Making Peer Prediction Practical

A DISSERTATION PRESENTED
BY
VICTOR SHNAYDER
TO
THE JOHN A. PAULSON SCHOOL OF ENGINEERING AND APPLIED SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
JUNE 2016

Thesis advisor: Professor David C. Parkes Victor Shnayder

Making Peer Prediction Practical

Abstract

My dissertation is on crowdsourcing—using crowds of people to accomplish tasks that are impractical or far more expensive otherwise. I focus specifically on crowdsourcing of information, where workers do tasks such as analyze images, translate sentences, report whether a cafe has public wi-fi, or assess the writing quality of essays. To encourage participation, workers can be paid, or given non-monetary rewards. In many applications, it is difficult to assess whether responses from a large crowd are accurate, and this can tempt workers into submitting nonsense, allowing them to complete tasks faster and get higher rewards. There are a number of ways to detect this and encourage worker effort and accurate reporting; I apply a technique called peer prediction, which rewards workers based on patterns of agreement among their reports.

I am particularly motivated by the challenge of providing education at scale: how to enable billions of people to learn what they want, at a cost even the very poor can afford. Specifically, I study peer assessment of open-ended assignments as a way to scale human feedback. I treat this as a crowdsourcing problem, and study how peer prediction can encourage effort and accurate assessment when students give feedback to their peers.

Previous work in peer prediction has highlighted the need for reward mechanisms where exerting effort and reporting truthfully is better for workers than other reporting strategies. I make three main contributions: I present a new Correlated Agreement mechanism for peer prediction in multi-signal environments, that guarantees that uninformed reporting is less attractive than being truthful. I show that replicator dynamics is a useful tool to analyze the likelihood of truthful behavior and its stability when workers are not assumed to be fully rational, and learn from experience instead. Finally, I analyze a dataset of three million peer assessments from online courses on the edX platform, studying several challenges for using peer prediction for peer assessment in education: reward variability, reward magnitude, and low-effort reporting. I compare several peer prediction mechanisms, and conclude that peer prediction is a promising technique in this domain when combined with other efforts to improve feedback quality.

Contents

Listing of figures

Acknowledgments

It's been a long and winding journey, and I would never have finished without encouragement and support from many people.

First, thanks to my advisor David, who first introduced me to EconCS, and later agreed to take me as a student when he was already overworked, and put up with my distractions by teaching and edX. Thank you for your quick and insightful feedback on research, your ability to rephrase convoluted ideas in clear language, and your permanently positive attitude.

Thanks to Greg for funding the first year of my (first!) return to grad school, and agreeing to still be on my committee while being a dean at Cornell. Thanks to Yiling for helping guide me and Mike on our decision markets project, for being on my committee, and for asking great questions.

Thanks to Matt Welsh for admitting me to grad school many years ago, and helping me start learning how to do research and write about it.

Thanks to Jim Waldo for great advice about passing quals, finishing dissertations, and life in general. Thanks to Brian Kernighan for advice from fifteen years ago, on the importance of getting to a can-declare-success point with projects.

Thanks to my co-authors on the research described here and on earlier papers: Raf Frongillo, for staying up past midnight to help submit two papers earlier this year, your knowledge of math, optimization, machine learning, and your ability to summarize the main story. Arpit Agarwal, thanks for proving theorems, calling in from India late at night, and asking good questions. Mike, Ian, Vikas, GWA, Konrad, Bor-rong, Mark—I learned a lot about research from working with you.

Thanks also to the anonymous reviewers for EC, IJCAI, and AAAI—your feedback helped improve this dissertation, and inspired research in new directions.

Thanks to many EconCS graduate students over the years for many interesting discussions, especially Jens, John L, Alice, and Bo.

Thanks to Ann Marie for always being friendly and helpful.

This dissertation is a research document, but there was a lot more along the way—learning and teaching, friendships, and staying sane.

John Girash and many others at the Bok Center helped me learn to teach and also to learn. Nancy Houfek taught me to value and enjoy public speaking, and to not pass out.

Thanks to Tegan, Alison, Jay, Mike H, Robin, Scot, Heather, Mackenzie, Phil, Tom, and everyone else in the Harvard Outing Club for joining in adventures, dinner parties, and sundry silliness.

Thanks all the instructors at Harvard Aikikai, and to Sally, Jan, and the rest of the aikido students over the years. It's a great place to be in the moment, and I learned a lot.

Thanks to Atul for pushing me to finish what I started, and many long conversations about how people and societies really work.

Thanks to Tom for discussions about education, ideas on what really matters based on your teaching experience.

Thanks to the OrangeSox.

My last detour in graduate school was to edX, and everyone there was supportive of my working on research part time. Thanks especially to Anant, Rob, Lou, Beth, Wendy, Tena, Nina, and Piotr, for encouragement, permission to study edX data, and good questions and suggestions. Thanks to the LMS, Analytics, and Product teams at edX for being great to work with and great to have fun with.

Thanks to the NumPy, Pandas, Python, and Mathematica teams for building great tools, and to Jordan Suchow for the LaTeX dissertation template.

Thanks to my parents and grandma for your unconditional encouragement and support.

Finally, thank you Anne for encouraging me, putting up with my distraction for the last few years, and for naming the delta matrix. Now, on to new projects!

1

Introduction

It is commonly said that two minds are better than one. By induction, it follows that many minds are even better, and an entire crowd must be best. In this dissertation, I study *crowdsourcing*—using crowds of people to accomplish tasks that are impractical or far too expensive otherwise. Specifically, I focus on encouraging effort in information reporting tasks via a technique called *peer prediction*. I will show that peer prediction is a practical way to encourage effort in crowdsourcing, including in peer assessment in education.

Crowdsourcing is primarily an internet-enabled phenomenon,[1] taking advantage of easy on-

[1] Though in some sense, crowdsourcing goes back at least to Charles Darwin's survey-based studies of the universality of emotion in the 1860s.

line coordination to recruit people to perform tasks. Tasks performed by the "crowd" range from the very complex and skilled, such as inventing solutions to mechanical or chemical problems posted by companies (e.g. InnoCentive), to simpler creative tasks, such as designing logos, icons, or other visuals (e.g. 99 Designs), to tasks that are easy for humans but still hard to automate, such as labeling images, or data entry (e.g. Captricity).

Crowdsourcing can reduce costs, enable novel solutions by matching people to problems, and educate and engage. Relatively simple tasks, such as labeling media, evaluating the quality of search results, and audio transcription, are usually crowdsourced to save money—it is cheaper to pay a crowd of workers than hire full-time employees, and crowds can easily adjust to variability in the timing and amount of work. In more complex tasks such as crowdsourcing solutions to problems, the goal is as much to reach a wide variety of people as to save money—it is unknown what previous experiences will inspire a solution, so hiring the "right" person is difficult, and a crowdsourcing platform may help bring the challenge to the attention of just the right person. Finally, some applications have an educational or engagement purpose: one example is crowdsourced citizen science, with hundreds of projects where non-scientists help track the migration of birds and whales, analyze astronomical images, and gather and analyze data in many other scientific projects. While reduced cost is often a benefit, making some projects feasible—e.g. there is no way a lab could hire enough staff to track the migration of birds across the whole country—getting more people to engage with real science is a key motivation.

I focus specifically on crowdsourcing for information elicitation—asking workers to report something, either based on their experiences, as in soliciting info about businesses they have visited, or based on examination or analysis of an object, as in translation of a sentence, labeling objects in an image, or identifying the topic of an essay.

I am especially motivated by the challenge of providing education at scale: how to enable billions of people to learn what they want, at a cost even the very poor can afford. The internet has been a massive help already, with general sites likes Wikipedia and YouTube as well as specific sites and online courses on many many topics. The trend of ever-increasing access to content goes back to public libraries and even the printing press, but access is only a small part of education. Feedback, mentoring, engagement, support, applying learning, and many other aspects are also crucial. Recent efforts with Massive Open Online Courses (MOOCs), such as edX, Coursera, Udacity, and others, start to tackle some of these aspects by providing community and guidance along with well-designed content, but there are still boundless opportunities and a critical need to do better.

1.1 Crowdsourcing in Education

Of course, scaling education is too big a problem for one person and definitely for one dissertation. There are many directions to pursue. In this dissertation, I follow one thread—to put this work in context, I will start with the high-level questions, and narrow down to the specific problems I study. There are many other questions that can be considered at each level, providing vast opportunities for other work in this space. While the discussion here is in the context of education, most of my contributions apply more broadly to crowdsourcing for information elicitation.

I start with a high level question: how can we effectively teach topics that cannot yet be automatically assessed, at large scale and with minimal intervention by professors and teaching assistants? As a concrete example, consider teaching writing, and set the bar high—not just to have students memorize rules and get multiple choice questions correct, but take students with minimal writing skills, and help them learn to craft effective, concise, professional-level prose. Putting together such a program would require assembling many components: explanations, examples, tutorials, motivational videos, simple assessments to choose the right word, and many more. One absolutely key component would be for students to actually write, and to receive useful feedback on their writing.

I take as given that hiring enough professional teachers to give such feedback is not always possible, and instead tackle this problem by assuming that the course team can do a lot of preparation and design, then monitor the course and intervene occasionally, but that the majority of feedback must come from the students themselves. Even with this constraint, there are many ways to organize such peer-to-peer learning. I focus on peer assessment, where students each write something, then evaluate the work of several peers.

Building a great peer evaluation system where students learn as much as possible requires many decisions—designing the user experience, creating questions and evaluation rubrics, fitting peer evaluation into the overall structure of the course, and choosing the feedback to solicit from peers and the feedback given to evaluators themselves. For peer assessment to work, students need to be prepared enough to do a good job, and motivated to do so. Proper preparation is a pedagogical issue—the course designer must provide enough scaffolding, so students know enough to assess competently. Motivation and effort can be tackled in a number of ways. I focus on explicit evaluation of peer feedback, computing grades that will reward students who provide accurate feedback. I treat peer evaluation as an information elicitation crowdsourcing task, and use peer prediction to encourage accurate reporting.

1.2 Peer Prediction and Related Research Areas

In this section, I introduce peer prediction in crowdsourcing and related research areas. The subsequent chapters each include a more detailed and more technical discussion of related research.

The basic idea of peer prediction is simple: when there is no ground truth that can be used to evaluate reported information, reward agreement with other reporters instead. A key challenge is to define the rewards so that it is difficult to cheat, getting high scores without reporting accurately. My dissertation describes a new peer prediction mechanism that meets this challenge, shows a new way to evaluate the stability of truthfulness of peer prediction mechanisms, and considers how peer prediction could be applied to peer assessment in education.

I will use three crowdsourcing applications to explain peer prediction and related issues. The first application is soliciting information for online mapping:

Example 1. *Modern mapping services like Google Maps provide information about businesses to help users search and decide where to go. Sometimes, this information is entered by the business owners; other times, it is sourced from users who have visited the business. An example question is "does* The Hungry Student Cafe *have free public wi-fi?" The possible answers are yes and no.*

Here, the task is to answer a simple, fairly objective[2] yes-no question. At the same time, note that it is difficult to check the answer without physically sending someone to the cafe.

Our second application is labeling the emotions of film snippets to train an automated classifier:

Example 2. *Researchers who study emotions have long created labeled data sets of images and videos corresponding to particular emotions. Today, crowdsourcing can make this process easier and cheaper, enabling larger scale, which could be used e.g., to train an automated classifier for improving video search. A typical task is "what emotion is most evident in the following two minute clip of* The Godfather*?", with the answer either open ended, or selected from a list: e.g. happiness, sadness, joy, confusion, surprise.*

Here, the task is less objective, especially if answers are open-ended. Doing it properly requires watching the video, which takes time.

Our final application is peer assessment in online courses:

Example 3. *Students in a course on environmental issues are asked to write a short essay on the concerns and conflicts in water usage in their country or local jurisdiction, then assess the submission of several peers. One part of such an assessment may be to evaluate the organization of the peer's essay, ranging from "weak: no evidence of structure; paragraphs and sentences are not connected" to "ok: essay is orga-*

[2] Though consider that wi-fi may be turned off at certain times, or may not work for a particular person.

nized logically, but lacks clear transitions between parts and structure is not made explicit" to "great: clear structure, made clear in the introduction, with smooth transitions between parts". The student selects a particular option, and has a space to give open ended feedback.

In peer assessment, a good rubric is crucial to make the evaluations of different students likely to agree with each other and with the assessment that members of the course team would give.

In these and other crowdsourcing applications, there is a question of trust: will the people performing the tasks do a good job even if no one checks their work? In applications like citizen science, where the user is a volunteer motivated by the scientific topic at hand, expects no reward, and has no reasons to corrupt the results, it is reasonable to trust users, though they may still be unqualified or unskilled. In our video emotion tagging example and other data entry and data tagging applications, users are often paid per task, there is less intrinsic satisfaction, and it may be tempting to maximize rewards by submitting random or constant answers, thus completing tasks faster. Peer assessment is an in-between situation: there is no immediate financial motivation to do it quickly, but reviewing can be hard work, and it can be tempting to take a cursory glance and pick a number. In traditional classes, this is mitigated by instructor oversight. In early MOOCs, the lack of such oversight was mitigated by the fact that students were self-motivated and usually there to learn, not to earn a credential. As the importance of credentials for online courses goes up, there will be more students who just want a good grade, and are happy to slack off or even deliberately cheat if they can get away with it, and so grading peer assessments will be more important.

So, what can we do to deter bad reporting? The simplest thing is to check student work, and punish bad behavior. Staff can spot check some evaluations, or in some cases, "gold standard" submissions with known answers can be intermixed with the real peer submissions. These approaches can work, but have some downsides. Spot checking requires hiring trusted staff, which can be expensive, and using gold standard submissions incurs the expense of creating the gold standards, and wastes work by thousands of students, who assess the same known-answer submission. There is also a risk that users will learn to recognize the gold standards. In applications where multiple users perform each task, such as peer assessment, peer prediction can be used to avoid or to complement other methods.

The simplest peer prediction mechanism is *output agreement*, described here informally—given a user's report on a task, pick a reference report from a different user who did the same task, and reward the user only if their report matches the reference report. Assuming that other users are likely to report the "right" answer, it is usually optimal to do the same.

As a user, this gives you a reason to try to match your peer's reports. It also raises some questions: will I be punished if my reference peer is incompetent? What if I see something in the task that I suspect my peer will miss? Should I report the more likely answer instead of what I really think? If users can coordinate, even implicitly, there is also a collusion concern—if all users report the same thing for each task, they always get rewarded, and the system gets no useful information.

A number of papers have studied these and other issues, resulting in many variants of peer prediction. I introduce the broad strokes here, and Chapter 2 has a more detailed survey.

Most of the work is based on economics and game theory. Here is a non-mathematical introduction to the key ideas: users of a crowdsourcing system are *agents* who play a *game*, where they observe a *signal* that corresponds to their observations of the task. They use a *strategy* to decide whether to report that signal and any other requested information truthfully, or report something else. The *mechanism* defines the reward each agent gets for their *report*, and agents are typically assumed to try to maximize their rewards. A mechanism is usually considered good if *truthfulness*, the strategy of carefully observing the agent's signal and reporting it straightforwardly, is an *equilibrium*, with each agent maximizing their expected reward assuming that all others are truthful too (See Leyton-Brown & Shoham (2008) for a general introduction to game theory). There is significant variation in the details—further assumptions and properties a mechanism guarantees—leading to a variety of approaches.

Mechanisms vary in the information elicited: some require agents to make a *prediction report* about the likely reports of others in addition to their signal (Prelec, 2004; Witkowski & Parkes, 2012a). This can make for stronger results, at the cost of complexity and more assumptions about the knowledge and rationality of agents. In contrast, other mechanisms are *minimal*, requiring only the signals to be reported (Miller et al., 2005; Jurca & Faltings, 2009). I am convinced that minimal mechanisms are far more practical—reporting probabilities about what other people might do is usually too much to ask—and only study this subclass.

Another dimension is the amount of information required to configure the mechanism. In some cases, the rewards depend not only on the reports, but also on a precise probabilistic model of how agents obtain signals (Miller et al., 2005). Other mechanisms do not use such information directly, but require certain assumptions about the world model to hold. For example, the output agreement mechanism does not use information about the model, but for it to be truthful, agents must believe that their signal is the most likely signal for their reference peer to observe (Waggoner & Chen, 2014). This may not hold, e.g. in a product review scenario, where users report whether the product is good—an agent may have a bad experience, but still believe that others are likely to have a good one. Some mechanisms manage to get strong results with only minimal assumptions on the signal patterns (typically that multiple agents' observations are correlated somehow, not independent) (Dasgupta & Ghosh, 2013).

A third distinction is whether the mechanism is *single-task* or *multi-task*—multi-task mechanisms require that there are several similar tasks, and have further variants: some require that each agent do several tasks, others that there are many tasks, though each agent can do just one. Multi-task mechanisms may not apply where individual tasks very unique, but when they can be used, they can let the mechanism learn report patterns and not require as much information, or take advantage of the fact that multiple reports on the same task should have different statistical patterns than reports on different tasks. For example, the mechanism in Dasgupta & Ghosh (2013) is multi-signal, and that is critical for getting the results just mentioned.

There are two other important distinctions: whether or not the mechanism can tolerate differences between agents, for example in how well they can do the tasks, and whether the mechanism provides any guarantees about the relative rewards of the all-truthful equilibrium and other strategies (e.g. Kamble et al. (2015)). As an example of the latter, recall the potential collusion problem with output agreement again–if all agents report the same value, they always match, guaranteeing them a higher reward than for truthful reporting. In the following chapters, I will describe *informed truthful* mechanisms, where such signal-independent reporting is less rewarding than being truthful.

When I started my research, there were no informed truthful mechanisms that worked when there were more than two possible signals. I developed one, in parallel with several other mechanisms[3] by other researchers. Unlike the others, my mechanism works even when there are only two agents and three tasks; additionally, it has lower reward variability when there are many signals, because it can reward approximate agreement between reports. On the education side, there have been a few small-scale experiments with peer prediction in education, but no one had done a careful study of a large dataset.

1.2.1 Related Research Areas

Several research areas complement studies of peer prediction. Peer prediction focuses on eliciting truthful information, and once reports are gathered, there is often a need to aggregate several reports for a task into a single value. Several machine learning and statistics techniques have been developed to improve the quality of the result by adjusting for agent biases or filtering out low

[3] Such as Radanovic & Faltings (2015a); Kamble et al. (2015); Radanovic et al. (2016); Kong & Schoenebeck (2016).

quality answers (an example in peer assessment is Piech et al. (2013)).

Another complementary use of machine learning is to automate certain tasks entirely, removing the need for crowdsourcing. In the decade or so since peer prediction was first invented, there has been tremendous progress in image and speech analysis, with many tasks that were once human-only now automatable with high quality (e.g., see the ImageNet challenge for visual object recognition (Russakovsky et al., 2015)). This trend will continue, but it is clear that many crowdsourcing problems will remain. Artificial intelligence is advancing rapidly, but there is still a huge gap to human abilities. Some applications require gathering information from the physical world, not just processing, so a fully automated solution will require autonomous robotics at competitive cost, and in some applications, such as citizen science and peer assessment, performing the tasks has an inherent value, engaging and educating the "workers."

Other research shares the goal of information elicitation, but studies settings where ground truth is available, and reports can be scored using ground truth rather than peer reports. One example is prediction markets (Hanson, 2003), where agents (implicitly) report probabilities over the outcome of a future event (e.g. the outcome of the next presidential election), and are rewarded once that event occurs.

A related area is mechanism design (Fudenberg & Tirole, 1991, Chapter 7), the study of eliciting agent preferences over a set of possible outcomes to enable a desirable decision to be made. Examples include designing auctions to maximize revenue or social welfare, designing voting rules that limit opportunities for manipulation, and building systems that match medical students to residencies. Mechanism design is similar to crowdsourcing in that is also concerned with truthfulness in agent reports; the key distinction is that the goal is eliciting preferences, not information about assigned tasks. This distinction matters because preferences are usually truly private, whereas ground truth is sometimes available in crowdsourcing. Another consequence is that agents in crowdsourcing usually do not care about how the information they report is used, whereas in mechanism design, reported preferences are used to make decisions that directly affect the agent—e.g. using an agent's bid in an auction to determine whether they win and how much they pay.

1.3 My Contributions

The goal of this dissertation is to show that peer prediction can be practical for encouraging effort in crowdsourcing, including in large-scale peer assessment for education. This section summarizes the key contributions in the three areas that will be discussed in subsequent chapters.

1.3.1 A New Minimal Mechanism for Peer Prediction

The first set of contributions define and analyse a new mechanism.

- I present the new *Correlated Agreement* mechanism for peer prediction. The mechanism is minimal, works for any non-trivial report model, and is *informed truthful*, guaranteeing that truthful reporting has expected payoff higher than that of any uninformed strategy. The base mechanism uses limited information – whether pairs of signals are positively or negatively correlated, not the full probabilistic model. The mechanism properties significantly extend those of an earlier mechanism due to Dasgupta and Ghosh (2013).

- I show that the mechanism can learn the model parameters from reports themselves, eliminating the knowledge requirements in many-task settings.

- The notion of informed truthfulness introduced here is of independent interest, allowing ties between informed strategies that depend on agents' signals, but ensuring that uninformed reporting has lower expected reward.

