write it as a Lebesgue-Stieltjes integral:

$$G'_i(p_i; p_{-i}) = \int_0^{p_i} d\xi_i(t; p_{-i}). \tag{2}$$

Then, we have

$$G_i(p_i; p_{-i}) = \int_0^{p_i} \int_0^t d\xi_i(s; p_{-i})\, dt = \int_0^{p_i} (p_i - t)\, d\xi_i(t; p_{-i}) \tag{3}$$

after a change of variables. Plugging the last two lines into line 1 yields

$$T((a,x_{-i}),p) = \kappa_i(x,p_{-i}) + \int_0^{p_i} (\bar{x}_{-i} - t)\, d\xi_i(t;p_{-i}), \tag{4}$$

which is closely related to the Schervish (1989) representation (see also Lambert (2011)). This representation prescribes the specific way that p_i and the proportion $\bar{x}_{-i}$ must interact for incentive compatibility, up to a weighting by ξ_i. For T to be neutral between A and B, we must have

$$T((b,x_{-i}),p) = \kappa_i(x,p_{-i}) - \int_0^{1-p_i} (1 - \bar{x}_{-i} - t)\, d\xi_i(t;1-p_{-i})$$

so $T(x,p) + T(1-x,1-p) = 1$. With this form for each agent, it follows by anonymity that

$$T(x,p) = \kappa(\bar{x}) + \sum_{i:x_i=a} \int_0^{p_i} (\bar{x}_{-i} - t)\, d\xi(t) - \sum_{i:x_i=b} \int_0^{1-p_i} (1 - \bar{x}_{-i} - t)\, d\xi(t),$$

since $\bar{x}$ contains all information preserved under permutations of x and ξ can't depend on the identity of the agent. Although ξ_i could have depended on the predictions of other agents to be a proper scoring rule for agent i, those predictions can only appear in their respective integrals to be proper for the remaining agents.

Again taking p_{-i} to be known conditional on x_{-i}, incentive compatibility implies T is higher in expectation when agent i reports her true type (a,p_i) than when reporting any (b,p_i'). Since the mechanism is anonymous, an agent's beliefs can be reduced to a distribution over the number of other agents $m = \sum_{j\neq -i} x_j$ with the a opinion rather than on x_{-i} directly, even if the underlying belief treats other agents asymmetrically. We have

$$\begin{aligned}
&\sum_{m=0}^{n-1}\pi(m)\Big(\kappa\left(\tfrac{m+1}{n}\right)+\int_0^{p_i}\left(\tfrac{m}{n-1}-t\right)d\xi(t)\\
&\qquad\qquad+\sum_{j:x_j=a}\int_0^{p_j}\left(\tfrac{m}{n-1}-t\right)d\xi(t)-\sum_{j:x_j=b}\int_0^{1-p_j}\left(1-\tfrac{m+1}{n-1}-t\right)d\xi(t)\Big)\\
&\geq\sum_{m=0}^{n-1}\pi(m)\Big(\kappa\left(\tfrac{m}{n}\right)-\int_0^{1-p'}\left(1-\tfrac{m}{n-1}-t\right)d\xi(t)\\
&\qquad\qquad+\sum_{j:x_j=a}\int_0^{p_j}\left(\tfrac{n_a-1}{n-1}-t\right)d\xi(t)-\sum_{j:x_j=b}\int_0^{1-p_j}\left(1-\tfrac{m}{n-1}-t\right)d\xi(t)\Big)
\end{aligned}\tag{5}$$

$$\begin{aligned}
\iff&\sum_{m=0}^{n-1}\pi(m)\left(\kappa\left(\tfrac{m+1}{n}\right)-\kappa\left(\tfrac{m}{n}\right)+\sum_{j:x_j=a}\int_0^{p_j}\tfrac{1}{n-1}\,d\xi(t)+\sum_{j:x_j=b}\int_0^{1-p_j}\tfrac{1}{n-1}\,d\xi(t)\right)\\
&\geq-\int_0^{p_i}\left(p_i-t\right)d\xi(t)-\int_0^{1-p'}\left(1-p_i-t\right)d\xi(t)
\end{aligned}\tag{6}$$