1.3.2 Replicator Dynamics and the Stability of Truthfulness

The next contributions are about using learning dynamics to study peer prediction.

- I motivate the use of learning dynamics in peer prediction, arguing that people do not compute equilibria, but are more likely to use strategies that have worked well for them or others before, and learn over time. If the system has a bad equilibrium, a population is more likely to get there if it is "easy to fall into."
- I confirm previous experimental results and theoretical concerns about equilibrium selection (Gao et al., 2014), showing for example that under output agreement, truthfulness is unstable.
- I show that the JF09 mechanism (Jurca & Faltings, 2009) that tried to deter constant-reporting equilibria is ineffective, and that truthfulness is more stable in recent mechanisms, including the correlated agreement mechanism.
- Using data from MOOC peer assessments, I quantitatively compare the stability of truthfulness in several peer prediction mechanisms.

1.3.3 Practical Peer Prediction in Online Peer Assessment

My final set of contributions are on the use of peer prediction in peer assessment in massive online courses.

- I analyze a large dataset of peer assessments in edX MOOCs, studying reporting patterns, agreement between peers, and other statistics that matter when consider the use of peer prediction.

- I show that the conditions on reporting patterns needed by several peer prediction mechanisms are frequently violated, so other mechanisms—e.g. the Correlated Agreement mechanism or the mechanism in Kamble et al. (2015)—that do not make such assumptions are preferable.

- I point out that fairness and reward variance are important design concerns for educational uses of peer prediction, and show that the Correlated Agreement mechanism has lower reward variation than output agreement and two recent mechanisms based on exact agreement (Kamble et al., 2015; Radanovic et al., 2016).

- I show that the benefit from exerting effort in evaluating peers is relatively low, due to the relatively low agreement between peers in my dataset, and suggest several directions for increasing agreement—using peer prediction or another effort-encouraging techniques, improving rubric design and student training, and implementing bias-compensating algorithms to map student reports to a common scale.

- Finally, I examine a particular type of mis-reporting, where students base their reports on "low-effort" signals that are easily agreed on, not on the assessment criteria they are asked to use. I find that this may be a significant issue that a practical system should expect and prepare to handle. Mitigations include improving agreement between peers as just described, allowing students to report unfair scores, and spot checking by the course team.

1.4 Bibliographic Note

Chapters 2 and 3 of this dissertation are based on the following papers, respectively:

1. Victor Shnayder, Arpit Agarwal, Rafael M. Frongillo, and David C. Parkes (2016). Strong Truthfulness in Multi-Task Peer Prediction. In *Proceedings of the 17th ACM Conference on*

Economics and Computation, EC-16.

2. Victor Shnayder, Rafael M. Frongillo, David C. Parkes (2016). Measuring Performance Of Peer Prediction Mechanisms Using Replicator Dynamics. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI-16*.

A summarized data set used for in 4 is available at `http://www.eecs.harvard.edu/~shnayder/`.

1.5 Dissertation Outline

Chapter 2 summarizes related work on peer prediction, describes the need for minimal multi-signal general-model peer prediction, and presents the *Correlated Agreement* (CA) mechanism as a solution. Chapter 3 describes my results on applying learning dynamics to peer prediction. Chapter 4 focuses on practical considerations in using peer prediction for peer assessment in education. These chapters are intended to be readable on their own, so there is some duplication of background and related work. Chapter 5 summarizes, describes open questions and areas for future work, and concludes.

2

Informed Truthfulness in Multi-Task Peer Prediction

We study the problem of information elicitation without verification ("peer prediction"). This challenging problem arises across a diverse range of multi-agent systems, in which participants are asked to respond to an information task, and where there is no external input available against which to score reports. Examples include completing surveys about the features of new products, providing feedback on the quality of food or the ambience in a restaurant, sharing emotions when watching video content, and peer assessment of assignments in Massive Open

Online Courses (MOOCs).

The challenge is to provide incentives for participants to choose to invest effort in forming an opinion (a “signal”) about a task, and to make truthful reports about their signals. In the absence of inputs other than the reports of participants, peer-prediction mechanisms make payments to one agent based on the reports of others, and seek to align incentives by leveraging correlation between reports (i.e., peers are rewarded for making reports that are, in some sense, predictive of the reports of others).

Some domains have binary signals, for example “was a restaurant noisy or not?”, and “is an image violent or not?”. We are also interested in domains with non-binary signals, for example:

- *Image labeling.* Signals could correspond to answers to questions such as “Is the animal in the picture a dog, a cat or a beaver”, or “Is the emotion expressed joyful, happy, sad or angry.” These signals are categorical, potentially with some structure: ‘joyful’ is closer to ‘happy’ than ‘sad’, for example.

- *Counting objects.* There could be many possible signals, representing answers to questions such as (“are there 0, 1-5, 6-10, 11-100, or $>$ 100 people in the picture”?). The signals are ordered.

- *Peer assessment in MOOCs*. Multiple students evaluate their peers’ submissions to an open-response question using a grading rubric. For example, an essay may be evaluated for clarity, reasoning, and relevance, with the grade for reasoning ranging from 1 (“wild flights of fancy throughout”), through 3 (“each argument is well motivated and logically defended.”)

We do not mean to take an absolute position that external “ground truth” inputs are never available in these applications. We do however believe it important to understand the extent to which such systems can operate using only participant reports.

The design of peer-prediction mechanisms assumes the ability to make payments to agents, and that an agent's utility is linear-increasing with payment and does not depend on signal reports other than through payment. Peer prediction precludes, for example, that an agent may prefer to misreport the quality of a restaurant because she is interested in driving more business to the restaurant. The payments need not be monetary; one could for example issue points to agents, these points conveying some value (e.g., redeemable for awards, or conveying status). On a MOOC platform, the payments could correspond to scores assigned as part of a student's overall grade in the class. What is needed is a linear relationship between payment (of whatever form) and utility, and expected-utility maximizers.

The challenge of peer prediction is timely. For example, Google launched *Google Local Guides* in November 2015. This provides participants with points for contributing star ratings and descriptions about locations. The current design rewards quantity but not quality and it will be interesting to see whether this attracts useful reports. After 200 contributions, participants receive a 1 TB upgrade of Drive storage (currently valued at $9.99/month.)

We are interested in *minimal* peer-prediction mechanisms, which require only signal reports from participants.[1] A basic desirable property is that truthful reporting of signals is a strict, correlated equilibrium of the game induced by the peer-prediction mechanism.[2] For many years, an Achilles heel of peer prediction has been the existence of additional equilibria that have higher

[1] While more complicated designs have been proposed (e.g. (Prelec, 2004; Witkowski & Parkes, 2012a; Radanovic & Faltings, 2015b)), in which participants are also asked to report their beliefs about the signals that others will report, we believe that peer-prediction mechanisms that require only signal reports are more likely to be adopted in practice. It is cumbersome to design user interfaces for reporting beliefs, and people are notoriously bad at reasoning about probabilities.

[2] It has been more common to refer to the equilibrium concept in peer-prediction as a Bayes-Nash equilibrium. But as pointed out by Jens Witkowski, there is no agent-specific, private information about payoffs (utility is linear in payment). In a correlated equilibrium, agents get signals and a strategy is a mapping from signals to actions. An action is a best response for a given signal if, conditioned on the signal, it maximizes an agent's expected utility. This equilibrium concept fits peer prediction: each agent receives a signal from the environment, signals are correlated, and strategies map signals into reported signals.

payoff than truthful behavior and reveal no useful information (Jurca & Faltings, 2009; Dasgupta & Ghosh, 2013; Radanovic & Faltings, 2015a). An uninformative equilibrium is one in which reports do not depend on the signals received by agents. Indeed, the equilibria of peer-prediction mechanisms must always include an uninformative, mixed Nash equilibrium (Waggoner & Chen, 2014). Moreover, with binary signals, a single task, and two agents, Jurca & Faltings (2005) show that an incentive-compatible, minimal peer-prediction mechanism will always have an uninformative equilibrium with a higher payoff than truthful reporting. Because of this, a valid concern has been that peer prediction could have the unintended effect that agents who would otherwise be truthful now adopt strategic misreporting behavior in order to maximize their payments.

In this light, a result due to Dasgupta & Ghosh (2013) is of interest: if agents are each asked to respond to multiple, independent tasks (with some overlap between assigned tasks), then in the case of binary signals there is a mechanism that addresses the problem of multiple equilibria. The binary-signal, multi-task mechanism is *strongly truthful*, meaning that truthful reporting yields a higher expected payment than any other strategy (and is tied in payoff only with strategies that report permutations of signals, which in the binary case means $1 \rightarrow 2, 2 \rightarrow 1$).

We introduce a new, slightly weaker incentive property of *informed truthfulness*: no strategy profile provides more expected payment than truthful reporting, and the truthful equilibrium is strictly better than any uninformed strategy (where agent reports are signal-independent, and avoid the effort of obtaining a signal). Informed truthfulness is responsive to what we consider to be the two main concerns of practical peer prediction design:

1. Agents should have strict incentives to exert effort toward acquiring an informative signal, and

2. Agents should have no incentive to misreport this information.

Relative to strong truthfulness, the relaxation to informed truthfulness is that there may be other informed strategies that match the expected payment of truthful reporting. Even so, informed truthfulness retains the property of strong truthfulness that there can be no other behavior strictly better than truthful reporting.

The binary-signal, multi-task mechanism of Dasgupta and Ghosh is constructed from the simple building block of a *score matrix*, with a score of '$1'$ for agreement and '$0'$ otherwise. Some tasks are designated without knowledge of participants as bonus tasks. The payment on a bonus task is 1 in the case of agreement with another agent. There is also a penalty of -1 if the agent's report on another (non-bonus) task agrees with the report of another agent on a third (non-bonus) task. In this way, the mechanism rewards agents when their reports on a shared (bonus) task agree more than would be expected based on their overall report frequencies. Dasgupta and Ghosh remark that extending beyond two signals "is one of the most immediate and challenging directions for further work."

Our main results are as follows:

- We study the *multi-signal extension of the Dasgupta-Ghosh mechanism* (DGMS), and show that DGMS is strongly truthful for domains that are *categorical*, where receiving one signal reduces an agent's belief that other agents will receive any other signal. We also show that (i) this categorical condition is tight for DGMS for agent-symmetric signal distributions, and (ii) the peer grade distributions on a large MOOC platform do not satisfy the categorical property.

- We generalize DGMS, obtaining the *Correlated Agreement (CA) mechanism*. This provides informed truthfulness in general domains, including domains in which the DGMS mechanism is neither informed- nor strongly-truthful. The CA mechanism requires the designer to know the correlation structure of signals, but not the full signal distribution. We further characterize domains where the CA mechanism is strongly truthful, and show that no mechanism with similar structure and information requirements can do better.

- For settings with a large number of tasks, we present a *detail-free CA mechanism*, in which the designer estimates the statistics of the correlation structure from agent reports. This mechanism is informed truthful in the limit where the number of tasks is large (handling the concern that reports affect estimation and thus scores), and we provide a convergence rate analysis for ε-informed truthfulness with high probability.

We believe that these are the first results on strong or informed truthfulness in domains with non-binary signals without requiring a large population for their incentive properties (compare with (Radanovic & Faltings, 2015a; Kamble et al., 2015; Radanovic et al., 2016)). The robust incentives of the multi-task DGMS and CA mechanisms hold for as few as two agents and three tasks, whereas these previous papers crucially rely on being able to learn statistics of the distribution from multiple reports. Even if given the true underlying signal distribution, the mechanisms in these earlier papers would still need to use a large population, with the payment rule based on statistics estimated from reports, as this is critical for incentive alignment in these papers. Our analysis framework also provides a dramatic simplification of the techniques used by Dasgupta & Ghosh (2013).

In a recent working paper, Kong & Schoenebeck (2016) show that a number of peer prediction mechanisms that provide variations on strong-truthfulness can be derived within a single information-theoretic framework, with scores determined based on the information they provide relative to reports in the population (leveraging a measure of mutual information between the joint distribution on signal reports and the product of marginal distributions on signal reports). Earlier mechanisms correspond to particular information measures. Their results use different technical tools, and also include a different, multi-signal generalization of Dasgupta & Ghosh (2013) that is independent of our results, outside of the family of mechanisms that we consider in

Section 2.4.3, and provides strong truthfulness in the limit of a large number of tasks.[3]

2.0.1 Related Work

The theory of peer prediction has developed rapidly in recent years. We focus on minimal peer-prediction mechanisms. Beginning with the seminal work of Miller et al. (2005), a sequence of results relax knowledge requirements on the part of the designer (Witkowski & Parkes, 2012a; Jurca & Faltings, 2011), or generalize, e.g. to handle continuous signal domains (Radanovic & Faltings, 2014). Simple output-agreement, where a positive payment is received if and only if two agents make the same report (as used in the *ESP image labeling game* (von Ahn & Dabbish, 2004)), has also received some theoretical attention (Waggoner & Chen, 2014; Jain & Parkes, 2013).

Early peer prediction mechanisms had uninformative equilibria that gave better payoff than honesty. Jurca & Faltings (2009) show how to remove uninformative, pure-strategy Nash equilibria through a clever three-peer design. Kong et al. (2016) show how to design strong truthful, minimal, single-task mechanisms with a known model when there are reports from a large number of agents.

In addition to Dasgupta & Ghosh (2013) and Kong & Schoenebeck (2016), several recent papers have tackled the problem of uninformative equilibria. Radanovic & Faltings (2015a) establish strong truthfulness amongst symmetric strategies in a large-market limit where both the number of tasks and the number of agents assigned to each task grow without bound. Radanovic et al. (2016) provide complementary theoretical results, giving a mechanism in which truthfulness is the equilibrium with highest payoff, based on a population that is large enough to estimate statistical properties of the report distribution. They require a self-predicting condition that limits

[3] While they do not state or show that the mechanism does not need a large number of tasks in any special case, the techniques employed can also be used to design a mechanism that is a linear transform of our CA mechanism, and thus informed truthful with a known signal correlation structure and a finite number of tasks (personal communication).

the correlation between differing signals. Each agent need only be assigned a single task. Kamble et al. (2015) describe a mechanism where truthfulness has higher payoff than uninformed strategies, providing an asymptotic analysis as the number of tasks grows without bound. The use of learning is crucial in these papers. In particular, they must use statistics estimated from reports to design the payment rule in order to align incentives. This is a key distinction from our work. Cai et al. (2015) work in a different model, showing how to achieve optimal statistical estimation from data provided by self-interested participants. These authors do not consider misreports and their mechanism is not informed- (or strongly-) truthful and is vulnerable to collusion. Their model is interesting, though, in that it adopts a richer, non-binary effort model. Witkowski & Parkes (2013) first introduced the combination of learning and peer prediction, coupling the estimation of the signal prior together with the shadowing mechanism.

Although there is disagreement in the experimental literature about whether equilibrium selection is a problem in practice, there is compelling evidence that it matters (Gao et al., 2014); see Faltings et al. (2014) for a study where uninformed equilibria did not appear to be a problem. One difference is that this later study was in a many-signal domain, making it harder for agents to coordinate on an uninformative strategy. Chapter 3 in this dissertation uses replicator dynamics as a model of agent learning to argue that equilibrium selection is indeed important, and that truthfulness is significantly more stable under mechanisms that ensure it has higher payoff than other strategies. Orthogonal to concerns about equilibrium selection, Gao et al. (2016) point out a modeling limitation—when agents can coordinate on some other, unintended source of signal, then this strategy may be better than truthful reporting. They suggest randomly checking a fraction of reports against ground truth as an alternative way to encourage effort. We discuss this in Section 2.4.6.

Turning to online peer assessment for MOOCs, research has primarily focused on evaluating students' skill at assessment and compensating for grader bias (Piech et al., 2013), as well as helping students self-adjust for bias and provide better feedback (Kulkarni et al., 2013). Other studies, such as the *Mechanical TA* (Wright & Leyton-Brown, 2015), focus on reducing TA workload in high-stakes peer grading. A recent paper (Wu et al., 2015) outlines an approach to peer assessment that relies on students flagging overly harsh feedback for instructor review. We are not aware of any systematic studies of peer prediction in the context of MOOCs, though Radanovic et al. (2016) present experimental results from an on-campus experiment.

2.1 Model

We consider two agents, 1 and 2, which are perhaps members of a larger population. Let $k \in M = \{1, \ldots, m\}$ index a task from a universe of $m \geq 3$ tasks to which one or both of these agents are assigned, with both agents assigned to at least one task. Each agent receives a signal when investing effort on an assigned task. The effort model that we adopt is binary: either an agent invests no effort and does not receive an informed signal, or an agent invests effort and incurs a cost and receives a signal.

Let S_1, S_2 denote random variables for the signals to agents 1 and 2 on some task. The signals have a finite domain, with $i, j \in \{1, \ldots, n\}$ indexing a realized signal to agents 1 and 2, respectively.

Each task is *ex ante* identical, meaning that pairs of signals are i.i.d. for each task. Let $P(S_1{=}i, S_2{=}j)$ denote the joint probability distribution on signals, with marginal probabilities $P(S_1{=}i)$ and $P(S_2{=}j)$ on the signals of agents 1 and 2, respectively. We assume exchangeability, so that the identity of agents does not matter in defining the signal distribution. The signal distribution is common knowledge to agents.[4]

We assume that the signal distribution satisfies *stochastic relevance*, so that for all $s' \neq s''$, there

[4] We assume common knowledge and symmetric signal models for simplicity of exposition. Our mechanisms do not require full information about the signal distribution, only the correlation structure of signals, and can tolerate some user heterogeneity, as described further in Section 2.4.5.

exists at least one signal s such that

$$P(S_1{=}s|S_2{=}s') \neq P(S_1{=}s|S_2{=}s''), \tag{2.1}$$

and symmetrically, for agent 1's signal affecting the posterior on agent 2's. If two signals are not stochastically relevant, they can be combined into one signal.

Our constructions and analysis will make heavy use of the following matrix, which encodes the correlation structure of signals.

Definition 1 (Delta matrix). *The* Delta matrix Δ *is an* $n \times n$ *matrix, with entry* (i, j) *defined as*

$$\Delta_{ij} = P(S_1{=}i, S_2{=}j) - P(S_1{=}i)P(S_2{=}j). \tag{2.2}$$

The Delta matrix describes the correlation (positive or negative) between different realized signal values. For example, if

$$\begin{aligned}\Delta_{1,2} &= P(S_1{=}1, S_2{=}2) - P(S_1{=}1)P(S_2{=}2) &(2.3)\\ &= P(S_1{=}1)(P(S_2{=}2|S_1{=}1) - P(S_2{=}2)) > 0, &(2.4)\end{aligned}$$

then $P(S_2{=}2|S_1{=}1) > P(S_2{=}2)$, so signal 2 is positively correlated with signal 1 (and by exchangeability, similarly for the effect of 1 on 2). If a particular signal value increases the probability that the other agent will receive the same signal then $P(S_1{=}i, S_2{=}i) > P(S_1{=}i)P(S_2{=}i)$, and if this holds for all signals the Delta matrix has a positive diagonal. Because the entries in a row i of joint distribution $P(S_1{=}i, S_2{=}j)$ and a row of product distribution $P(S_1{=}i)P(S_2{=}j)$ both sum to $P(S_1{=}i)$, each row in the Δ matrix sums to 0 as the difference of the two. The same holds for columns.

The CA mechanism will depend on the sign structure of the Δ matrix, without knowledge of the specific values. We will use a sign operator $\mathrm{Sign}(x)$, with value 1 if $x > 0$, 0 otherwise.[5]

Example 4. *If the signal distribution is*

$$P(S_1, S_2) = \begin{bmatrix} 0.4 & 0.15 \\ 0.15 & .3 \end{bmatrix}$$

with marginal distribution $P(S) = [0.55; 0.45]$, *we have*

$$\Delta = \begin{bmatrix} 0.4 & 0.15 \\ 0.15 & .3 \end{bmatrix} - \begin{bmatrix} 0.55 \\ 0.45 \end{bmatrix} \cdot \begin{bmatrix} 0.55 & 0.45 \end{bmatrix} \approx \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 0.1 \end{bmatrix},$$

and

$$\mathrm{Sign}(\Delta) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

An agent's *strategy* defines, for every signal it may receive and each task it is assigned, the signal it will report. We allow for mixed strategies, so that an agent's strategy defines a distribution over signals. Let R_1 and R_2 denote random variables for the *reports* by agents 1 and 2, respectively, on some task. Let matrices F and G denote the mixed strategies of agents 1 and 2, respectively, with $F_{ir} = P(R_1{=}r|S_1{=}i)$ and $G_{jr} = P(R_2{=}r|S_2{=}j)$ to denote the probability of making report r given signal i is observed (signal j for agent 2). Let $r_1^k \in \{1, \ldots, n\}$ and $r_2^k \in \{1, \ldots, n\}$ refer to the realized report by agent 1 and 2, respectively, on task k (if assigned).

Definition 2 (Permutation strategy). *A* permutation strategy *is a deterministic strategy in which an agent adopts a bijection between signals and reports, that is, F (or G for agent 2) is a permutation matrix.*

Definition 3 (Informed and uninformed strategies). *An* informed strategy *has $F_{ir} \neq F_{jr}$ for some $i \neq j$, some $r \in \{1, \ldots, n\}$ (and similarly for G for agent 2). An* uninformed strategy *has the same*

[5] Note that this differs from the standard sign operator, which has value -1 for negative inputs.

report distribution for all signals.