for all $p_i, p_i', p_j(x_{-i})$, and beliefs π such that $\mathrm{E}_\pi[m/(n-1)] = p_i$. The last statement is true only if

$$\sum_{n_a=0}^{n-1}\pi(m)\left(\kappa\left(\tfrac{m+1}{n}\right)-\kappa\left(\tfrac{m}{n}\right)\right)\geq-\int_0^{p_i}\left(p_i-t\right)d\xi(t)-\int_0^{1-p'}\left(1-p_i-t\right)d\xi(t),\tag{7}$$

taking $p_j(a)=0$ and $p_j(b)=1$. This inequality says that the expectation of κ's first differences must be greater than a function of the mean of the distribution. Following a similar argument for agents with opinion b yields the differences in κ having the lower bound

$$\sum_{m=0}^{n-1}\pi(m)\left(\kappa\left(\tfrac{m+1}{n}\right)-\kappa\left(\tfrac{m}{n}\right)\right)\geq-\int_0^{p'}\left(p_i-t\right)d\xi(t)-\int_0^{1-p_i}\left(1-p_i-t\right)d\xi(t)\tag{8}$$

Since the right-hand side of each lower bound is quasi-convex in p' (non-increasing at $p'<p_i$ and non-decreasing at $p'>p_i$), each inequality is satisfied for all p' if and only if it holds for $p'\in\{0,1\}$. Combined, these yields

$$\begin{aligned}
\sum_{m=0}^{n-1}\pi(m)\left(\kappa\left(\tfrac{m+1}{n}\right)-\kappa\left(\tfrac{m}{n}\right)\right)\geq\max\Big\{&-\int_0^{p_i}\left(p_i-t\right)d\xi(t),\\
&-\int_0^{p_i}\left(p_i-t\right)d\xi(t)-\int_0^{1}\left(1-p_i-t\right)d\xi(t),\\
&-\int_0^{1-p_i}\left(1-p_i-t\right)d\xi(t),\\
&-\int_0^{1}\left(p_i-t\right)d\xi(t)-\int_0^{1-p_i}\left(1-p_i-t\right)d\xi(t)\Big\}
\end{aligned}\tag{9}$$

for all $p_i \in [0,1]$ and all π such that $\mathrm{E}_\pi[m/(n-1)] = p_i$.

A lower bound on the expectations of κ's differences for all distributions is equivalent to the differences being separated from the right-hand side by some convex function of p_i. The four quantities in the lower bound are each concave in p_i. The first and fourth are maximized at zero while the second and third are maximized at one, as can be seen by taking first-order conditions via Leibniz's rule. Since the first and fourth are symmetric around $\frac{1}{2}$ with the third and second respectively, attention can be restricted to the second and third terms when considering $p_i \geq \frac{1}{2}$.

Since the terms of the lower bound are concave in p_i, the least restrictive convex upper bound for each term is a supporting line at some point in $[\frac{1}{2}, 1]$. The supporting line of the second term at ϕ_1 is

$$\begin{aligned}&(p_i - \phi_1)\left(-\int_0^{\phi_1} d\xi(t) + \int_0^1 d\xi(t)\right) - \int_0^{\phi_1} \left(\phi_1 - t\right) d\xi(t) - \int_0^1 \left(1 - \phi_1 - t\right) d\xi(t) = \\ &-\int_0^{\phi_1} \left(p_i - t\right) d\xi(t) - \int_0^1 \left(1 - p_i - t\right) d\xi(t)\end{aligned} \tag{10}$$

and the supporting line of the third term at ϕ_2 is

$$(p_i - \phi_2)\left(\int_0^{1-\phi_2} d\xi(t)\right) - \int_0^{1-\phi_2} \left(1 - \phi_2 - t\right) d\xi(t) = -\int_0^{1-\phi_2} \left(1 - p_i - t\right) d\xi(t) \tag{11}$$

The pointwise maximum of the supporting lines is convex and increasing, so this provides a minimal bound of the differences in κ above $\frac{1}{2}$. Define $\delta(m)$ for $m+1 \geq \lceil \frac{n}{2} \rceil$ as

$$\begin{aligned}\delta(m) = \max\bigg\{ &-\int_0^{\phi_1} \left(\tfrac{m}{n-1} - t\right) d\xi(t) - \int_0^1 \left(1 - \tfrac{m}{n-1} - t\right) d\xi(t), \\ &-\int_0^{1-\phi_2} \left(1 - \tfrac{m}{n-1} - t\right) d\xi(t)\bigg\}\end{aligned} \tag{12}$$

Then, we have $\kappa\left(\frac{m+1}{n}\right) \geq \kappa\left(\frac{m}{n}\right) + \delta(m)$ for $m+1 \geq \lceil \frac{n}{2} \rceil$ when the expectation in line (9) is evaluated at degenerate distributions. Neutrality implies $\kappa\left(\frac{1}{2}\right) = \frac{1}{2}$ and $\kappa(\bar{x}) + \kappa(1-\bar{x}) = 1$, so without loss of generality

$$\kappa\left(\tfrac{n_a}{n}\right) = \frac{1}{2} + \tau\left(\tfrac{n_a}{n} - \tfrac{1}{2}\right) + \mathbb{1}(n \text{ odd})\frac{\delta\left(\frac{n-1}{2}\right)}{2} + \sum_{m=\lceil n/2 \rceil}^{n_a - 1} \delta(m) \tag{13}$$

for $n_a \geq \lceil \frac{n}{2} \rceil$ with non-decreasing $\tau : [0, \frac{1}{2}] \to \mathbb{R}_+$ to account for excess differences in κ above $\delta(m)$. Since the base score κ must be negatively symmetric around $\frac{1}{2}$, we then have

$$\kappa\left(\tfrac{n_a}{n}\right) = \frac{1}{2} + \operatorname{sign}\left(\tfrac{n_a}{n} - \tfrac{1}{2}\right)\left(\tau\left(\left|\tfrac{n_a}{n} - \tfrac{1}{2}\right|\right) + \mathbb{1}(n \text{ odd})\frac{\delta\left(\frac{n-1}{2}\right)}{2} + \sum_{m=\lceil n/2 \rceil}^{\max\{n_a, n_b\}-1} \delta(m)\right) \tag{14}$$

for all n_a. Without loss of generality, a scaling factor of $\frac{1}{n}$ could have been applied to each scoring rule originally and carried through, resulting in the statement of the theorem. □

Sufficiency of Proposition 1.3. The sufficiency of this representation follows from iterated deletion of interim dominated strategies in the direct mechanism. Consider an agent of type (a, p_i) who conjectures the average proportion of reported opinions is $\hat{p}_i$. By the conditions on the base score, a report of $(a, \hat{p}_i)$ weakly prefers good as all reports (b, p'). A comparison of lines (5) and (6) above shows the agent will strictly prefer $(a, \hat{p}_i)$ to (b, p') as long as the agent thinks there is some chance that p_j and $1 - p_j$ (when $x_j = a$ and $x_j = b$, respectively) are outside a neighborhood of zero where $\xi(t)$ is uniformly zero. Otherwise, a strict incentive from a strictly increasing τ or partial honesty is necessary to guarantee dominance. An analogous argument for agents of type (b, p_i) rules out all (a, p'). Since each agent prefers submitting their true opinion, it follows that each agent weakly prefers submitting their true prediction of the opinions of other agents since T is a proper scoring rule for each agent. Consequently, honest reporting always survives iterated deletion of weakly interim dominated strategies. Other strategies might also survive if agents are indifferent between these reports and honesty, but all will result in the same outcome as honest reporting indifference occurs only when T is constant, with ξ uniformly zero in some interval containing those reports. Therefore, the unique dominance solvable outcome for type profile (x, p) conincides with $T(x, p)$. □

REFERENCES

ABDULKADIROĞLU, A., PATHAK, P. A., AND ROTH, A. E. 2005a. The new york city high school match. *American Economic Review*, 364–367.