Permutation strategies are merely relabelings of the signals; in particular, truthfulness (denoted $\mathbb{I}$ below) is a permutation strategy. Note also that by definition, deterministic uniformed strategies are those that give the same report for all signals.

Each agent is assigned to two or more tasks, and the agents overlap on at least one task. Let $M_b \subseteq M$ denote a non-empty set of "bonus tasks", a subset of the tasks to which both agents are assigned. Let $M_1 \subseteq M \setminus M_b$ and $M_2 \subseteq M \setminus M_b$, with $M_1 \cap M_2 = \emptyset$ denote non-empty sets of tasks to which agents 1 and 2 are assigned, respectively. These will form the "penalty tasks." For example, if both agents are assigned to each of three tasks, A, B and C, then we could choose $M_b = \{A\}$, $M_1 = \{B\}$ and $M_2 = \{C\}$.

We assume that tasks are *a priori* identical, so that there is nothing to distinguish two tasks other than their signals. In particular, agents have no information about which tasks are shared, or which are designated bonus or penalty. This can be achieved by choosing M_b, M_1 and M_2 randomly after task assignment. This can also be motivated in largely anonymous settings, such as peer assessment and crowdsourcing.

A *multi-task peer-prediction mechanism* defines a total payment to each agent based on the reports made across all tasks. The mechanisms that we study assign a total payment to an agent based on the sum of payments for each bonus task, but where the payment for a bonus task is adjusted downwards by the consideration of its report on a penalty task and that of another agent on a different penalty task.

For the mechanisms we consider in this paper, it is without loss of generality for each agent to adopt a uniform strategy across each assigned task. Changing a strategy from task to task is equivalent in terms of expected payment to adopting a linear combination over these strategies, given that tasks are presented in a random order, and given that tasks are equivalent, conditioned on signal.

This result relies on the random order of tasks as presented to each agent, preventing coordination. Tasks will be indexed as $1, \ldots, k \ldots, m$ from the first agent's point of view. The second agent will see them reshuffled using a permutation π chosen uniformly at random: $\pi(1), \ldots, \pi(m)$.

Let $\vec{F}$ be the first agent's strategy vector, with F_k the first agent's strategy on task k. Fix the second agent's vector of strategies $\vec{G}$. Let J_{ij} be the joint signal distribution. Then, for a broad class of mechanisms, it is without loss of generality to focus on agents having a single per-task strategy applied to all tasks.

Let K, K', K'' be random variables corresponding to a task id, with uniform probability of value $1, \ldots, m$. Let $\mathcal{M}$ be a *linear* mechanism if its expected score function is a linear function of $\Pr(R_1^K = r_1, R_2^K = r_2)$ and $\Pr(R_1^{K'} = r_1, R_2^{K''} = r_2 | K' \neq K'')$, for all set of report pairs r_1, r_2. For example, the DGMS mechanism we describe later has expected score

$$\Pr(R_1^K = R_2^K) - \Pr(R_1^{K'} = R_2^{K''} | K' \neq K'') = \tag{2.5}$$

$$= \sum_{r=1}^{n} \Pr(R_1^K = r, R_2^K = r) - \Pr(R_1^{K'} = r, R_2^{K''} = r | K' \neq K''), \tag{2.6}$$

which fits this condition. The multi-task mechanism we define below is also linear. The expectation is with respect to the signal model, agent strategies, the random task order, and any randomization in the scoring mechanism itself.

Lemma 1. *Let $\mathcal{M}$ be a linear mechanism. Let $\vec{F}$ be a vector of strategies. Then for any $\vec{G}$, $\bar{F} = mean(\vec{F})$ will have the same expected score as $\vec{F}$.*

Proof. We prove equivalence of expected value of $\Pr(R_1^K = r_1, R_2^K = r_2)$ and $\Pr(R_1^{K'} = r_1, R_2^{K''} = r_2 | K' \neq K'')$ for all r_1, r_2, and equivalence for any $\mathcal{M}$ follows by linearity.

Fix r_1, r_2. Then $\Pr(R_1^K = r_1, R_2^K = r_2)$ has the same expected value for $\vec{F}$ and $\bar{F}$:

$$\Pr(R_1^K = r_1, R_2^K = r_2) = \tag{2.7}$$

$$= \frac{1}{m}\sum_{k=1}^{m} \Pr(R_1^k = r_1, R_2^k = r_2) \tag{2.8}$$

$$= \frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} \Pr(S_1^k = i, S_2^k = j)\Pr(R_1^k = r_1|s_1 = i)\Pr(R_2^k = r_2|s_2 = j) \tag{2.9}$$

$$= \frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} J_{ij}F_{ir_1}^{k}G_{jr_2}^{\pi(k)}, \tag{2.10}$$

Taking the expectation over π, we get

$$= \frac{1}{m!}\sum_{\pi}\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} J_{ij}F_{ir_1}^{k}G_{jr_2}^{\pi(k)} \tag{2.11}$$

where the sum is over all $m!$ possible permutations of the tasks. By symmetry, we know that each element of G will be used for task k with equal probability $1/m$:

$$= \frac{1}{m}\sum_{\ell}\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} J_{ij}F_{ir_1}^{k}G_{jr_2}^{\ell} \tag{2.12}$$

and reordering the sums, we get:

$$= \frac{1}{m}\sum_{\ell}\sum_{i=1}^{n}\sum_{j=1}^{n} J_{ij}G_{jr_2}^{\ell}\frac{1}{m}\sum_{k=1}^{m} F_{ir_1}^{k}. \tag{2.13}$$

Using the definition of $\bar{F}$ as the mean of $\vec{F}$, (2.14)

$$= \frac{1}{m}\sum_{\ell}\sum_{i=1}^{n}\sum_{j=1}^{n} J_{ij}G_{jr_2}^{\ell}\bar{F}_{ir_1} \tag{2.15}$$

$$= \Pr(R_1^K = r_1, R_2^K = r_2|\text{using } \bar{F} \text{ instead of } \vec{F}) \tag{2.16}$$

The same argument works for $\Pr(R_1^{K'} = r_1, R_2^{K''} = r_2|K' \neq K'')$, substituting $\Pr(S_1 =$

$i)\Pr(S_2 = j)$ for J_{ij}. The key to the proof is the random permutation of task order in line 2.12, which prevents coordination between the per-task strategies of the two agents. □

Given this uniformity, we write $E(F, G)$ to denote the expected payment to an agent for any bonus task. The expectation is taken with respect to both the signal distribution and any randomization in agent strategies. Let $\mathbb{I}$ denote the truthful reporting strategy, which corresponds to the identity matrix.

Definition 4 (Strictly Proper). *A multi-task peer-prediction mechanism is* proper *if and only if truthful strategies form a correlated equilibrium, so that* $E(\mathbb{I}, \mathbb{I}) \geq E(F, \mathbb{I})$, *for all strategies* $F \neq \mathbb{I}$, *and similarly when reversing the roles of agents* 1 *and* 2. *For* strict properness, *the inequality must be strict.*

This insists that the expected payment on a bonus task is (strictly) higher when reporting truthfully than when using any other strategy, given that the other agent is truthful.

Definition 5 (Strongly-truthful). *A multi-task peer-prediction mechanism is* strongly-truthful *if and only if for all strategies* F, G *we have* $E(\mathbb{I}, \mathbb{I}) \geq E(F, G)$, *and equality may only occur when* F *and* G *are both the same permutation strategy.*

In words, strong-truthfulness requires that both agents being truthful has strictly greater expected payment than any other strategy profile, unless both agents play the same permutation strategy, in which case equality is allowed.[6] From the definition, it follows that any strongly-truthful mechanism is strictly proper.

Definition 6 (Informed-truthful). *A multi-task peer-prediction mechanism is* informed-truthful *if and only if for all strategies* F, G, $E(\mathbb{I}, \mathbb{I}) \geq E(F, G)$, *and equality may only occur when both* F *and* G *are informed strategies.*

In words, informed-truthfulness requires that the truthful strategy profile has strictly higher

[6] Permutation strategies are not of practical concern—they require coordination and provide no benefit over being truthful.

expected payment than any profile in which one or both agents play an uninformed strategy, and weakly greater expected payment than all other strategy profiles. It follows that any informed-truthful mechanism is proper.

Although weaker than strong-truthfulness, informed truthfulness is responsive to the primary, practical concern in peer-prediction applications: avoiding equilibria where agents achieve the same (or greater) payment as a truthful informed agent but without putting in the effort of forming a careful opinion about the task. For example, it would be undesirable for agents to be able to do just as well or better by reporting the same signal all the time. Once agents exert effort and observe a signal, it is reasonable to expect them to make truthful reports as long as this is an equilibrium and there is no other equilibrium with higher expected payment. Informed-truthful peer-prediction mechanisms provide this guarantee.[7]

2.2 Multi-Task Peer-Prediction Mechanisms

We define a class of multi-task peer-prediction mechanisms that is parametrized by a *score matrix*, $S : \{1, \ldots, n\} \times \{1, \ldots, n\} \to \mathbb{R}$, that maps a pair of reports into a score, the same score for both agents. This class of mechanisms extends the binary-signal multi-task mechanism due to Dasgupta & Ghosh (2013) in a natural way.

Definition 7 (**Multi-task mechanisms**). *These mechanisms are parametrized by score matrix S.*

1. *Assign each agent to two or more tasks, with at least one task in common, and at least three tasks total.*

2. *Let r_1^k denote the report received from agent* 1 *on task k (and similarly for agent* 2*). Designate one*

[7] For simplicity of presentation, we do not model the cost of effort explicitly, but it is a straightforward extension to handle the cost of effort as suggested in previous work (Dasgupta & Ghosh, 2013). In our proposed mechanisms, an agent that does not exert effort receives an expected payment of zero, while the expected payment for agents that exert effort and play the truthful equilibrium is strictly positive. With knowledge of the maximum possible cost of effort, scaling the payments appropriately incentivizes effort.

or more tasks assigned to both agents as bonus tasks (set M_b). Partition the remaining tasks into penalty tasks M_1 and M_2, where $|M_1| > 0$ and $|M_2| > 0$ and M_1 tasks have a report from agent 1 and M_2 a report from agent 2.

3. *For each bonus task $k \in M_b$, pick a random $\ell \in M_1$ and $\ell' \in M_2$. The payment to both agent 1 and agent 2 for task k is $S(r_1^k, r_2^k) - S(r_1^\ell, r_2^{\ell'})$.*

4. *The total payment to an agent is the sum total payment across all bonus tasks.*[8]

As discussed above, it is important that agents do not know which tasks will become bonus tasks and which become penalty tasks. The expected payment on a bonus task for strategies F, G is

$$\begin{aligned} E(F,G) &= \sum_{i=1}^{n}\sum_{j=1}^{n} P(S_1{=}i, S_2{=}j) \sum_{r_1=1}^{n}\sum_{r_2=1}^{n} P(R_1{=}r_1|S_1{=}i)P(R_2{=}r_2|S_2{=}j)S(r_1,r_2) \\ &\quad - \sum_{i=1}^{n}\sum_{j=1}^{n} P(S_1{=}i)P(S_2{=}j) \sum_{r_1=1}^{n}\sum_{r_2=1}^{n} P(R_1{=}r_1|S_1{=}i)P(R_2{=}r_2|S_2{=}j)S(r_1,r_2) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij} \sum_{r_1=1}^{n}\sum_{r_2=1}^{n} S(r_1,r_2)F_{ir_1}G_{jr_2}. \end{aligned} \tag{2.17}$$

The expected payment can also be written succinctly as $E(F,G) = \mathrm{tr}(F^\top \Delta G S^\top)$. In words, the expected payment on a bonus task is the sum, over all pairs of possible signals, of the product of the correlation (negative or positive) for the signal pair and the (expected) score given the signal pair and agent strategies.

For intuition, note that for the identity score matrix which pays \$1 in the case of matching reports and \$0 otherwise, agents are incentivized to give matching reports for signal pairs with

[8] A variation with the same expected payoff and the same incentive analysis is to compute the expectation of the scores on all pairs of penalty tasks, rather than sampling. We adopt the simpler design for ease of exposition. This alternate design would reduce score variance if there are many non-bonus tasks, and may be preferable in practice. See Chapter 4.

positive correlation and non-matching reports for signals with negative correlation. Now consider a general score matrix S, and suppose that all agents always report 1. They always get $S(1,1)$ and the expected value $E(F,G)$ is a multiple of the sum of entries in the Δ matrix, which is exactly zero. Because individual rows and columns of Δ also sum to zero, this also holds whenever a single agent uses an uninformed strategy. In comparison, truthful behavior provides payment $E(\mathbb{I},\mathbb{I}) = \sum_{ij} \Delta_{ij} S(i,j)$, and will be positive if the score matrix is bigger where signals are positively correlated than where they are not.

While agent strategies in our model can be randomized, the linearity of the expected payments allows us to restrict our attention to deterministic strategies.

Lemma 2. *For any world model and any score matrix S, there exists a deterministic, optimal joint strategy for a multi-task mechanism.*

Proof. The proof relies on solutions to convex optimization problems being extremal. The game value can be written $V = \max_F \max_G h(F,G)$, where

$$h(F,G) = \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij} \sum_{r_1=1}^{n} \sum_{r_2=1}^{n} S(r_1,r_2) F_{ir_1} G_{jr_2} \ .$$

Note that h is linear in both F and G separately. Now letting $V(F) = \max_G h(F,G)$ be the value for the G player for a fixed F, we have $V = \max_F V(F)$ by definition. As $h(F,\cdot)$ is linear, and the strategy space for G, all binary row-stochastic matrices, is convex, there exists a maximizer at an extreme point. These extreme points are exactly the deterministic strategies, and thus for all F there exists an optimal $G = G^{\text{opt}}$ which is deterministic. Considering the maximization over F, we see that $V(F) = \max_G h(F,G)$ is a pointwise supremum over a set of linear functions, and is thus convex. V is therefore optimized by an extreme point, some deterministic $F = F^{\text{opt}}$, and for that F^{opt} there exists a corresponding deterministic G^{opt} by the above. □

Lemma 2 has several consequences:

- It is without loss of generality to focus on deterministic strategies when establishing strongly truthful or informed truthful properties of a mechanism.
- There is a deterministic, perhaps asymmetric equilibrium, because the optimal solution that maximizes $E(F, G)$ is also an equilibrium.
- It is without loss of generality to consider deterministic deviations when checking whether or not truthful play is an equilibrium.

We will henceforth assume deterministic strategies. By a slight abuse of notation, let $F_i \in \{1, \ldots, n\}$ and $G_j \in \{1, \ldots, n\}$ denote the reported signals by agent 1 for signal i and agent 2 for signal j, respectively. The expected score then simplifies to

$$E(F, G) = \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij} S(F_i, G_j). \tag{2.18}$$

We can think of deterministic strategies as mapping signal pairs to reported signal pairs. Strategy profile (F, G) picks out a report pair (and thus score) for each signal pair i, j with its corresponding Δ_{ij}. That is, strategies F and G map signals to reports, and the score matrix S maps reports to scores, so together they map signals to scores, and we then dot those scores with Δ.

2.3 The Dasgupta-Ghosh Mechanism

We first study the natural extension of the Dasgupta & Ghosh (2013) mechanism from binary to multi-signals. This multi-task mechanism uses as the score matrix S the identity matrix ('$1'$ for agreement, '$0'$ for disagreement.)

Definition 8 (The Multi-Signal Dasgupta-Ghosh mechanism (DGMS)). *This is a multi-task mechanism with score matrix $S(i, j) = 1$ if $i = j$, 0 otherwise.*

Example 5. *Suppose agent* 1 *is assigned to tasks* $\{A, B\}$ *and agent* 2 *to tasks* $\{B, C, D\}$*, so that* $M_b = \{B\}, M_1 = \{A\}$ *and* $M_2 = \{C, D\}$*. Now, if the reports on* B *are both* 1*, and the reports on* A, C*, and* D *were* $0, 0$*, and* 1*, respectively, the expected payment to each agent for bonus task* B *is* $1 - (1 \cdot 0.5 + 0 \cdot 0.5) = 0.5$*. In contrast, if both agents use an uninformed coordinating strategy and always report* 1*, the expected score for both is* $1 - (1 \cdot 0.5 + 1 \cdot 0.5) = 0$*.*

The expected payment in the DGMS mechanism on a bonus task is

$$E(F, G) = \sum_{i,j} \Delta_{ij} 1_{[F_i = G_j]}, \tag{2.19}$$

where $1_{x=y}$ is 1 if $x = y$, 0 otherwise. An equivalent expression is $\operatorname{tr}(F^\top \Delta G)$.

Definition 9 (Categorical model). *A world model is* categorical *if, when an agent sees a signal, all other signals become less likely than their prior probability; i.e.,* $P(S_2 = j | S_1 = i) < P(S_2 = j)$*, for all i, for all* $j \neq i$ *(and analogously for agent* 2*). This implies positive correlation for identical signals:* $P(S_2 = i | S_1 = i) > P(S_2 = i)$*.*

Two equivalent definitions of categorical are that the Delta matrix has positive diagonal and negative off-diagonal elements, or that $\operatorname{Sign}(\Delta) = \mathbb{I}$.

Theorem 1. *If the world is categorical, then the DGMS mechanism is strongly truthful and strictly proper. Conversely, if the Delta matrix* Δ *is symmetric and the world is not categorical, then the DGMS mechanism is not strongly truthful.*

Proof. First, we show that truthfulness maximizes expected payment. We have $E(F, G) = \sum_{i,j} \Delta_{ij} 1_{[F_i = G_j]}$. The truthful strategy corresponds to the identity matrix $\mathbb{I}$, and results in a payment equal to the trace of Δ: $E(\mathbb{I}, \mathbb{I}) = \operatorname{tr}(\Delta) = \sum_i \Delta_{ii}$. By the categorical assumption, Δ has positive diagonal and negative off-diagonal elements, so this is the sum of all the positive elements of Δ. Because $1_{[F_i = G_j]} \leq 1$, this is the maximum possible payment for any pair of strategies.

To show strong truthfulness, first consider an asymmetric joint strategy, with $F \neq G$. Then there exists i s.t. $F_i \neq G_i$, reducing the expected payment by at least $\Delta_{ii} > 0$. Now consider symmetric, non-permutation strategies $F = G$. Then there exist $i \neq j$ with $F_i = F_j$. The expected payment will then include $\Delta_{ij} < 0$. This shows that truthfulness and symmetric permutation strategies are the only optimal strategy profiles. Strict properness follows from strong truthfulness.

For the tightness of the categorical assumption, first consider a symmetric Δ with positive off-diagonal elements Δ_{ij} and Δ_{ji}. Then agents can benefit by both "merging" signals i and j. Let $\bar{F}$ be the strategy that is truthful on all signals other than j, and reports i when the signal is j. Then $E(\bar{F}, \bar{F}) = \Delta_{ij} + \Delta_{ji} + \mathrm{tr}(\Delta) > E(\mathbb{I}, \mathbb{I}) = \mathrm{tr}(\Delta)$, so DGMS is not strongly truthful. Now consider a Δ where one of the on-diagonal entries is negative, say $\Delta_{ii} < 0$. Then, because all rows and columns of Δ must add to 0, there must be a j such that $\Delta_{ij} > 0$, and this reduces to the previous case where "merging" i and j is useful. □

For binary signals ('1' and '2'), any positively correlated model, such that $\Delta_{1,1} > 0$ and $\Delta_{2,2} > 0$, is categorical, and thus we obtain a substantially simpler proof of the main result in Dasgupta & Ghosh (2013).

2.3.1 Discussion: Applicability of the DGMS Mechanism

Which world models are categorical? One example is a noisy observation model, where each agent observes the "true" signal t with probability q greater than $1/n$, and otherwise makes a mistake uniformly at random, receiving any signal $s \neq t$ with probability $(1 - q)/(n - 1)$. Such model makes sense for classification tasks in which the classes are fairly distinct. For example, we would expect a categorical model for a question such as "Does the animal in this photo swim, fly, or walk?"

On the other hand, a classification problem such as the ImageNet challenge (Russakovsky et al., 2015), with 1000 nuanced and often similar image labels, is unlikely to be categorical. For example, if "Ape" and "Monkey" are possible labels, one agent seeing "Ape" is likely to increase the probability that another says "Monkey", when compared to the prior for "Monkey" in a generic set of photos. The categorical property is also unlikely to hold when signals have a natural order, which we dub *ordinal* worlds.

Example 6. *If two evaluators grade essays on a scale from one to five, when one decides that an essay should get a particular grade, e.g. one, this may increase the likelihood that their peer decides on that or an adjacent grade, e.g. one or two. In this case, the sign of the delta matrix would be*

$$\text{Sign}(\Delta) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{2.20}$$

Under the DGMS mechanism, evaluators increase their expected payoff by agreeing to always report one whenever they thought the score was either one or two, and doing a similar "merge" for other pairs of reports. We will return to this example below.

The categorical condition is a stronger requirement than previously proposed properties in the literature, such as those assumed in the analyses of the Jurca & Faltings (2011) and Radanovic et al. (2016) "1/prior" mechanism and the Witkowski & Parkes (2012a) shadowing mechanism. The 1/prior mechanism requires the self-predicting property

$$\Pr(S_2 = j | S_1 = i) < \Pr(S_2 = j | S_1 = j),$$

whereas the categorical property insists on a upper bound of $\Pr(S_2 = j)$, which is tighter than $\Pr(S_2 = j|S_1 = j)$ in the typical case where the model has positive correlation.The shadowing mechanism requires

$$\Pr(S_2 = i|S_1 = j) - \Pr(S_2 = i) < \Pr(S_2 = j|S_1 = j) - \Pr(S_2 = j),$$

which says that the likelihood of signal $S_2 = i$ cannot go up "too much" given signal $S_1 = j$, whereas the categorical property requires the stronger condition that $\Pr(S_2 = i|S_1 = j) - \Pr(S_2 = i) < 0$.

To see how often categorical condition holds in practice, we look at the correlation structure in a dataset from a large MOOC provider, focusing on 104 questions with over 100 submissions each, for a total of 325,523 assessments from 17 courses. Each assessment consists of a numerical score, which we examine, and an optional comment, which we do not study here. As an example, one assessment task for a writing assignment asks how well the student presented their ideas, with options "Not much of a style at all", "Communicative style", and "Strong, flowing writing style", and a paragraph of detailed explanation for each. These correspond to 0, 1, and 2 points on this rubric element. While we only see student reports, we take as an assumption that these reasonably approximate the true world model. As MOOCs develop along with valuable credentials based on their peer-assessed work, we believe it will nevertheless become increasingly important to provide explicit credit mechanisms for peer assessment.

We estimate Δ matrices on each of the 104 questions from the assessments. We can think about each question as corresponding to a different signal distribution, and assessing a particular student's response to the question as an information task that is performed by several peers. The questions in our data set had five or fewer rubric options (signals), with three being most common (Figure 2.1).

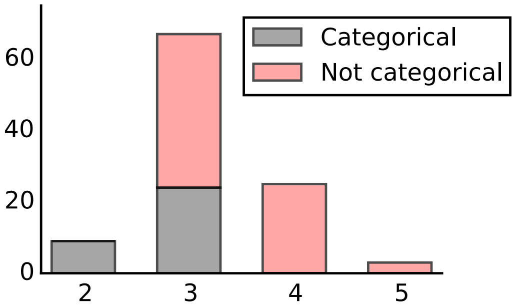


Figure 2.1: MOOC peer evaluation is an ordinal scoring setting, so most models with 3 or more signals are not categorical.

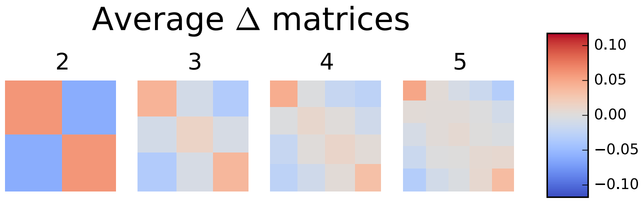


Figure 2.2: Averaged Δ matrices, grouped by the number of signals in a domain. The positive diagonals show that users tend to agree on their assessments. For models of size 4 and 5, the ordinal nature of peer assessment is clear (e.g., an assessment of $2/5$ is positively correlated with an assessment of $3/5$).

This analysis confirms that the categorical condition only holds for about one third of our three-signal models and for none of the larger models (Figure 2.1). We also computed the average Δ matrix for each model size, as visualized in Figure 2.2. The bands of positive correlation around the diagonal are typical of what we refer to as an ordinal rather than categorical domain.

2.4 Handling the General Case

We showed that the DGMS mechanism is only strongly truthful in categorical domains. The natural follow-on question is what can be done in non-categorical domains. In this section, we present a mechanism that is informed-truthful for general domains.

2.4.1 An Illustrative Reduction

We first discuss an attempt to reduce the general multi-signal setting to binary, and then use the binary Dasgupta-Ghosh mechanism. This is inspired by the multi-signal shadowing method (Witkowski, 2014, Section 3.6.2), and while it turns out not to work, it provides useful intuition. The idea is as follows: split the set of signals into two non-empty subsets A and B, and treat all signals in each subset as identical. Because binary settings are naturally categorical, this seems on the surface to give a simple general mechanism. Let us look at an example:

There are three signals: $0, 1, 2$. We will work with a multiple of the delta matrix to have whole numbers.

$$\Delta \approx \begin{bmatrix} 2 & 1 & -3 \\ 1 & 2 & -3 \\ -3 & -3 & 6 \end{bmatrix} \tag{2.21}$$

This model is not categorical, so DGMS is not truthful—reporting truthfully has expected score 10 (the sum of the diagonal entries of Δ), whereas both agents reporting 1 when their signals are either 0 or 1 would get expected score 12 (the sum of all positive entries of Δ). Our proposed mechanism tries to do such combining or "merging" of signals for the agents. For example, if the system makes $A = \{0, 1\}$, and $B = \{2\}$, we get a "reduced" delta:

$$\Delta_{\{0,1\},\{2\}} = \begin{bmatrix} 6 & -6 \\ -6 & 6 \end{bmatrix} \tag{2.22}$$

The top left element, corresponding to signal pairs in A, A, is the sum of the top-left four elements of the original Δ, including the off-diagonal elements. For this signal partition, the expected score for truthful reporting is 12, matching the best manipulation of DGMS.

Other partitions do not do as well: if the system chooses $A = \{0\}$, and $B = \{1, 2\}$, we get reduced delta:

$$\Delta_{\{0\},\{1,2\}} = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix} \tag{2.23}$$

This has a lower truthful score 4, and now agents can manipulate by misreporting their signals: $0, 1 \to 0; 2 \to 2$ will again give expected score 12.

One might think that by randomizing over all splits, we can prevent agents from coordinating to improve their score. As foreshadowed, that too fails: when two signals $s \neq s'$ are positively correlated on shared tasks (i.e. the corresponding off-diagonal entry of delta is positive), agents can benefit by ensuring that pair of signals is rewarded. Being truthful with a mechanism that randomizes the signal partitions will do this sometimes, but agents can guarantee it by merging s and s' into one or the other. We can make this concrete by continuing our example. The mechanism now picks a random split of signals into A, B, both non-empty. There are three options for A: $\{0\}$, $\{0, 1\}$, and $\{0, 2\}$ (plus the three symmetric options with A and B swapped). We looked at the resulting reduced delta matrices for the first two already. The last is:

$$\Delta_{\{0,2\},\{1\}} = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix} \tag{2.24}$$

If agents are truthful, their expected score is the average of 12, 4, and 4: $20/3$. Merging 0 and 1 as above: $0, 1 \rightarrow 0; 2 \rightarrow 2$ will have scores 12, 12, and 0 for the three splits, for an average of $24/3 > 20/3$. By merging, agents change the distribution over reports to ensure that positively correlated pairs of signals are always rewarded. This inspires our next mechanism, that directly rewards such correlated reports.

2.4.2 The Correlated Agreement Mechanism

Based on the intuition given in Section 2.2, and the success of DGMS for categorical domains, it seems promising to base the construction of a mechanism on the correlation structure of the signals, and in particular, directly on Δ itself. This is precisely our approach. In fact, we will see that essentially the simplest possible mechanism following this prescription is informed-truthful for *all* domains.

Definition 10 (CA mechanism). *The* Correlated Agreement (CA) mechanism *is a multi-task mechanism with score matrix* $S = \text{Sign}(\Delta)$.

Theorem 2. *The CA mechanism is informed-truthful and proper for all worlds.*

Proof. The truthful strategy F^*, G^* has higher payment than any other pair F, G:

$$E(F^*, G^*) = \sum_{i,j} \Delta_{i,j} S(i,j) = \sum_{i,j:\Delta_{ij}>0} \Delta_{i,j} \geq \sum_{i,j} \Delta_{i,j} S(F_i, G_j) = E(F, G),$$

where the inequality follows from the fact that $S(i,j) \in \{0, 1\}$.

The truthful score is positive, while any uninformed strategy has score zero. Consider an uninformed strategy F, with $F_i = r$ for all i. Then, for any G,

$$E(F, G) = \sum_i \sum_j \Delta_{i,j} S(r, G_j) = \sum_j S(r, G_j) \sum_i \Delta_{i,j} = \sum_j S(r, G_j) \cdot 0 = 0,$$

where the next-to-last equality follows because rows and columns of Δ sum to zero. □

While informed-truthful, the CA mechanism is not always strictly proper. As discussed at the end of Section 2.1, we do not find this problematic; let us revisit this point. The peer prediction literature makes a distinction between proper and strictly proper, and insists on the latter. This comes from two motivations: (i) properness is trivial in standard models: one can simply pay the same amount all the time and this would be proper (since truthful reporting would be as good as anything else); and (ii) strict properness provides incentives to bother to acquire a useful signal or belief before making a report. Neither (i) nor (ii) is a critique of the CA mechanism; consider (i) paying a fixed amount does not give informed truthfulness, and (ii) the mechanism provides strict incentives to invest effort in acquiring a signal.

Example 7. *Continuing with Example 6, we can see why CA is not manipulable. CA considers signals that are positively correlated on bonus tasks (and thus have a positive entry in Δ) to be matching, so there is no need to agents to misreport to ensure matching. In simple cases, e.g. if only the two signals* 1 *and* 2 *are positively correlated, they are "merged," and reports of one treated equivalently to the other. In cases such as Equation 2.20, the correlation structure is more complex, and the result is not simply merging.*

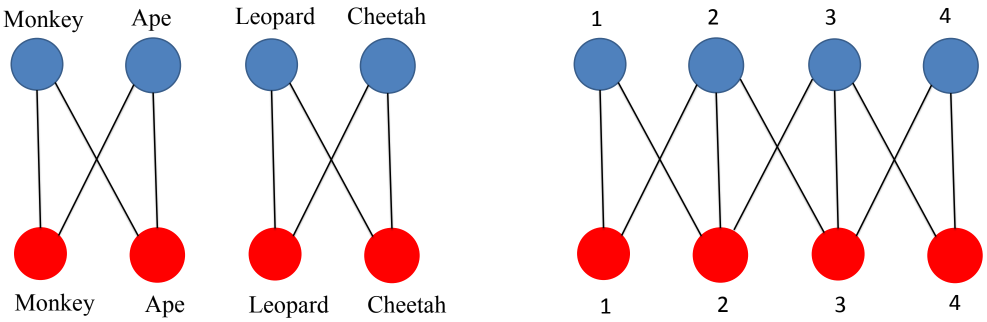


Figure 2.3: The blue and red nodes represent signals of agent 1 and 2, respectively. An edge between two signals represents that there is positive correlation between those signals. Left: A signal distribution for an image classification task with clustered signals. Right: A signal distribution for a MOOC peer assessment task or object counting task with ordinal signals and without clustered signals.

2.4.3 Strong Truthfulness of the CA Mechanism

The CA mechanism is always informed truthful. In this section we characterize when it is also strongly truthful (and thus strictly proper), and show that it is maximal in this sense across a large class of mechanisms.

Definition 11 (Clustered signals). *A signal distribution has* clustered signals *when there exist at least two identical rows or columns in* $\mathrm{Sign}(\Delta)$.

Equivalently, two signals i and i' of an agent are clustered if i is positively correlated with the same set of matched agent's signals as i'.

Example 8. *See Figure 2.3. The first example corresponds to an image classification task where there are categories such as "Monkey", "Ape", "Leopard", "Cheetah" etc. The signals "Monkey" and "Ape" are clustered: for each agent, seeing one is positively correlated with the other agent having one of the two, and negatively correlated with the other possible signals. The second example concerns models with ordinal* *signals, such as peer assessment or counting objects. In this example there are no clustered signals for either agent. For example, signal* 1 *is positively correlated with signals* 1 *and* 2*, while signal* 2 *with signals* 1*,* 2*, and* 3*.*

Lemma 3. *If $\Delta_{ij} \neq 0$, $\forall i, j$, then a joint strategy where at least one agent uses a non-permutation strategy and matches the expected score of truthful reporting exists if and only if there are clustered signals.*

Proof. Suppose clustered signals, so there exists $i \neq i'$ such that $\text{Sign}(\Delta_{i,\cdot}) = \text{Sign}(\Delta_{i',\cdot})$. Then if agent 2 is truthful, agent 1's expected score is the same for being truthful or for reporting i' whenever she receives either i or i'. Formally, consider the strategies $G = \mathbb{I}$ and F formed by replacing the i-th row in $\mathbb{I}$ by the i'-th row. Observe that $S(i, j) = S(F_i, G_j)$ as the i-th and i'-th row in S are identical. Hence, $E(F, G) = E(\mathbb{I}, \mathbb{I})$. The same argument holds for clustered signals for agent 2.

If the world does not have clustered signals, any agent using a non-permutation strategy leads to lower expected score than being truthful. Suppose F is a non-permutation strategy, such that $E(F, G) = E(\mathbb{I}, \mathbb{I})$ for some G. Then there exist signals $i \neq i'$ such $F_i = F_{i'} = r$, for some r. No clustered signals implies that $\exists j$ such that $\text{Sign}(\Delta_{i,j}) \neq \text{Sign}(\Delta_{i',j})$. Let $G(j) = j'$, for some j'. Without loss of generality assume that $\Delta(i, j) > 0$, then we get $\Delta(i', j) < 0$ as $\Delta(i', j) \neq 0$. The score for signal pair $(S_1 = i, S_2 = j)$ is $S(r, j')$ and for $(S_1 = i', S_2 = j)$ is also $S(r, j')$. Either $S(r, j') = 1$ or $S(r, j') = 0$. In both cases the strategy profile F, G will lead to a strictly smaller expected score as compared to the score of truthful strategy, since $\Delta(i, j) > 0$ and $\Delta(i', j) < 0$. Similarly, we can show that if the second agent uses a non-permutation strategy, that also leads to strictly lower expected scores for both agents. □

We now give a condition under which there are asymmetric permutation strategy profiles that give the same expected score as truthful reporting.

Definition 12 (Paired permutations). *A signal distribution has* paired permutations *if there exist distinct permutation matrices P, Q s.t. $P \cdot \text{Sign}(\Delta) = \text{Sign}(\Delta) \cdot Q$.*

Lemma 4. *If $\Delta_{ij} \neq 0$, $\forall i, j$, then there exist asymmetric permutation strategy profiles with the same expected score under the CA mechanism as truthful reporting if and only if the signal distribution has paired permutations.*

Proof. First we show that if the world has paired permutations then there exist asymmetric permutation strategy profiles that have the same expected score as truthful strategies. Consider $F = P$ and $G = Q$. From the paired permutations condition it follows that $S(i, j) = S(F_i, G_j)$, $\forall i, j$, since $S(F_i, G_j)$ is the (i, j)-th entry of the matrix $F \cdot S \cdot G^\top$ which is equal to S. Therefore, $E[F, G] = E[\mathbb{I}, \mathbb{I}]$.

To prove the other direction, let F and G be the permutation strategies of agent 1 and 2, respectively, with $F \neq G$. If the world does not have paired permutations, then $F \cdot S \cdot G^\top \neq S$. Let $\hat{S} = F \cdot S \cdot G^\top$. The expected score for F, G is

$$E[F, G] = \sum_{i,j} \Delta_{i,j} \cdot \hat{S}(i, j) \; ,$$

and the expected score for truthful strategies is

$$E[\mathbb{I}, \mathbb{I}] = \sum_{i,j} \Delta_{i,j} \cdot S(i, j) \; .$$

Combining the facts that $E[\mathbb{I}, \mathbb{I}] \geq E[F, G]$; $\Delta_{ij} \neq 0, \forall i, j$; and $\hat{S}$ differs from S by at least one entry, $E[F, G]$ will be strictly less than $E[\mathbb{I}, \mathbb{I}]$. □

Lemma 3 shows that when the world has clustered signals, the CA mechanism cannot differentiate between individual signals in a cluster, and is not strongly truthful. Similarly, Lemma 4 shows that under paired permutations this mechanism is not able to distinguish whether an agent is reporting the true signals or a particular permutation of the signals. In domains without clustered signals and paired permutations, all strategies (except symmetric permutations) lead to a strictly lesser score than truthful strategies, and hence, the CA mechanism is strongly truthful.

The CA mechanism is informed truthful, but not strongly truthful, for the image classification example in Figure 2.3 as there are clustered signals in the model. For the peer assessment example, it is strongly truthful because there are no clustered signals and a further analysis reveals that there are no paired permutations.

A natural question is whether we can do better by somehow 'separating' clustered signals from each other, and 'distinguishing' permuted signals from true signals, by giving different scores to different signal pairs, while retaining the property that the designer only needs to know $\text{Sign}(\Delta)$. Specifically, can we do better if we allow the score for each signal pair $(S_1 = i, S_2 = j)$ to depend on i, j in addition to $\text{Sign}(\Delta_{ij})$? We show that this extension does not add any additional power over the CA mechanism in terms of strong truthfulness.

Theorem 3. *If $\Delta_{ij} \neq 0$, $\forall i, j$, then CA is maximally strong truthful amongst multi-task mechanisms that only use knowledge of the correlation structure of signals, i.e. mechanisms that decide $S(i, j)$ using $\text{Sign}(\Delta_{ij})$ and index (i, j).*

Proof. We first show that the CA mechanism is strongly truthful if the signal distribution has neither clustered signals nor paired permutations. This follows directly from Lemmas 3 and 4, as strategy profiles in which any agent uses a non-permutation strategy or both agents use an asymmetric permutation strategy lead to strictly lower expected score than truthful strategies.

Next we show maximality by proving that if a signal distribution has either clustered signals or paired permutations then there do not exist any strong truthful multi-task mechanisms that only use the correlation structure of signals.

We prove this by contradiction. Suppose there exists a strongly truthful mechanism for the given signal distribution which computes the scoring matrix using the correlation structure of signals. Let the scoring matrix for the signal distribution be S.

If the signal distribution has clustered signals then at least two rows or columns in $\mathrm{Sign}(\Delta)$ are identical.

Suppose that there exist $i \neq i'$, such that the i-th and i'-th row in $\mathrm{Sign}(\Delta)$ are identical. We will construct another delta matrix Δ' representing a signal distribution that has clustered signals, for which this mechanism cannot be simultaneously strongly truthful.

Let Δ' be computed by exchanging rows i and i' of Δ. Clearly, Δ' has clustered signals. Now, the scoring matrix for both Δ and Δ' is the same, since the sign structure is the same for both. Let $G = \mathbb{I}$ and F be computed by exchanging rows i and i' of $\mathbb{I}$.

Strong truthfulness for Δ implies that

$$E_{\Delta}[\mathbb{I}, \mathbb{I}] > E_{\Delta}[F, G] \,. \tag{2.25}$$

However, observe that $E_{\Delta}[\mathbb{I}, \mathbb{I}] = E_{\Delta'}[F, G]$ and $E_{\Delta'}[\mathbb{I}, \mathbb{I}] = E_{\Delta}[F, G]$. Strong truthfulness for Δ' implies that

$$E_{\Delta'}[\mathbb{I}, \mathbb{I}] > E_{\Delta'}[F, G] \implies E_{\Delta}[\mathbb{I}, \mathbb{I}] < E_{\Delta}[F, G] \,. \tag{2.26}$$

Equation 2.25 and 2.26 lead to a contradiction, implying that the above mechanism cannot be strongly truthful.

Similarly, we can show that if two columns in $\mathrm{Sign}(\Delta)$ are identical, then there exists another delta matrix Δ' formed by exchanging the columns of the Δ for $j \neq j'$ such that the j-th and j'-th column of $\mathrm{Sign}(\Delta)$ are identical. A similar contradiction can be reached using strong truthfulness on Δ and Δ'.

The interesting case is when the signal distribution satisfies paired permutations, i.e. there exist permutation matrices $P \neq Q$ such that $P \cdot S \cdot Q^\top = S$. Consider $\Delta' = (P^{-1}) \cdot \Delta \cdot (Q^{-1})^\top$, $F = P$, and $G = Q$. We need to argue that Δ' represents a correct signal distribution and that it has paired permutations.

To see this, observe that exchanging the columns or rows of a delta matrix leads to a valid delta matrix, and pre-multiplying or post-multiplying a matrix with permutation matrices only exchanges rows or columns, respectively. Observe that the sign structure of Δ' is the same as the sign structure of Δ since $S = (P^{-1}) \cdot S \cdot (Q^{-1})^\top$, and therefore, the scoring matrix for both Δ and Δ' is the same. Due to this Δ' has paired permutations.

Strong truthfulness for Δ implies that

$$E_\Delta[\mathbb{I}, \mathbb{I}] > E_\Delta[F, G] \, . \tag{2.27}$$

However, again observe that $E_\Delta[\mathbb{I}, \mathbb{I}] = E_{\Delta'}[F, G]$ and $E_{\Delta'}[\mathbb{I}, \mathbb{I}] = E_\Delta[F, G]$. Strong truthfulness for Δ' implies that

$$E_{\Delta'}[\mathbb{I}, \mathbb{I}] > E_{\Delta'}[F, G] \implies E_\Delta[\mathbb{I}, \mathbb{I}] < E_\Delta[F, G] \, . \tag{2.28}$$

Equation 2.27 and 2.28 lead to a contradiction, implying that the above mechanism cannot be strongly truthful.