ABDULKADIROĞLU, A., PATHAK, P. A., ROTH, A. E., AND SÖNMEZ, T. 2005b. The boston public school match. *American Economic Review*, 368–371.

ACKERMANN, H., GOLDBERG, P. W., MIRROKNI, V. S., RÖGLIN, H., AND VÖCKING, B. 2011. Uncoordinated two-sided matching markets. *SIAM Journal on Computing 40,* 1, 92–106.

ASHLAGI, I., KANORIA, Y., AND LESHNO, J. D. Forthcoming. Unbalanced random matching markets: The stark effect of competition. *Journal of Political Economy*.

AUSTEN-SMITH, D. 1993. Interested experts and policy advice: multiple referrals under open rule. *Games and Economic Behavior 5,* 1, 3–43.

AUSTEN-SMITH, D. AND BANKS, J. S. 1996. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review 90,* 1, 34–45.

BATTAGLINI, M. 2004. Policy advice with imperfectly informed experts. *Advances in Theoretical Economics 4,* 1.

BRIER, G. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78,* 1, 1–3.

CARVALHO, A. AND LARSON, K. 2012. Sharing rewards among strangers based on peer evaluations. *Decision Analysis 9,* 3, 253–273.

CHWE, M. S.-Y. 2010. Anonymous Procedures for Condorcet's Model: Robustness, Nonmonotonicity, and Optimality. *Quarterly Journal of Political Science 5,* 1, 45–70.

CRAWFORD, V. P. AND SOBEL, J. 1982. Strategic information transmission. *Econometrica 50,* 6, 1431–1451.

CRÉMER, J. AND MCLEAN, R. 1988. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society 56,* 6, 1247–1257.

DASGUPTA, A. AND GHOSH, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proc. of 22nd Int. World Wide Web Conf. (WWW 2013).* 319–329.

DAWES, R. M. 1989. Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology 25,* 1, 1–17.

DUGGAN, J. AND MARTINELLI, C. 2001. A Bayesian Model of Voting in Juries. *Games and Economic Behavior 37,* 2, 259–294.

FEDDERSEN, T. AND PESENDORFER, W. 1997. Voting behavior and information aggregation in elections with private information. *Econometrica 65,* 5, 1029–1058.

FEDDERSEN, T. AND PESENDORFER, W. 1998. Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting. *American Political Science Review 92,* 1, 23–35.

GALE, D. AND SHAPLEY, L. S. 1962. College admissions and the stability of marriage. *The American Mathematical Monthly 69,* 1, 9–15.

GERARDI, D. 2000. Jury verdicts and preference diversity. *American Political Science Review 94,* 2, 395–406.

GERARDI, D., MCLEAN, R., AND POSTLEWAITE, A. 2009. Aggregation of expert opinions. *Games and Economic Behavior 65,* 2, 339–371.

GHOSH, A. AND ROTH, A. 2015. Selling privacy at auction. *Games and Economic Behavior 91,* 334–346.

GLAZER, J. AND RUBINSTEIN, A. 2004. On optimal rules of persuasion. *Econometrica 72,* 6, 1715–1736.

GNEITING, T. AND RAFTERY, A. E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association 102,* 477, 359–378.

GOOD, I. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological) 14,* 1, 107–114.

GROFMAN, B., OWEN, G., AND FELD, S. L. 1983. Thirteen theorems in search of the truth. *Theory and Decision 15,* 261–278.

GUARNASCHELLI, S., MCKELVEY, R. D., AND PALFREY, T. R. 2000. An Experimental Study of Jury Decision Rules. *American Political Science Review 94,* 2, 407–423.

HOFFMAN, M., MOELLER, D., AND PATURI, R. 2013. Jealousy graphs: Structure and complexity of decentralized stable matching. In *Web and Internet Economics*. Springer, 263–276.