Therefore, if the signal distribution has either clustered signals or paired permutations there exist no strongly truthful scoring mechanism that assigns scores based on the correlation structure of Δ. □

This result shows that if a multi-task mechanism only relies on the correlation structure and is strongly truthful in some world model then the CA mechanism will also be strongly truthful in

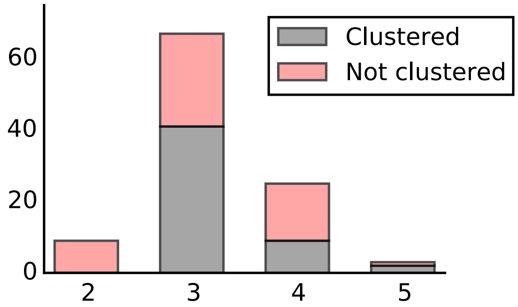


Figure 2.4: Number of MOOC peer assessment models with clustered signals (CA is informed truthful) and without clustered signals (CA is strongly truthful up to paired permutations).

that world model. Therefore, even if one uses $2 \cdot n^2$ parameters in the design of scoring matrices from $\mathrm{Sign}(\Delta)$, one can only be strongly truthful in the worlds where CA mechanism is strongly truthful, which only uses two parameters.

A remaining question is whether strongly truthful mechanisms can be designed when the score matrix can depend on the exact value of the Δ matrix. We answer this question negatively.

Theorem 4. *There exist symmetric signal distributions such that no multi-task mechanism is strongly truthful.*

Proof. Let $n = 3$, and consider any symmetric Δ matrix of the form:

$$\Delta = \begin{bmatrix} x & y & -(x+y) \\ y & x & -(x+y) \\ -(x+y) & -(x+y) & 2(x+y) \end{bmatrix},$$

for some $0 < y < x \leq 0.5$, and let

$$S = \begin{bmatrix} a & b & e \\ c & d & f \\ g & h & i \end{bmatrix},$$

for some $a, b, c, d, e, f, g, h, i$ which can be selected using complete knowledge of Δ.

We will consider three strategy profiles $(F^1, G^1), (F^2, G^2), (F^3, G^3)$, with

$$F^1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad G^1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$F^2 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad G^2 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and

$$F^3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad G^3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Using strong truthfulness condition $E[\mathbb{I}, \mathbb{I}] > E[F^1, G^1]$, we get

$$\begin{aligned} ax + by + cy + dx &> cx + dy + ay + bx \\ (a+d)(x-y) &> (c+b)(x-y) \\ a + d &> c + b \qquad , \end{aligned} \tag{2.29}$$

where the last inequality follows due to the fact that $x > y$.

Using strong truthfulness condition $E[\mathbb{I}, \mathbb{I}] > E[F^2, G^2]$, we get

$$by + cy > -dx + a(2y + x) + (g + e - f - h)(-x - y) \tag{2.30}$$

and again using strong truthfulness condition $E[\mathbb{I}, \mathbb{I}] > E[F^3, G^3]$, we get

$$by + cy > -ax + d(2y + x) + (f + h - g - e)(-x - y) \tag{2.31}$$

Now, multiplying equation 2.29 by y and combining equation it with equation 2.30, we get

$$\begin{aligned} -dx + a(2y + x) + (g + e - f - h)(-x - y) \quad &< \quad by + cy \quad < \quad ay + dy \\ \implies -dx + a(2y + x) + (g + e - f - h)(-x - y) \quad &< \quad ay + dy \\ \implies a(x + y) \quad &< \quad d(x + y) + (f + h - g - e)(-x - y) \end{aligned} \tag{2.32}$$

Similarly, equation 2.29 by y and combining equation it with equation 2.31, we get

$$\begin{aligned} -ax + d(2y + x) + (f + h - g - e)(-x - y) \quad &< \quad by + cy \quad < \quad ay + dy \\ \implies -ax + d(2y + x) + (f + h - g - e)(-x - y) \quad &< \quad ay + dy \\ \implies d(x + y) + (f + h - g - e)(-x - y) \quad &< \quad a(x + y) \end{aligned} \tag{2.33}$$

Equation 2.32 and 2.33 lead to a contradiction, implying that there does not exist any $a, b, c, d, e, f, g, h, i$ that can satisfy these equations simultaneously. Therefore, for matrices of the above form there do not exist any strongly truthful scoring matrices. □

Figure 2.4 evaluates the sign structure of the Δ matrix for the 104 MOOC questions described earlier. The CA mechanism is strongly truthful up to paired permutations when signals are not clustered, and thus in roughly half of the worlds.

2.4.4 Detail-Free Implementation of the CA Mechanism

So far we have assumed that the CA mechanism has access to the sign structure of Δ. In practice, the signs may be unknown, or partially known (e.g. the designer may know or assume that the diagonal of Δ is positive, but be uncertain about other signs).

The CA mechanism can be made detail-free in a straightforward way by estimating correlation and thus the score matrix from reports; it remains informed truthful if the number of tasks is large (even allowing for the new concern that reports affect the estimation of the distribution and thus the choice of score matrix.)

Definition 13 (The CA Detail-Free Mechanism (CA-DF)). *As usual, we state the mechanism for two agents for notational simplicity:*

1. *Each agent completes m tasks, providing m pairs of reports.*

2. *Randomly split the tasks into sets A and B of equal size.*

3. *Let T^A, T^B be the empirical joint distributions of reports on the bonus tasks in A and B, with $T^A(i,j)$ the observed frequency of signals i,j. Also, let T^A_M, T^B_M be the empirical marginal distribution of reports computed on the penalty tasks in A and B, respectively, with $T^A_M(i)$ the observed frequency of signal i. Note that we only take one sample per task to ensure the independence of samples.*

4. *Compute the empirical estimate of the Delta matrix, based on reports rather than signals: $\Gamma^A_{ij} = T^A(i,j) - T^A_M(i)T^A_M(j)$, and similarly for Γ^B.*

5. *Define score matrices,* swapping task sets*: $S^A = \mathrm{Sign}(\Gamma^B)$, $S^B = \mathrm{Sign}(\Gamma^A)$. Note that S^A does not depend on the reports on tasks in A.*

6. *Apply the CA mechanism separately to tasks in set A and set B, using score matrix S^A and S^B for tasks in set A and B, respectively.*

Lemma 5. *For all strategies F, G and all score matrices $S \in \{0, 1\}^{n \times n}$, $E(S^*, \mathbb{I}, \mathbb{I}) \geq E(S, F, G)$ in the multi-task mechanism, where $E(S, F, G)$ is the expected score of the mechanism with a fixed score matrix S.*

Proof. The expected score for arbitrary score matrix and strategies is:

$$E(S, F, G) = \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij} S(F_i, G_j)$$

The expected score for truthful reporting with S^* is

$$E(S^*, \mathbb{I}, \mathbb{I}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij} \,\mathrm{Sign}(\Delta)_{ij} = \sum_{i,j:\Delta_{ij}>0} \Delta_{ij} \geq \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij} S(F_i, G_j),$$

where the inequality follows because S is a $0/1$ matrix. □

The lemma gives the main intuition for why CA-DF is informed truthful for large m: even if agents could set the score matrix completely independently of their strategies, the “truthful” score matrix S^* is the one that maximizes payoffs. To get a precise result, the following theorem shows that a score matrix “close” to S^* will be chosen with high enough probability.

Theorem 5 (Mechanism CA-DF is (ε, δ)-informed truthful). *Let $\varepsilon > 0$ and $\delta > 0$ be parameters. Then there exists a number of tasks $m = O(n^3 \log(1/\delta)/\varepsilon^2)$ (for n signals), such that with probability at least $1 - \delta$, there is no strategy profile with expected score more than ε above truthful reporting, and any uninformed strategy has expected score strictly less than truthful. Formally, with probability at least*

$1-\delta$, $E(F,G) \leq E(\mathbb{I},\mathbb{I})+\varepsilon$, for all strategy pairs F, G; for any uninformed strategy F_0 (equivalently G_0), $E(F_0,G) < E(\mathbb{I},\mathbb{I})$.

Proof. Let H^A and H^B be the (unobserved) joint signal frequencies, which are a sample from the true joint distribution. Let M^A and M^B be the (unobserved) marginal signal frequencies, which are a sample from the true marginal distribution. Finally, let Δ^A and Δ^B the corresponding empirical Delta matrices. Fixing strategies F, G, S^A is a function of H^B and M^B, and independent of H^A and M^A. This means that we can write the expected score for tasks in A as

$$E(S^A,F,G)=\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}S^A(F_i,G_j). \tag{2.34}$$

By Lemma 5, we know that $E(S^*,\mathbb{I},\mathbb{I}) \geq E(S,F,G)$ for all S, F, G, and will show that once m is large enough, being truthful gets close to this score with high probability. We have

$$\begin{aligned}|E(S_A,\mathbb{I},\mathbb{I})-E(S^*,\mathbb{I},\mathbb{I})| &= |E(\mathrm{Sign}(\Delta^B),\mathbb{I},\mathbb{I})-E(\mathrm{Sign}(\Delta),\mathbb{I},\mathbb{I})| &(2.35)\\ &= |\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}(\mathrm{Sign}(\Delta^B)_{ij}-\mathrm{Sign}(\Delta)_{ij})|\,. &(2.36)\end{aligned}$$

Therefore, for some accuracy ε and confidence δ, with $m=O(n^3\log(1/\delta)/\varepsilon^2)$, we want

$$|\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}(\mathrm{Sign}(\Delta^B)_{ij}-\mathrm{Sign}(\Delta)_{ij})| \leq \varepsilon\,. \tag{2.37}$$

Observe that

$$|\sum_{i,j}\Delta_{ij}(\mathrm{Sign}(\Delta^B)_{ij}-\mathrm{Sign}(\Delta)_{ij})| \leq \sum_{i,j}|\Delta_{ij}(\mathrm{Sign}(\Delta^B)_{ij}-\mathrm{Sign}(\Delta)_{ij})| \tag{2.38}$$

$$\leq \sum_{i,j} |\Delta_{ij} - \Delta^B_{ij}| \,. \tag{2.39}$$

Therefore, it is sufficient to learn Δ^B such that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} |\Delta_{ij} - \Delta^B_{ij}| \leq \varepsilon \,. \tag{2.40}$$

We now use a standard result (see e.g. (Devroye & Lugosi, 2001), Theorems 2.2 and 3.1), that any distribution over finite domain Ω is learnable within L1 distance d in $O(|\Omega|/d^2)$ samples with high probability, specifically $1 - \delta$ with an additional $\log(1/\delta)$ factor.

Using this result we can learn the joint signal distribution of the agents using $O(9n^2/\varepsilon^2)$ samples with accuracy $\varepsilon/3$. We can also learn the marginal distribution of agents' signals using $O(9n^3/\varepsilon^2)$ samples from the true marginal distribution with accuracy $\varepsilon/3n$. With high probability, after these many samples from each of these distributions, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} |P_{ij} - H^B_{ij}| \leq \frac{\varepsilon}{3} \tag{2.41}$$

$$\sum_{i=1}^{n} |P_i - M^B_i| \leq \frac{\varepsilon}{3n} \,. \tag{2.42}$$

Now,

$$\sum_{i,j} |\Delta_{ij} - \Delta^B_{ij}| = \sum_{i,j} |P_{ij} - H^B_{ij} - (P_i P_j - M^B_i M^B_j)| \tag{2.43}$$

$$\leq \sum_{i,j} |P_{ij} - H^B_{ij}| + \sum_{i,j} |P_i P_j - M^B_i M^B_j| \quad \text{(Triangle Ineq.)} \tag{2.44}$$

$$\leq \frac{\varepsilon}{3} + \sum_{i,j} |P_i P_j - M_i^B \left(P_j \pm \frac{\varepsilon}{3n}\right)| \qquad \text{(Using Eq. 2.41 \& 2.42)} \tag{2.45}$$

$$= \frac{\varepsilon}{3} + \sum_{i,j} |P_i P_j - M_i^B P_j \pm M_i^B \frac{\varepsilon}{3n}| \tag{2.46}$$

$$\leq \frac{\varepsilon}{3} + \sum_{i,j} | \left(P_i - M_i^B\right) P_j| + \sum_{i,j} M_i^B \frac{\varepsilon}{3n} \quad \text{(Triangle Ineq.)} \tag{2.47}$$

$$= \frac{\varepsilon}{3} + \sum_{i,j} P_j |P_i - M_i^B| + \sum_{i,j} M_i^B \frac{\varepsilon}{3n} \tag{2.48}$$

$$= \frac{\varepsilon}{3} + \sum_{i,j} P_j |P_i - M_i^B| + \sum_{j} \frac{\varepsilon}{3n} \tag{2.49}$$

$$\leq \frac{\varepsilon}{3} + \sum_{j=1}^{n} \sum_{i=1}^{n} |P_i - M_i^B| + n \frac{\varepsilon}{3n} \qquad (|P_j| \leq 1) \tag{2.50}$$

$$\leq \frac{\varepsilon}{3} + \sum_{j=1}^{n} \frac{\varepsilon}{3n} + \frac{\varepsilon}{3} \qquad \text{(Using Eq. 2.42)} \tag{2.51}$$

$$= \varepsilon\,. \tag{2.52}$$

We now conclude

$$|E(S_A, \mathbb{I}, \mathbb{I}) - E(S^*, \mathbb{I}, \mathbb{I})| \quad \leq \quad \sum_{i=1}^{n} \sum_{j=1}^{n} |\Delta_{ij} - \Delta_{ij}^B| \quad \leq \varepsilon\,, \tag{2.53}$$

which implies $E(S_A, \mathbb{I}, \mathbb{I}) + \varepsilon \geq E(S, F, G)$ for all S, F, G.

Finally, note that the expected value of uninformed strategies is 0, because $E(S, F^0, G) = 0$ for any uninformed F^0, regardless of score matrix, while ε can always be set small enough ensuring that being truthful has positive expected payoff. □

2.4.5 Agent Heterogeneity

The CA mechanism only uses the signs of the entries of Δ to compute scores, not the exact values. This means that the results can handle some variability across agent "sensing technology," as long as the sign structure of the Δ matrix is uniform across all pairwise matchings of peers. In the binary signal case, this reduces to agents having positive correlation between their signals, giving exactly the heterogeneity results in Dasgupta & Ghosh (2013). Moreover, the agents themselves do not need to know the detailed signal model to know how to act; as long as they believe that the scoring mechanism is using the correct correlation structure, they can be confident in investing effort and simply report their signals truthfully.

2.4.6 Unintended Signals

Finally, we discuss a seemingly pervasive problem in peer prediction: in practice, tasks may have many distinctive attributes on which agents may base their reports, in addition to the intended signal, and yet all models in the literature assume away the possibility that agents can choose to acquire such unintended signals. For example, in online peer assessment where students are asked to evaluate the quality of student assignments, students could instead base their assessments on the length of an essay or the average number of syllables per word. In an image categorization system, users could base their reports on the color of the top-left pixel, or the number of kittens present (!), rather than on the features they are asked to evaluate. Alternative assessments can benefit agents in two ways: they may require less effort, and they may result in higher expected scores via more favorable Delta matrices.[9]

We can characterize when this kind of manipulation cannot be beneficial to agents in the CA mechanism. The idea is that the amount of correlation coupled with variability across tasks should be large enough for the intended signal. Let η represent a particular *task evaluation strategy*, which may involve acquiring different signals from the task than intended. Let Δ^{η} be the corresponding Δ matrix that would be designed if this was the signal distribution. This is defined on a domain of signals that may be distinct from that in the designed mechanism. In comparison, let η^*

[9] This issue is related to the perennial problem of spurious correlations in classification and regression.

define the task evaluation strategy intended by the designer (i.e., acquiring signals consistent with the mechanism's message space), coupled with truthful reporting. The expected payment from this behavior is $\sum_{ij:\Delta_{ij}^{\eta^*}>0} \Delta_{ij}^{\eta^*}$.

The maximal expected score for an alternate task evaluation strategy η may require a strategy remapping signal pairs in the signal space associated with η to signal pairs in the intended mechanism (e.g., if the signal space under η is different than that provided by the mechanism's message space). The expected payment is bounded above by $\sum_{ij:\Delta_{ij}^{\eta}>0} \Delta_{ij}^{\eta}$. Therefore, if the expected score for the intended η^* is higher than the maximum possible score for other η, there will be no reason to deviate.

2.5 Conclusion

We study the design of peer prediction mechanisms that leverage signal reports on multiple tasks to ensure informed truthfulness, where truthful reporting is the joint strategy with highest payoff across all joint strategies, and strictly higher payoff than all uninformed strategies (i.e., those that do not depend on signals or require effort). We introduce the CA mechanism, which is informed-truthful in general multi-signal domains. The mechanism reduces to the Dasgupta & Ghosh (2013) mechanism in binary domains, is strongly truthful in categorical domains, and maximally strongly truthful among a broad class of multi-task mechanisms. We also present a detail-free version of the mechanism that works without knowledge of the signal distribution while retaining ε-informed truthfulness. Interesting directions for future work include: (i) adopting a non-binary model of effort, and (ii) combining learning with models of agent heterogeneity.

3

Measuring Performance Of Peer Prediction Mechanisms Using Replicator Dynamics

Peer prediction formalizes the challenge of eliciting information from agents in settings without verification. Whereas scoring rules (Gneiting & Raftery, 2007) and prediction markets (Hanson, 2003; Chen et al., 2007) can be used to elicit beliefs about observable events (e.g., the outcome of the U.S. Presidential election), peer prediction addresses settings without direct access to the ground truth. Consider, for example, eliciting information about noise in a restaurant, about the quality of an e-commerce search algorithm, or the suggested grade for a student's assignment in an online course, where obtaining ground truth is either not possible or costly.

The theory of peer prediction has developed rapidly in recent years. From the simple approach of output agreement (von Ahn & Dabbish, 2004; Waggoner & Chen, 2014), the field has moved to scoring-rule based approaches with varying knowledge requirements on the part of the designer (Miller et al., 2005; Witkowski & Parkes, 2012a), later relaxing the requirement of a common prior (Witkowski & Parkes, 2012b; Radanovic & Faltings, 2013; Kamble et al., 2015). These early mechanisms all had *uninformative* equilibria, where agents could make reports without looking at their assigned task, and yet get a higher score than by being truthful. Several recent papers propose mechanisms that ensure that truthfulness is not only a strict correlated equilibrium, but has higher payoff than certain other strategies.

Jurca and Faltings [2009] discourage strategies where all agents report identically, by rewarding near-agreement rather than complete agreement with peers. Radanovic and Faltings [2015a] present the logarithmic peer truth serum, with a large population and many peers performing each task, comparing an agent's agreement with their peers to their agreement with the population as a whole. Dasgupta and Ghosh [2013] propose a *multi-task* approach, where each agent completes multiple tasks, and compare agreement on overlapping tasks to expected agreement on non-overlapping ones, showing that truthfulness is optimal for settings with binary reporting. Chapter 2 extended this method to settings with more than two possible reports.

There is also experimental work on peer prediction. One study (Gao et al., 2014) showed that Mechanical Turk workers are able to coordinate on an uninformative equilibrium in some peer prediction mechanisms, while behaving in an unpredictable way in a design inspired by Jurca & Faltings (2009). A second experimental study is more positive, showing that simple scoring mechanisms can encourage effort, and that workers do not seem to coordinate on uninformative equilibria (Faltings et al., 2014).[1]

We adopt replicator dynamics as a model of population learning in peer prediction mechanisms. Our interest is to understand the robustness of different designs when, rather than precomputing equilibria, participants adjust their behavior via a simple dynamic. Learning is widely used to study behavior in games, giving a useful measure of the likelihood that various equilibria emerge in repeated play of a mechanism, as well as the stability of those equilibria. Intuitively, these dynamics capture how players may adjust their behavior slightly each round depending on the success of their previous actions. While truthfulness may be an equilibrium of the game, if learning dynamics steer away from it, one may not expect to see (long-lasting) truthful behavior in practice.

Analyzing models derived from peer evaluation data in several massive online courses, we confirm concerns about uninformative equilibria in early peer prediction mechanisms: despite the existence of a truthful equilibrium, learning dynamics move toward uninformative equilibria in these mechanisms. The learning dynamics still tend toward all participants adopting the same uniformed report in the approach of Jurca and Faltings [2009]. In contrast, the multi-task mechanisms do better, with a larger basin of attraction of the truthful equilibrium. Truthfulness is most stable under the *correlated agreement* mechanism (see Chapter 2), which generalizes the method of Dasgupta and Ghosh [2013], while the logarithmic peer truth serum (Radanovic & Faltings, 2015a) does not work well unless each task is performed by a comparatively large number of agents.

[1] A possible reason for the difference in results is that the environment in this second study had many possible reports, making it harder to coordinate.

3.0.1 Case Study: Peer Grading

To choose realistic parameters for our experimental study, we use data from peer evaluation in several Massive Open Online Courses (MOOCs).[2] Organizations such as edX, Coursera, and many others around the world are scaling online learning to tens of thousands of students per course without a corresponding expansion in course staff. A key challenge is to scalably teach topics that are difficult to automatically assess, such as writing, judgement, or design. Peer evaluation is a promising tool—students submit assignments, which are evaluated by several peers using an instructor-created rubric. Peers provide scores as well as written feedback.

In today's systems, the evaluators are not scored, though participation can be coupled with being able to see feedback from their peers. This means that students can (and do) submit minimal feedback without giving it much thought.[3] This setting fits the peer prediction model—it is expensive for staff to make "ground truth" evaluations by grading submissions, and because several peers evaluate each submission, their assessments are naturally correlated and can be compared.

Other research on scalable peer evaluation evaluates students' assessment skills, identifies and compensates for their biases (Piech et al., 2013), and helps students self-adjust for bias (Kulkarni et al., 2013). The Mechanical TA project (Wright & Leyton-Brown, 2015) aims to reduce TA workload in high-stakes peer grading.

3.0.2 Background on Replicator Dynamics

We use one of the simplest models of evolutionary population dynamics, which were first introduced to study evolution (Smith, 1972; Sandholm, 2009; Gintis, 2009). Such models track segments of a population, gradually adjusting behavior in response to feedback. Evolutionary

[2] A anonymous, summarized version of the data set used in this chapter is available at `http://www.eecs.harvard.edu/~shnayder/`.

[3] This is a well-known issue in on-campus peer-evaluation settings as well, though there, instructors can review the feedback and intervene. In MOOCs, that kind of oversight may not be scalable.

dynamics have been used in many applications besides evolutionary biology. For example, Erev & Roth (1998) show that learning dynamics can capture key features of human behavior in economic games, and they have many applications in multi-agent systems (Bloembergen et al., 2015).

Replicator dynamics track a continuous population of agents playing a game over time, with each agent adopting a pure strategy and probabilistically switching to higher-payoff strategies in proportion to the gain in expected payoff. Nash equilibria are known to be fixed points of replicator dynamics, but the converse need not hold (Easley & Kleinberg, 2010, Thm 12.6). These dynamics also provide an appealing model for learning at the individual level, as they are a continuous-time limit of the *multiplicative-weights* learning algorithm, and guarantee no regret (Hofbauer et al., 2009, Prop 4.1 and Prop 6.2). See Arora et al. (2012) for more about the multiplicative weights algorithm.

Replicator dynamics have been used to compute the symmetric, mixed equilibria in empirical game theory (Reeves et al., 2005). Recently, replicator dynamics have been applied to assess the likelihood or stability of various equilibria in games (Panageas & Piliouras, 2014) (see also (Kleinberg et al., 2011, 2009)). We employ this latter interpretation; specifically, we adopt the basin of attraction of the truthful equilibrium, meaning the set of strategy profiles leading eventually to the equilibrium, as a proxy for how likely and how stable truthfulness would be under repeated play.

3.1 Model

There is a continuum of agents, representing a distribution over strategies observed in the population. At each time t, finite groups of agents are sampled from this distribution, and each group is assigned to a particular task (e.g., label an image, evaluate a particular homework submission, judge the mood of a video clip, etc), which has a hidden type $h \in H$.

Each agent i privately observes a signal $s_i \in S = \{0, 1, \ldots, n-1\}$, identically and independently distributed, conditioned on type h. Let $\Pr(h)$ denote the type prior and let $\Pr(s|h)$ denote the signal distribution conditioned on type. For simplicity, we assume that the number of types is equal to the number of signals. For example, in a peer evaluation setting, the hidden type would be the "true" quality of a submission, and the signal a student's assessment of the quality, both on a scale of e.g. 0, 1, or 2. We assume that $\Pr(h)$ and $\Pr(s|h)$ are the same for all tasks and all agents, though the methodology extends to heterogeneous agent populations with non-identical signal models.

Once the agents observe their signals, they use a strategy, θ, to compute a report $r_i = \theta(s_i)$. In general, θ can be randomized, but we focus on deterministic strategies, relying on the random sampling from the population for mixing. A peer-prediction mechanism, without knowing the hidden type or the observed signals, computes a score σ_i for each agent based on reports. This score can depend on the reports of *peer* agents who did the same task, as well as on the overall set of reports across all tasks. A good scoring rule leads agents to maximize expected score by truthfully revealing their signals, and is robust to alternate equilibria as well as misreports or noise from other agents. A special concern is to prevent high-payoff, *uninformed equilibria*, where agents adopt signal-independent strategies; e.g., "always report 1."

We represent the *population strategy profile* as a distribution $x = (x_1, \ldots, x_m)$, where x_k is the fraction of agents who adopt strategy θ_k, and m is the total number of strategies. Let $U(k, x)$ denote the expected payoff from strategy θ_k given population profile x. The average population payoff is defined as

$$A(x) = \sum_{k=1}^{m} x_k U(k, x), \tag{3.1}$$

leading to the replicator dynamics differential equation:

$$\dot{x}_k = x_k(U(k,x) - A(x)). \tag{3.2}$$

We numerically solve this equation for particular starting strategy profiles to predict whether the population will tend toward the all-truthful profile. The set of profiles that lead to all-truthful is the *basin of attraction* of truthfulness.

Replicator dynamics does not itself require a model of agent beliefs—it simply assumes a particular pattern of strategy evolution. However, there is a complementary way to think about beliefs: if an agent believes the overall population strategy distribution to be in the basin of attraction of the all-truthful equilibrium, being truthful is a simple and robust strategy that will maximize reward. A large basin of attraction for truthfulness means that agents who believe that the majority of agents will be truthful will not be tempted by niche strategies. In contrast, if truthfulness has a small basin of attraction, believing that even a small fraction of agents are being strategic can make joining them attractive.

3.1.1 Peer Prediction Mechanisms

We focus on *strictly proper* peer prediction mechanisms, where truthful reporting is a strict correlated equilibrium.

Single-task mechanisms. We first define mechanisms that only depend on the reports for a single task.

(1) **Output Agreement (OA)** (von Ahn & Dabbish, 2004). The system picks a reference agent j for each agent i, and defines $\sigma_i(r_i, r_j) = A(r_i, r_j)$, where $A(x, y)$ is 1 if $x = y$, 0 otherwise. The OA mechanism is only strictly proper when the model is *self-dominant*—observing a signal s makes s the most likely signal for a reference agent as well (see Frongillo & Witkowski (2016) for an elaboration). A useful property of OA is that it is *detail-free*, requiring no knowledge of the probabilistic model of the world.

(2) **MRZ**. The peer prediction method (Miller et al., 2005) (MRZ), which uses proper scoring rules (Gneiting & Raftery, 2007) to achieve strict properness. In MRZ, the system gets a report r_i, picks a reference peer j, and uses a proper scoring rule R based on the likelihood of r_j given r_i. By the properties of proper scoring rules, this makes truthful reporting a strict correlated equilibrium. In our experiments, we use the *log scoring rule* $R(\gamma, o) = \log(\gamma_o)$, where γ is a probability distribution over outcomes, and o is the observed outcome. MRZ is not detail-free, as computing γ requires knowledge of the world model.

(3) **JF09**. A problem with both OA and MRZ is that they also have uninformative, pure-strategy symmetric Nash equilibria, one of which always results in the highest possible payoff (Jurca & Faltings, 2005). The JF09 (Jurca & Faltings, 2009) mechanism removes these (pure) Nash equilibria in binary settings, relying on four or more peers doing a single task. To evaluate a report r_i in a binary signal setting ($S = \{0, 1\}$), the mechanism picks three reference agents, defines z_i as the total number of 1 reports among them, and gives score $\sigma_i(r_i, z_i) = M[r_i, z_i]$, where M is the matrix

$$\begin{pmatrix} 0 & \alpha & 0 & \varepsilon \\ \varepsilon & 0 & \beta & 0 \end{pmatrix}. \tag{3.3}$$

α and β are set based on the world parameters to preserve strict properness, while the form of the payoff matrix ensures that if all agents coordinate on 0 or 1, they get score 0.[4] JF09 is not detail free because the designer needs the world model to compute the score matrix.

Multitask mechanisms. The next two mechanisms are *strong truthful*, meaning that all agents being truthful is an equilibrium with higher payoff than any other strategy profile, with the inequality strict except for signal permutations.

(4) **RF15**. The RF15 (Radanovic & Faltings, 2015a) mechanism scores an agent based on the

[4] There are no results about mixed equilibria. Our analysis in Section 3.2 shows them to be problematic.

statistical significance of the agent's report compared to the reports of their peers and the distribution of reports in the entire population across multiple tasks. Given report r_i and the fractions $z_{\text{peer}}, z_{\text{global}}$ of reference peers and global population respectively reporting r_i, the agent's score is $\sigma_i = \log(z_{\text{peer}}/z_{\text{global}})$. As the number of reference peers goes to infinity, this approaches $\log(\Pr(r_{peer} = r_i)/\Pr(r_i))$.[5] RF15 is detail-free.

(5) **DG13**. The DG13 mechanism (Dasgupta & Ghosh, 2013) is detail-free and multi-task, so each agent reports on several tasks. It is defined for binary signals. An agent is rewarded for being more likely to match the reports of peers doing the same task than the reports of peers doing other tasks.

We present a slightly generalized form, parametrized by a score matrix Λ. The mechanism is described, w.l.o.g., for two agents, 1 and 2:

1. Assign the agents to three or more tasks, with each agent to two or more tasks, including at least one overlapping task. Let M_s, M_1, and M_2 denote the shared, agent-1 and agent-2 tasks, respectively.

2. Let r_1^k denote the report received from agent 1 on task k (and similarly for agent 2). The payment to both agents for a shared task $k \in M_s$ is

$$\sigma_i = \Lambda(r_1^k, r_2^k) - \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \Lambda(i,j) \cdot h_{1,i} \cdot h_{2,j},$$

where $\Lambda : \{0,\ldots,n-1\} \times \{0,\ldots,n-1\} \to \mathbb{R}$ is a score matrix, $h_{1,i} = \frac{|\{\ell \in M_1 | r_1^\ell = i\}|}{|M_1|}$ is the empirical frequency with which agent 1 reports signal i in tasks in set M_1, and $h_{2,j} = \frac{|\{\ell \in M_2 | r_2^\ell = j\}|}{|M_2|}$ is the empirical frequency with which agent 2 reports signal j in tasks in set

[5] Because log is non-linear, the expected score with a finite number of reference peers is lower than this limit, even in a continuous population, and this affects the attractiveness of different strategies. We examine this effect in Section 3.2.

M_2.

3. The total payment to an agent is the sum of the payments across all shared tasks.

In the DG13 mechanism, Λ is the identity matrix ('1' for agreement, '0' for disagreement.) For binary signals and positive correlation between signals, DG13 is strong truthful.

(6) **DGMS.** Shnayder et al. (see Chapter 2) extend the DG13 mechanism in two ways. The first is DGMS, the direct extension of DG13 to multiple signals, using the identity matrix for scoring. DGMS is strong truthful when the world satisfies a *categorical* property, where, given an agent's signal, the likelihood of peers having any other signal goes down: $\Pr(s'|s) < \Pr(s')$ for all $s' \neq s$; this property holds trivially for binary signal models with positive correlation.

(7) **Correlated Agreement (CA).** The second extension of DG13 yields the CA mechanism, which adopts a different scoring rule. Rather than the identity matrix, CA sets $\Lambda(i,j) = 1$ if $\Pr(s_j|s_i) > \Pr(s_i)$, and 0 otherwise; it rewards agreement on positively correlated signals. CA reduces to DGMS in categorical settings. In general settings, it is proper (not strictly), and *informed truthful*. The payoff for truthfulness is weakly higher than any other strategy profile, and strictly higher than any uninformed, signal-independent reporting strategy. CA only requires that the designer know the direction of correlation between pairs of signals, not the entire world model.

(8) **Robust Peer Truth Serum (RPTS)** (Radanovic et al., 2016). This is a version of OA, with scores scaled based on observed report frequencies; the system collects all reports, computes the empirical prior $\hat{P}(r)$ of each report r, and defines score $\sigma(r, r') = 1/\hat{P}(r)A(r, r')$, where r' is a reference report, just as in OA. This results in extra reward for matches on uncommon reports, which has two benefits, when the number of tasks is large enough to make the empirical estimates accurate: the model is now truthful if the world model is self-predicting—observing a signal s increases the likelihood that a peer also sees s—rather than self-dominant, and constant reporting now has lower expected score than truthfulness.

(9)**Kamble** (Kamble et al., 2015). This is another scaled version of OA. The system collects all reports, computes the empirical joint $\hat{P}(r, r')$, and defines $\sigma(r, r') = 1/\sqrt{\hat{P}(r, r)}A(r, r')$, or 0 if $\hat{P}(r, r)$ is exactly 0 or 1. This has similar benefits to RPTS: the model is now truthful for any non-trivial world model, and constant reporting again has lower expected score than truthfulness.

3.1.2 Strategy Selection

To fully define the replicator dynamics, we need to instantiate a finite set of strategies available to the population. In mechanisms where agents do multiple tasks per round, each agent uses the same strategy for each task. We omit permutation strategies, which exchange the names of signals in a 1-to-1 mapping, from our analysis. These are unnatural in practice, and do not give higher payoffs than the remaining strategies in the mechanisms we study.

With two signals, the remaining pure strategies are *const0, const1, T*, corresponding to agents always reporting *0, 1*, or being truthful, respectively. For three or more signals, there are more strategies possible, and we include the monotonic strategies that overreport, underreport, or merge adjacent signals, using *const0, const1, const2, merge01, merge12, bias+, bias-, T*. *merge01* reports 0 for signals 0 and 1. *merge12* reports 1 for signals 1 and 2. *bias+* over-reports, mapping signal i to $\min(i + 1, n - 1)$. *bias-* maps i to $\max(i - 1, 0)$. For four signals, we add *mergeAdj*, which reports 0 for signals 0 and 1, and 2 for signals 2 and 3. For five signals, we add *mergeEach3* which rounds down to the nearest multiple of three. The merging strategies lose information and increase the frequency of agreement, and are an intermediate step between truthfulness and constant reports.

As a simple model of effort, we distinguish between *informed* and *uninformed* strategies. An informed strategy depends on the agent's signal. In contrast, constant strategies such as *const1* are uninformed. The distinction reflects that it takes effort to obtain a signal, so informed misreporting strategies are less appealing to agents than uninformed ones.

World	Pr(h)	Pr$(s \mid h)$	Description
W2a	[0.5, 0.5]	$\begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}$	Strong correlation
W2b	[0.5, 0.5]	$\begin{pmatrix} 0.4 & 0.6 \\ 0.1 & 0.9 \end{pmatrix}$	Bias toward 1
W3a	[0.3, 0.3, 0.4]	$\begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$	Unbiased noise
W3b	[0.3, 0.3, 0.4]	$\begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.4 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$	0 and 1 correlated

Figure 3.1: Our manually selected world models.

For certain strategy profiles and mechanisms, there may be multiple strategies with equal payoff. When there is a tie between truthfulness and another informed strategy, we believe it is natural for agents to be truthful— it is simpler because it does not require strategic reasoning, while the effort of signal acquisition is needed either way. To model this, we add a tiny cost to the expected payoffs for non-truthful informed strategies, so as to break such ties in favor of truthfulness.[6]

3.1.3 World Models

Our initial qualitative analysis compares the mechanisms in four world models, selected to illustrate common scenarios; the worlds vary the correlation between agent signals and include bias toward particular values (Figure 3.1).

3.2 Replicator Dynamics of Peer Prediction

Starting with the single-task mechanisms, we show that in OA, MRZ, and JF09, non-truthful equilibria are attractors of replicator dynamics and the basin of attraction of truthfulness is small.

[6] The consequence for replicator dynamics is that areas of the strategy simplex where the derivative between truthful and another informed strategy was exactly zero now tend toward truthful.

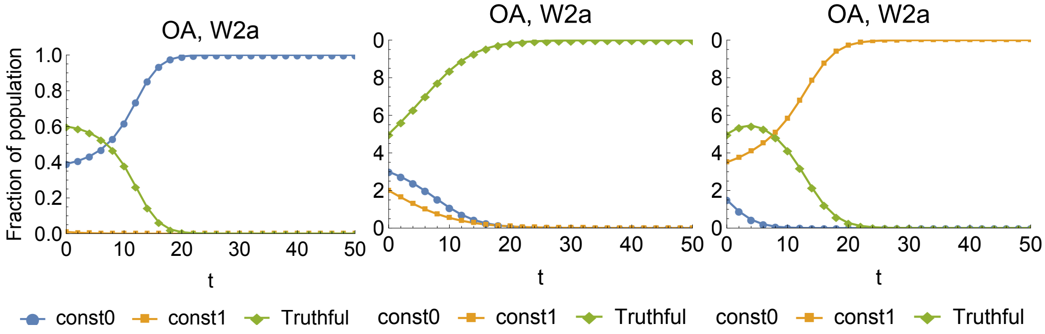


Figure 3.2: Replicator dynamics of OA for different initial strategy distributions. Even when a large fraction of the population starts out truthful, the dynamics can converge to all-ones or all-zeros uninformative equilibria.

3.2.1 Single-Task Mechanisms

We first look at the replicator dynamic for OA in the W2a world (Figure 3.2). This illustrates replicator dynamics for different initial values. At least half the population starts out truthful in each plot, but the dynamics can still converge to an uninformative strategy where all agents say 0 or 1.

It is difficult to understand the overall dynamics of a mechanism from plots of strategies versus time, because each only shows a particular starting point. A better visualization for analyzing convergence is a flow plot of the derivatives of the replicator equation, as shown in Figure 3.3. The area in green shows the *basin of attraction*, the set of starting points from which the dynamics converge to truthful play (the bottom-left corner). From the plots for W2b, we see that OA is not strictly proper, and that the all-0 and all-1 corners are much stronger attractors than truthfulness for MRZ.

From the JF09 plots, we can clearly see that even though the $(1, 0)$ and $(0, 1)$ corners are not equilibria, there are still attractors very nearby. This illustrates how replicator dynamics complement equilibrium analysis, showing that the pure-strategy-only theoretical guarantees of JF09 are

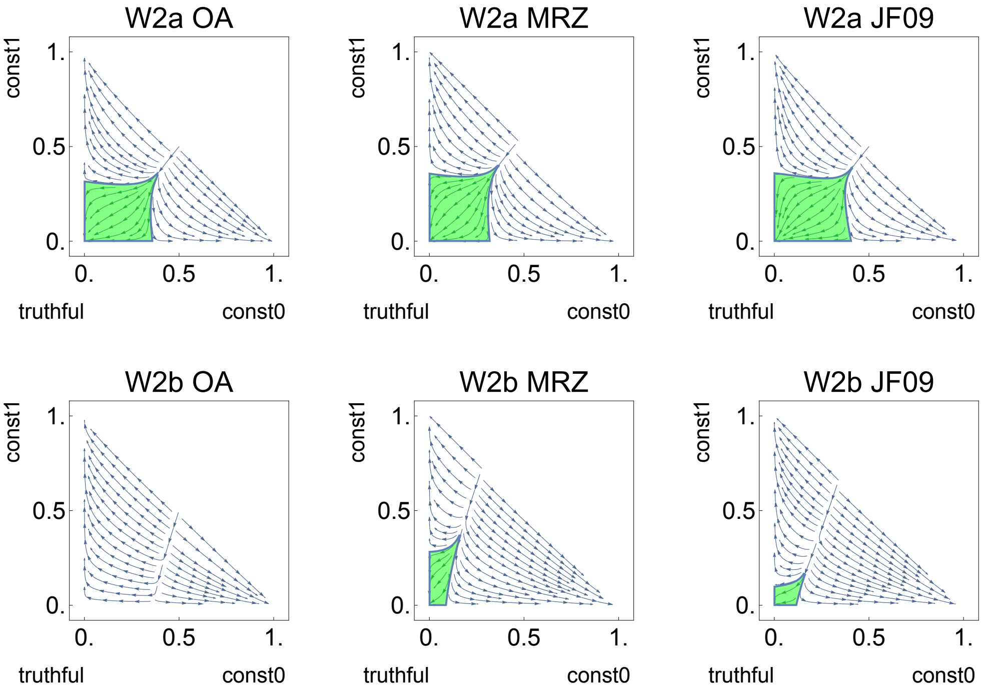


Figure 3.3: Flow plots of the derivative of the replicator equation (Eqn 3.2) for W2a and W2b, with OA, MRZ, and JF09. The all-truthful strategy profile is at $(0, 0)$, with its basin of attraction shown by the green shaded area. OA is not truthful for W3. Even when the mechanisms are truthful, the basins of attraction can be quite small.