HOSSAIN, T. AND OKUI, R. 2013. The binarized scoring rule. *The Review of Economic Studies 80,* 3, 984–1001.

JURCA, R. AND FALTINGS, B. 2007. Collusion-resistant, incentive-compatible feedback payments. In *Proc. of 8th ACM Conf. on Electronic Commerce (EC 2007)*. 200–209.

JURCA, R. AND FALTINGS, B. 2011. Incentives for answering hypothetical questions. In *Proc. of 1st Workshop on Social Computing and User Generated Content (EC 2011 Workshop)*.

KRISHNA, V. AND MORGAN, J. 2001. A model of expertise. *Quarterly Journal of Economics 116,* 2, 747–775.

LAMBERT, N. S. 2011. Elicitation and evaluation of statistical forecasts.

LI, H., ROSEN, S., AND SUEN, W. 2001. Conflicts and Common Interests in Committees. *American Economic Review 91,* 5, 1478–1497.

MARKS, G. AND MILLER, N. 1987. Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin 102,* 1, 72–90.

MILLER, N., RESNICK, P., AND ZECKHAUSER, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science 51,* 9, 1359–1373.

NITZAN, S. AND PAROUSH, J. 1982. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review 23,* 2, 289–297.

PITTEL, B. 1989. The average number of stable matchings. *SIAM Journal on Discrete Mathematics 2,* 4, 530–549.

PRELEC, D. 2004. A Bayesian truth serum for subjective data. *Science 306,* 5695, 462–466.

PRELEC, D. AND SEUNG, H. S. 2007. An algorithm that finds truth even if most people are wrong.

PRELEC, D., SEUNG, H. S., AND MCCOY, J. 2014. Finding truth even if the crowd is wrong.

RADANOVIC, G. AND FALTINGS, B. 2013. A robust Bayesian truth serum for non-binary signals. In *Proc. of the 27th AAAI Conf. on Artificial Intelligence (AAAI 2013)*. 833–839.

RILEY, B. 2014. Minimum Truth Serums with Optional Predictions.

ROTH, A. E. 1984. The evolution of the labor market for medical interns and residents: a case study in game theory. *The Journal of Political Economy*, 991–1016.

ROTH, A. E. 2008. Deferred acceptance algorithms: History, theory, practice, and open questions. *international Journal of game Theory 36,* 3-4, 537–569.

ROTH, A. E. AND VANDE VATE, J. H. 1990. Random paths to stability in two-sided matching. *Econometrica: Journal of the Econometric Society*, 1475–1480.

ROWAN, T. 1990. Functional Stability Analysis of Numerical Algorithms. Ph.D. thesis, University of Texas at Austin.

SAVAGE, L. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66,* 336, 738–801.

SCHERVISH, M. J. 1989. A general method for comparing probability assessors. *The Annals of Statistics 17,* 4, 1856–1879.

SCHLAG, K. AND VAN DER WEELE, J. 2009. Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. *Theoretical Economics Letters 2013,* February, 38–42.

WITKOWSKI, J., BACHRACH, Y., KEY, P., AND PARKES, D. 2013. Dwelling on the negative: Incentivizing effort in peer prediction. In *Proc. of 1st AAAI Conf. on Human Computation and Crowdsourcing*.

WITKOWSKI, J. AND PARKES, D. 2012a. A robust Bayesian truth serum for small populations. In *Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI 2012)*.

WITKOWSKI, J. AND PARKES, D. 2012b. Peer prediction without a common prior. In *Proc. of the 13th ACM Conference on Electronic Commerce (EC 2012)*.

WOLINSKY, A. 2002. Eliciting information from multiple experts. *Games and Economic Behavior 41,* 1, 141–160.

YANIV, I. AND KLEINBERGER, E. 2000. Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational behavior and human decision processes 83,* 2, 260–281.

ZHANG, P. AND CHEN, Y. 2014. Elicitability and knowledge-free elicitation with peer prediction. In *Proc. of the 2014 Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2014)*. 245–252.