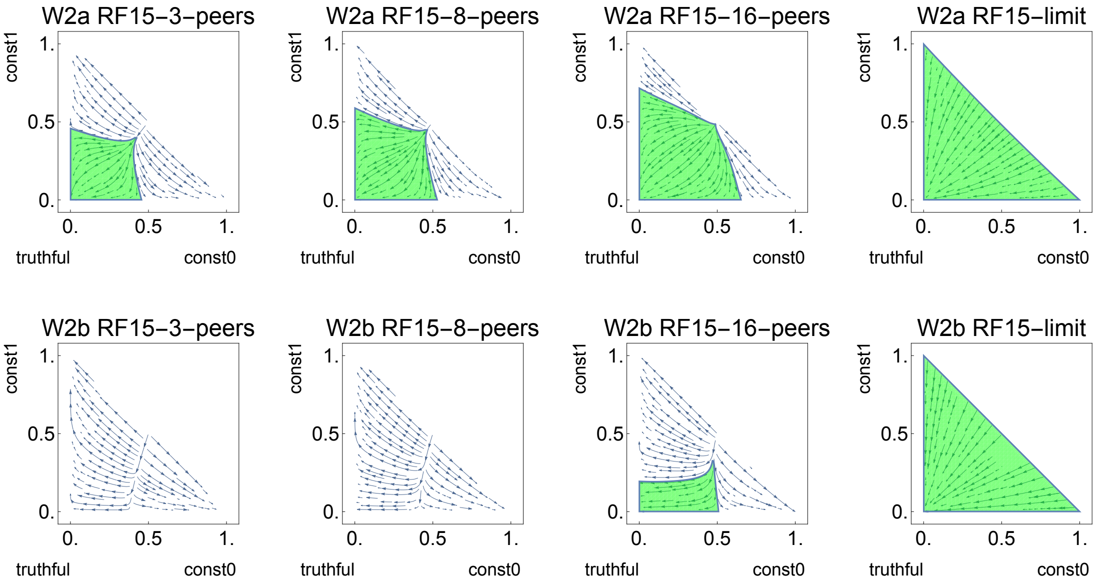


Figure 3.4: RF15 with finite sets of peers. The non-linearity of the log function makes RF15 far less robust with small numbers of peers, with much smaller basins of attraction for truthfulness.

not robust.

3.2.2 Multi-Task Mechanisms

Multi-task mechanisms leverage reports across multiple tasks to make coordination on uninformed behavior less attractive to agents.

We first confirmed that constant reporting is longer an attractive strategy in replicator dynamics under DG13 and RF15 for any binary world with correlated signals, including W2a and W2b. Instead, the basin of attraction of truthful play covers the entire strategy simplex.

However, for RF15, this is in the large-population limit, as both the total population and the number of reference peers for each task go to infinity. Figure 3.4 shows what happens when the population is large (formally, a continuum), but the number of peers per task is finite. We see that a large group of reference peers is needed for RF15 to behave as in its limit— even with 16 peers,

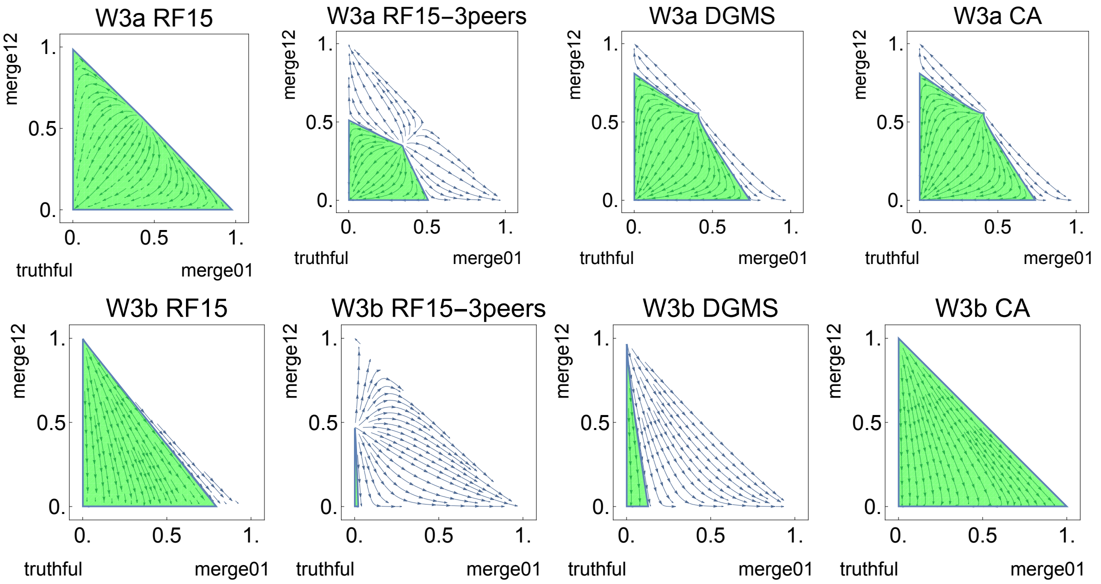


Figure 3.5: Flow plots for W3a, a categorical world, and W3b, a non-categorical one. *merge01* is the highest payoff strategy under DGMS, and is a strong attractor.

the non-linearity of the log function in the definition of the score rule makes constant reporting attractive if enough of the population agrees. Going forward, when using RF15 with a finite number of reference peers we fix this number to three and study *RF15-3-peer*; for motivation, consider that it is typical for 3-5 students to assess a peer's work for peer assessment in online courses.

We now look at settings with more than two signals, and examine the recent extensions of DG13 to multi-signal settings. The strategy space quickly grows, so we cannot visualize the full basin of attraction in the same way. Instead, we first consider *T* along with two non-truthful strategies at a time, looking to develop qualitative understanding through representative examples. We will then adopt a quantitative metric, which estimates the basin size for more than three strategies by sampling.

First we compare W3a, a categorical three-signal model, and W3b, a non-categorical model, showing dynamics for *merge01, merge12,* and *T* (Figure 3.5). For W3a, the CA and DGMS mech-

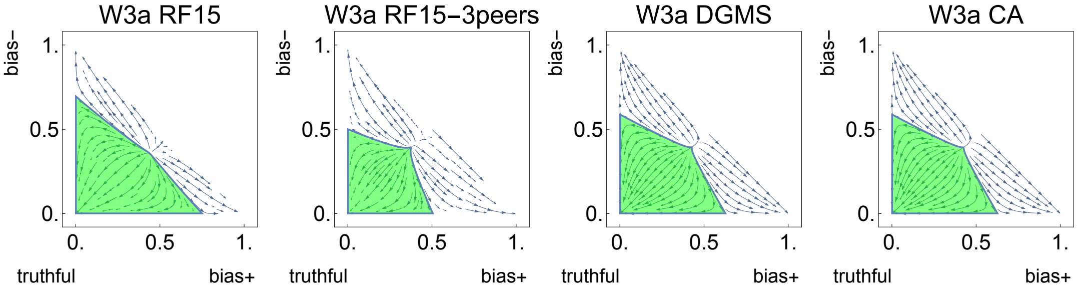


Figure 3.6: Flow plots for W3a, now using the *bias+* and *bias-* strategies. The difference in the basins of attraction compared to the top row of Figure 3.5 shows the limitations of two-dimensional flow plots in a many-strategy setting.

anisms are identical, and both converge to truthfulness from a large set of starting values. For W3b, *merge01* has higher payoff than *T* under DGMS, and the dynamics converge to *merge01* from almost the whole space.[7] Figure 3.6 parallels the W3a plots just discussed, but now adopting different strategies. Here, the basins of attraction for truthfulness are smaller. This illustrates the need to examine many combinations of strategies to understand a mechanism's behavior.

We now compare CA with the other two informed-truthful mechanisms we introduced earlier: Kamble and RPTS. Figure 3.7 confirms that as expected, constant strategies are unattractive for all three mechanisms. Figure 3.8 and Figure 3.9 look at the merging and bias strategies, and reveal some differences. Overall, truthfulness appears quite robust. Kamble and RPTS are even more resistant than to merging for worlds like W3a. On the other hand, in worlds like W3b, they are less resistant to biased strategies, *bias-* in this case. Both Kamble and RPTS depend on report frequencies, The more detailed analysis in the next section will show some differences when these are compared in a larger dataset. As a preview, recall that RPTS guarantees about truthfulness require that the world satisfy a self-predicting condition.

[7] CA gives equal payoff for *T* and *merge01*. The tiny boost to truthfulness described in Section 3.1 breaks the tie toward the truthful corner.

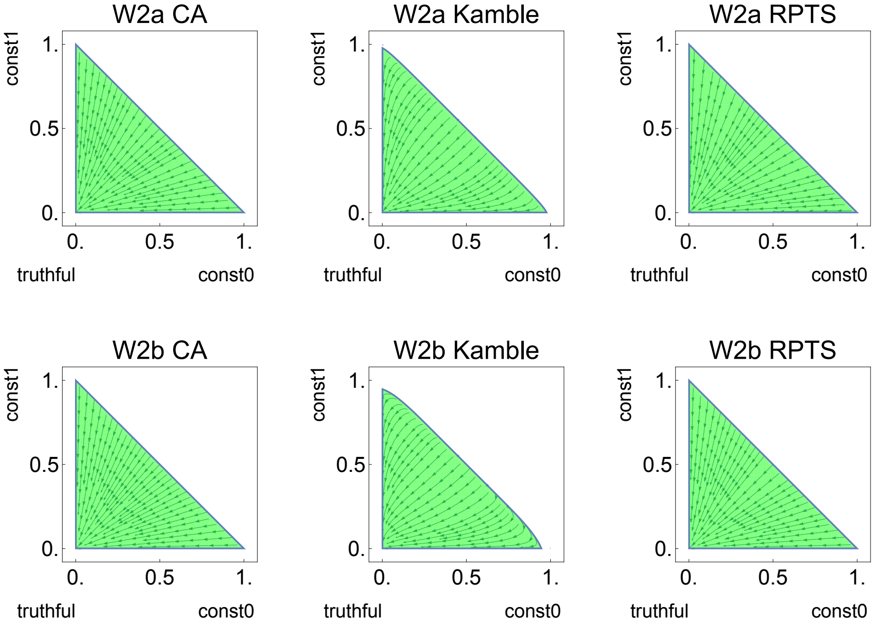


Figure 3.7: [Flow plots for W2a and W2b, comparing CA, Kamble, and RPTS with constant strategies.]Flow plots for W2a and W2b, comparing CA, Kamble, and RPTS with constant strategies. Truthfulness is always more attractive than constant strategies.

3.3 Peer Assessment in MOOCs

Our qualitative analysis suggests that the RF15, RPTS, Kamble, and CA are robust across a range of strategies and models, while non-truthful strategies can be attractors for OA, MRZ, and JF09. We now examine these patterns quantitatively on realistic world models. We study 325,523 peer assessments from 17 courses from a major MOOC platform. These comprise 104 questions, each with a minimum of 100 evaluations. There are 9, 67, 25, and 3 questions with 2, 3, 4, and 5 signals, respectively. We use maximum likelihood estimation to generate a probabilistic model, $\Pr(h)$ and $\Pr(s|h)$, for each question.

We base our model fit on student reports, not the unobservable signals, which are not available.

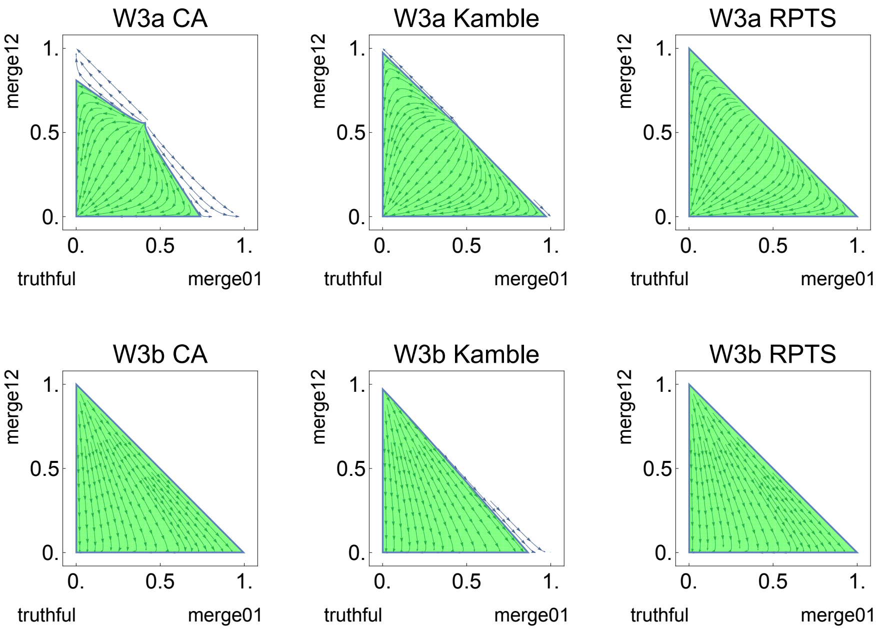


Figure 3.8: Flow plots for W3a and W3b, comparing CA, Kamble, and RPTS with merging strategies.

For the current work, in the absence of better data sets, we will simply stipulate that these are representative of true world models. This gives us a set of observed, non-hand-selected distributions, and provides a systematic way to compare the performance of the various mechanisms. Our analysis remains robust as long as the observed reports do not vary too much from the true signals learners would get if they all invested effort. We believe that as MOOCs start to provide valuable credentials based on peer-assessed work, there will be more incentive to cheat, and this condition may no longer hold without explicit credit mechanisms for peer assessment.

To ensure that our earlier observations were not specific to the particular strategies chosen for each plot, we look at dynamics with many strategies at once. For one more qualitative example, see Figure 3.10, which shows an example for W3b and CA, now with eight strategies. Despite the small fraction of the population starting out truthful, the dynamics converge to the truthful

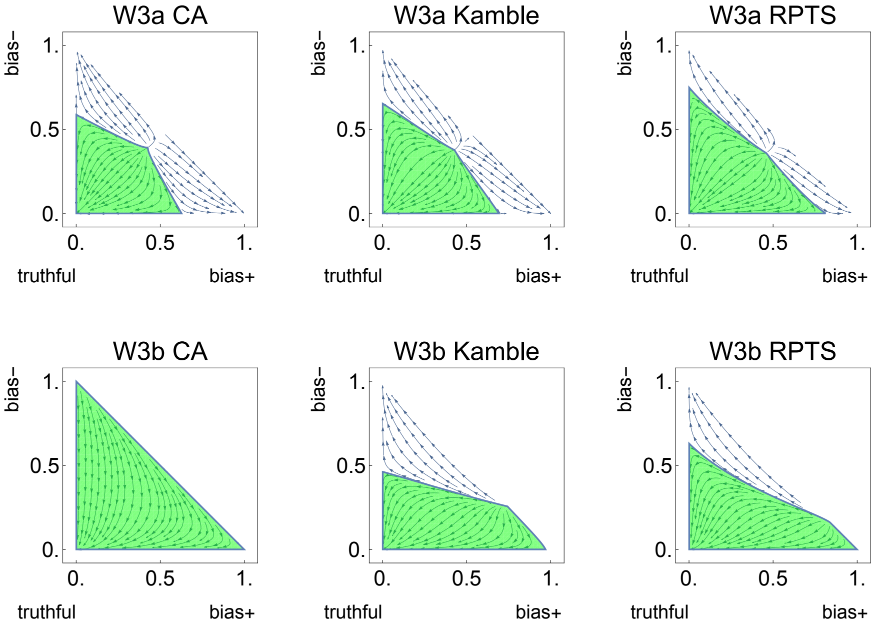


Figure 3.9: Flow plots for W3a and W3b, comparing CA, Kamble, and RPTS with bias+ and bias- strategies.

equilibrium.

To quantitatively compare the mechanisms, we estimate the size of the basin of attraction of truthfulness for each question and mechanism pair: we choose 100 starting strategy profiles uniformly at random in the strategy simplex, and measure the percentage for which the dynamics converge to truthful. We exclude JF09 because it is only defined for binary signals while the MOOC models have up to five signals. For each model, we use the corresponding strategy set from Section 3.1.[8]

This gives us a distribution of 104 basin sizes for each mechanism, shown as box plots in Figure 3.11. DGMS basin sizes span a large range because many of the estimated models are non-

[8] Due to computational limitations in simulating RF15-3-peer, we do not include the full strategy set in its analysis, using only *const0,const1,mergeAdj,bias-,T*. Our comparison thus favors RF15-3-peer, as other potentially attractive strategies are excluded.

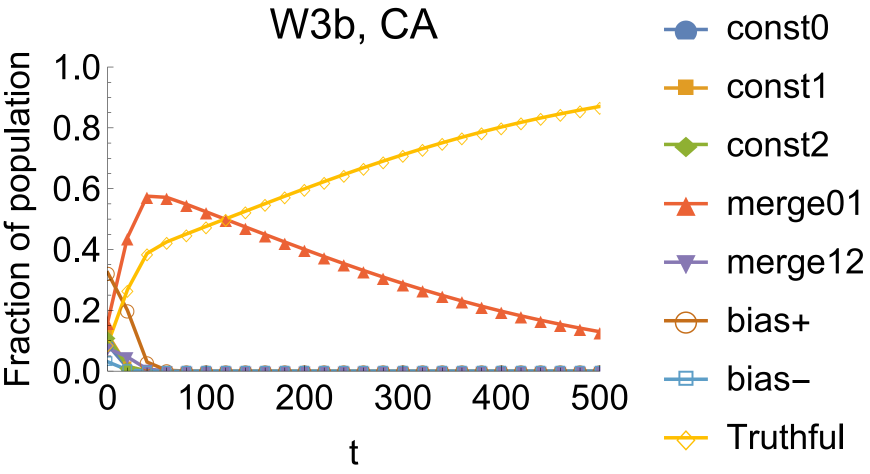


Figure 3.10: Dynamics with many strategies. We cannot easily visualize the many-dimensional simplex, but can sample to estimate the size of the basin of attraction of an equilibrium.

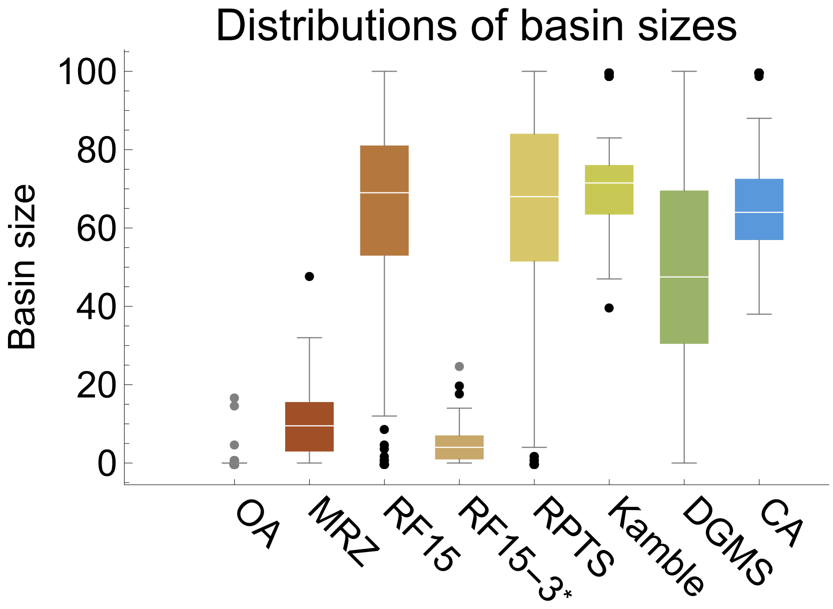


Figure 3.11: For each mechanism on the horizontal axis, and each of 104 MOOC-based world models, we estimate the size of the basin of attraction of T (as a fraction of the full space, on the vertical axis) for that world. The result is a distribution of 104 basin sizes for each mechanism, which we illustrate with a box plot above each mechanism name. The results match our earlier qualitative analyses.

categorical. RPTS has similar variation because many models are not self-predicting. RF15 is fairly good, but recall that RF15 is defined in the limit as the number of peers per task grows large, and is thus not a good fit for this domain. RF15-3-peer, which is a better match for the domain, does not do as well. The CA and Kamble mechanisms have the most robust performance, and appear promising in terms of their ability to robustly promote the convergence of population learning strategies to informed, truthful play.

3.4 Conclusions

We show that replicator dynamics are a good complement to equilibrium analysis and experiments for studying peer prediction mechanisms. We confirm that single-task mechanisms such as OA, MRZ, and JF09 can be very unstable with a learning population, even when being truthful is an equilibrium.We show that the newer, multi-task mechanisms (DG13, Kamble, RPTS, DGMS and CA) are much better at avoiding uninformative equilibria. We also support the need for large peer-group sizes with RF15, a point already made in (Radanovic & Faltings, 2015a).

In an analysis of models estimated from real peer assessment data, we show that Kamble and CA are promising candidates for real applications. Given over 100 distributions from this peer assessment data, we can have some confidence that our findings will generalize; although distributions in other application domains will differ, the size of the differences between the mechanisms suggest that our qualitative findings will be robust. Our results here do not give a reason to prefer one over the other. Chapter 4 compares these mechanisms using additional desiderata like fairness and score variability, in the peer assessment setting, and shows some differences–CA has lower variability, while Kamble only requires a single reference report for a particular report, not both a bonus and penalty reports.

Our analysis can be extended in various directions. We assume the same world model over many rounds of learning, which may not apply if the types of tasks change over time (imagine different homework assignments during a course). In addition, the replicator dynamics model ignores variance and small-population effects. Also of interest is the behavior of peer-prediction mechanisms with more complex models of human learning from behavioral economics or cognitive neuroscience. Finally, there is a need to validate these results with real people, in the lab, in real online courses, or in other crowdsourcing applications.

4

Practical Peer Prediction for Peer Assessment

We study the crowdsourcing of information in applications where it is difficult or expensive to verify contributions. There are many possible settings, including reporting information about businesses to improve products such as Google Maps, assessing peer work in large-scale education, and eliciting emotional reactions to video content or images. The objective is to encourage individuals to invest effort and make reports that reflect their viewpoint, even when this may be a minority viewpoint. Given reports, algorithmic methods can be used to aggregate the information in different ways.

The paradigm of *peer prediction* adopts explicit rewards to promote effort and truthful reports. In the absence of gold standard answers and the ability to verify reports, these rewards are determined based on comparisons between reports from different participants. Peer prediction has been studied for more than a decade, and there are now a number of mechanisms that have attractive theoretical properties—needing minimal information to operate, having broad domains of applicability, placing low reporting burden on participants, and avoiding undesirable "groupthink" style equilibria.

However, as far as we know, peer prediction has not yet been deployed in any large-scale application.[1] Peer assessment in MOOCs, in particular, is an exciting application domain for peer prediction—done well, it can help enable low-cost and thus broadly accessible education in subjects that are difficult to automatically assess today, such as writing, design, and public speaking.

Previous work on peer prediction has focused on the design of mechanisms that are *proper* (truthfulness is an equilibrium) and *strong truthful* (truthfulness is the equilibrium with highest score). We study several previously unexplored mechanism properties that matter for practical deployment. First, the magnitude of the benefit of exerting effort and being truthful over uninformed strategies is important: in educational settings, scaling the scores arbitrarily to increase the relative value of effort is impossible within a fixed grade range. Second, participants may be risk averse, and prefer strategies with relatively lower expected scores if they are more certain. Last, it is important to strive for fairness: specifically, participants who perform equally well at evaluating their peers should be rewarded equally. These concerns are of interest in education applications and beyond.

[1] There is some empirical work on peer prediction: Gao et al. (2014) study equilibrium selection in a simple binary setting, showing that agents find collusive equilibria. In contrast, Faltings et al. (2014) study a many-signal setting where uninformed equilibria did not appear to be a problem. We are not aware of any systematic studies of peer prediction in MOOCs, though Radanovic et al. (2016) present some initial positive experimental results from an on-campus experiment.

As a step toward deployment, we evaluate four candidate peer prediction mechanisms on a dataset of three million peer assessments from the edX MOOC platform. The comparison mechanisms include the classic output agreement mechanism as well as more recent designs from Chapter 2 and Kamble et al. (2015) and Radanovic et al. (2016). Our analysis is not experimental—we take existing peer assessments from a system that does not evaluate scorers, and compute what peer prediction mechanisms would do given these reports. Our key results:

- The benefit from exerting effort in evaluating peers is relatively low across all candidate mechanisms, due to relatively low agreement between peer scores.
- The *correlated agreement mechanism* (Chapter 2) has lower reward variation than other candidate mechanisms, because it rewards reports even without exact agreement between peers.
- The low peer agreement in our data set makes all mechanisms susceptible to student coordination on easy-to-see but unintended signals, as described by Gao et al. (2016).

In all cases, increasing agreement between peers would make peer prediction more practical. This can be accomplished by rubric design and student training, as well as using peer prediction itself to encourage effort.

4.1 Peer Assessment in Education

Peer assessment has a long history in education (see e.g. Goldfinch & Raeside (1990) and Falchikov (1995)) and is part of the much broader field of peer learning, which includes many types of peer-to-peer interaction in formal and informal settings.

For readers unfamiliar with peer assessment, we briefly summarize some lessons from its use in the classroom, to give a broader context for our incentive-focused study. There are two primary concerns about the scores given in peer assessments. The first is *reliability*, whether peers agree with each other. If not, ratings will have high variance, and many graders will be needed for each assignment to get a good estimate. The second is *validity*, whether the average peer score is “right” (Cho et al., 2006). In the typical situation where there is no absolute notion of right, it is typical to compare with instructor grades.

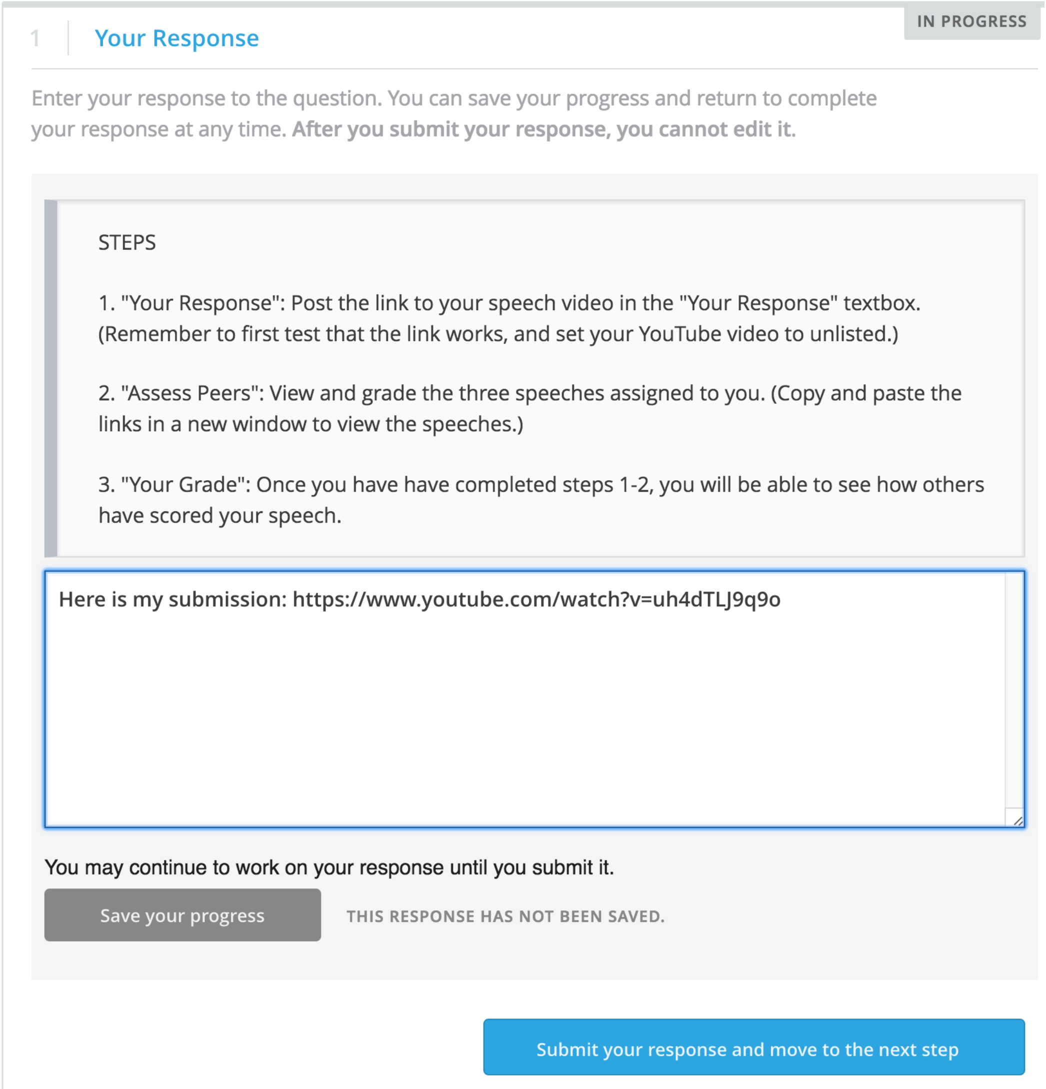
Calibrated peer review (Russell, 2004) helps improve validity and reliability: before students assess each other, they practice grading three instructor-created samples of varying quality until they give the right grades. Good rubric design is also critical to reliability. Orsmond & Merry (1996) note that objective evaluation criteria are easier to assess, especially if the rater does not need to be an expert in the subject to distinguish between the possibilities.

In the last several years, peer assessment has been deployed in massive online courses at much larger scales than before. As a concrete example, Figure 4.1 shows a screenshot of the edX peer assessment system. Students submit their responses, and are paired randomly for review.

Research in large scale peer assessment has focused primarily on evaluating students’ skill at assessment and compensating for grader bias (Piech et al., 2013), as well as helping students self-adjust for bias and provide better feedback (Kulkarni et al., 2013). Piech et al. (2013) test several models of student bias and reliability, testing for temporal coherence in bias as well as correlation between high scoring and being more reliable as a grader. Kulkarni et al. (2013) compare peer and staff grading, and find that the median peer grades are quite close to staff grades.

Other recent studies focus on other aspects of peer assessment. PeerStudio (Kulkarni et al., 2015) improves learning by ensuring fast feedback in large scale peer assessment. The *Mechanical TA* (Wright & Leyton-Brown, 2015) study focuses on reducing TA workload in high-stakes peer grading by reducing the need to spot-check peer grades.

Outside of peer assessment, many behavioral economics studies have shown that expected reward is not all-important in determining how people behave in practice (see Erev & Roth (2014) for a review). In our study, we are particularly motivated by *risk aversion*, which causes people to



Your Response

Enter your response to the question. You can save your progress and return to complete your response at any time. **After you submit your response, you cannot edit it.**

STEPS

1. "Your Response": Post the link to your speech video in the "Your Response" textbox. (Remember to first test that the link works, and set your YouTube video to unlisted.)

2. "Assess Peers": View and grade the three speeches assigned to you. (Copy and paste the links in a new window to view the speeches.)

3. "Your Grade": Once you have have completed steps 1-2, you will be able to see how others have scored your speech.

Here is my submission: https://www.youtube.com/watch?v=uh4dTLJ9q9o

You may continue to work on your response until you submit it.

Save your progress
THIS RESPONSE HAS NOT BEEN SAVED.
Submit your response and move to the next step

Figure 4.1: Screenshot from the edX peer assessment system, for a public speaking assignment.

choose lower expected reward for more consistency, with high payoff variability leading to more random choices (Erev & Barron, 2005; Busemeyer & Townsend, 1993).

4.2 Peer Prediction Mechanisms

Peer prediction mechanisms are modelled as follows: agents are assigned to tasks, and observe a *signal* for each task that encodes the information the system wants to elicit. The signal model includes a *signal prior* $P(s)$, the probability that an agent observes signal s, as well as a *signal* joint $P(s, s')$, the probability that two agents who do the same task get signals s and s', respectively.

Agents report their signals, either truthfully as observed, or strategically to increase their expected score or avoid the effort of observing the signal precisely in the first place. Some mechanisms also require reporting information beyond the observed signal.

The mechanism compares reports, and computes a reward for each report. A basic goal is for the mechanism to be (strictly) *proper*, so that truthful reporting is a (strict) correlated equilibrium. We restrict our study to *minimal* mechanisms, which do not require any additional information beyond a signal report, as these are more practical. We only include *detail-free* mechanisms, where the reward computation does not depend on precise details of the probabilistic signal model.[2]

We compare the following mechanisms:

Output Agreement (OA) (von Ahn & Dabbish, 2004). For each report r, the system picks a reference report r' on the same task, and defines score $\sigma(r, r') = A(r, r')$, where $A(x, y)$ is the agreement function, defined to be 1 if $x = y$, and 0 otherwise. The OA mechanism is only strictly proper when the signal distribution is *self-dominant*, meaning that a user's observation is also the most likely observation for their reference peers.

[2] We omit the scoring-rule based mechanism of Miller et al. (2005), because it is not detail-free and not strong truthful, and non-minimal mechanisms (Prelec, 2004; Witkowski & Parkes, 2012a,b; Radanovic & Faltings, 2013). We also omit the minimal, strong truthful mechanism in Radanovic & Faltings (2015a), because it requires many more reports per task than are typical in peer assessment.

We include OA in our study because of its simplicity. However, it and other early peer prediction mechanisms allow agents to coordinate and get higher rewards by reporting untruthfully. In OA, all agents simply reporting the same thing each time guarantees maximal reward.[3]

The next two mechanisms use the empirically observed report prior and joint distributions. We denote these $\hat{P}(\cdot)$ and $\hat{P}(\cdot, \cdot)$, respectively.

Robust Peer Truth Serum (RPTS) (Radanovic et al., 2016). This is a version of OA in which scores are scaled based on observed report frequencies; the system collects all reports, computes the empirical prior $\hat{P}(r)$ of each report r, and defines score $\sigma(r, r') = A(r, r')/\hat{P}(r)$, where r' is a reference report, just as in OA. This results in higher scores for matches on uncommon reports, which has two benefits: the mechanism requires a weaker *self-predicting* condition on the signal model—seeing a signal should increase the likelihood peer agents observe the same signal—and constant reporting now has lower expected score than truthfulness. RPTS requires that the number of tasks is large enough to make the empirical prior accurate.

Kamble (Kamble et al., 2015). This is another scaled version of OA. The system collects all reports, computes the empirical joint $\hat{P}(r, r')$, and defines score

$$\sigma(r, r') = A(r, r')/\sqrt{\hat{P}(r, r)},$$

or 0 if $\hat{P}(r, r)$ is exactly 0 or 1. Similarly to RPTS, constant reporting again has lower expected score than truthfulness. Additionally, the mechanism is proper for general signal distributions.

Correlated Agreement (CA) (Chapter 2). The CA mechanism is *multi-task*, so each agent reports on several tasks (at least two). An agent is rewarded for being more likely to match reports

[3] Jurca & Faltings (2009) attempted to fix this by rewarding near-agreement, not perfect agreement, with several peers. Dasgupta & Ghosh (2013) went further to design the first mechanism that guaranteed *strong truthfulness*, where the truthful equilibrium has higher payoff than all other equilibria, in settings with binary reports. Peer assessment uses non-binary reports, so we study several newer mechanisms that provide similar guarantees with arbitrary numbers of signals.

of peers doing the same task than the reports of peers doing other tasks. Let r_i^k denote the report received from agent i on task k. The mechanism is described, w.l.o.g., for two agents, 1 and 2:

1. Assign the agents to three or more tasks, with each agent to two or more tasks, including at least one overlapping task. Let M_s, M_1, and M_2 denote the shared, agent-1 and agent-2 tasks, respectively.

2. The score for a shared task $k \in M_s$ to each agent is

$$\sigma_k = \Lambda(r_1^k, r_2^k) - \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \Lambda(i,j) \cdot h_{1,i} \cdot h_{2,j}, \tag{4.1}$$

where $\Lambda : \{0, \ldots, n-1\} \times \{0, \ldots, n-1\} \to \mathbb{R}$ is a score matrix with $\Lambda(s, s') = 1$ if $P(s, s') > P(s)P(s')$, and 0 otherwise, $h_{1,i} = \frac{|\{\ell \in M_1 | r_1^\ell = i\}|}{|M_1|}$ is the empirical frequency with which agent 1 reports signal i in tasks in set M_1, and $h_{2,j} = \frac{|\{\ell \in M_2 | r_2^\ell = j\}|}{|M_2|}$ is the empirical frequency with which agent 2 reports signal j in tasks in set M_2. This definition for Λ rewards agreement on positively correlated pairs of signals.

3. The total score to an agent is the sum of the score across all shared tasks.

CA is proper (not strictly), and *informed truthful*: the payoff for all agents being truthful is weakly higher than any other strategy profile, and strictly higher than any uninformed (signal-independent) reporting strategy. CA works with small numbers of tasks if the designer knows the direction of correlation between pairs of signals, needed to define Λ, or can learn these correlations from agent reports when there are many tasks.

4.2.1 Scaling Scores

To be practical for peer assessment, a mechanism's scores must be positive, and have bounded range—like any other grade, course teams need a way to say that assessing peers on an assignment counts for a particular number of points, and it is unreasonable to tell students that they may get an unboundedly high score with a very small probability, compensating for much more likely low scores.[4]

We set the scores for all mechanisms to be in $[0, 1]$ to make comparisons consistent. For RPTS and the Kamble mechanism, we do this by "clamping"—imposing a minimum on the report prior $\hat{P}(r)$ and the joint factor $\sqrt{\hat{P}(r, r)}$, respectively, and scaling to ensure the resulting score is in $[0, 1]$. We choose the minimum value to balance between effective score range and frequency of clamping in our dataset—whenever clamping applies, it breaks the mechanism's theoretical guarantees, effectively underpaying for unlikely reports.[5] An undesirable side-effect of this adjustment is that typical reports, with high priors by definition, will only use a small fraction of the score range, and only unlikely reports with prior close to the minimum will get scores close to 1.

For the CA mechanism, we remapped scores from the base range of $[-1, 1]$ into $[0, 1]$. The effect is that the expected score for uninformed reporting is 0.5, regardless of the reports of other learners.

All the mechanisms use a single reference peer as described, and can be modified to give the average score over several such peers. For example, for OA, given a set of reference reports $r_1, \ldots, r_n$ from n different peers on the same task, the mechanism could instead give score

$$\sigma(r) = \frac{1}{n} \sum_{i=1}^{n} A(r, r_i), \tag{4.2}$$

and similarly for the other mechanisms. Choosing a random reference peer or averaging across

[4] The bounded range means that the standard theoretical trick of linearly scaling payoffs until the difference between truthful reporting and other strategies is big enough is not viable.

[5] For RPTS, we make the minimal prior value 0.1, and divide scores by 10. This makes the expected score for uninformed reporting is 0.1. For Kamble, we make the minimum value of $\sqrt{\hat{P}(r, r)}$ 0.25 and divide scores by 4. These values balance between clamping too often and having the typical scores use a significant fraction of the $[0, 1]$ range.

Category	Example	Count
Courses	*"Eating, Then and Now"*	254
Submission prompts	*"What is food?"*	682
Evaluation criteria	*"Correct grammar"*	1983
Submissions	*"Cheese is the best food"*	354312
Peer assessments	*3/5 points*	3090452

Table 4.1: Dataset summary, with examples of each item. There are approximately 1500 assessments for an average evaluation criterion.

all reference peers has no effect on the expected score of a mechanism, but does affect the score variability. We study both variants below in Section 4.6.

4.3 The edX Dataset

The dataset in our study consists of peer assessments from edX, a site that offers open online courses, and includes data from 2014-2016.[6] Each peer assessment is a tuple

```
(course, item, submitter, submission,
submission_time, scorer, criterion,
points),
```

corresponding to a scorer assessing the given submission along a particular evaluation criterion, and giving it a score.

As a preprocessing pass, we keep only the latest evaluation for each `(scorer, submission)` pair, and discard criteria with fewer than 100 assessments. This leaves just over three million assessments, across about 2000 evaluation criteria, in 254 courses. Table 4.1 shows summary counts. Most courses in our dataset only used peer assessment one or

[6] A summarized dataset of the joint report probabilities for each of the 1983 evaluation criteria will be published with this paper. The full dataset of individual students' assessments is sensitive, and cannot be shared.

two times; a few courses had weekly or bi-weekly peer assessment assignments. Figure 4.2 shows the full distribution of the number of prompts per course.

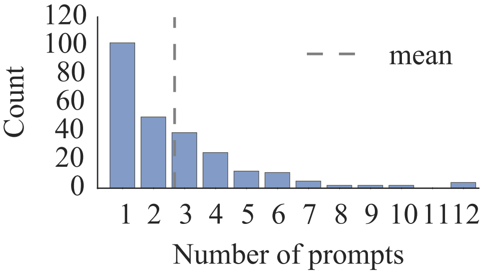


Figure 4.2: Prompts per course.

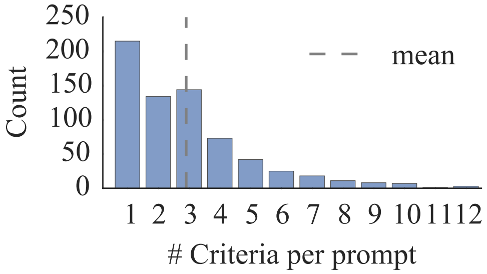


Figure 4.3: Histogram of evaluation criteria per prompt.

Submissions for each prompt are assessed on several evaluation criteria. For example, a short essay in a writing class may be judged on four criteria: grammar, style, argument, and appropriate citations. Figure 4.3 shows the number of evaluation criteria per prompt. Most prompts have four or fewer evaluation criteria. For each evaluation criterion, students can select a point value corresponding to a particular rubric option (e.g. 5/5, "Perfect grammar."), and each criterion induces a separate, empirical signal distribution.

Many evaluation criteria have several hundred assessments (median 733), with a few from large courses having ten thousand or more (Figure 4.4). The mean is about 1500.

Some peer prediction mechanisms rely on each user doing multiple tasks. The system was not explicitly set up to ensure this, but course teams can require that learners do a minimum number of peer evaluations before they see their peers' evaluations of their own work. The typical recommended values are 3 or 5, and this shows in distribution of submissions assessed by each scorer in Figure 4.5. Some course teams choose a lower number, and some students stop before they finish the assignment, so a practical system would need to handle these cases, likely with a low default score. On the other end of the distribution, learners are permitted to assess more than the

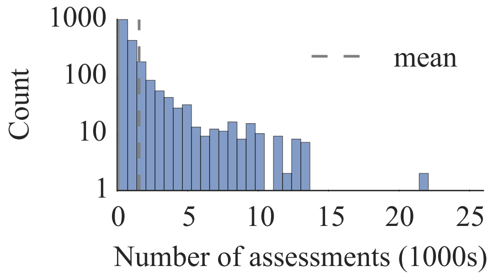


Figure 4.4: Histogram of the number of assessments across evaluation criteria. Note the log scale: a few evaluation criteria have more than 20,000 assessments, while the majority have less than 1500.

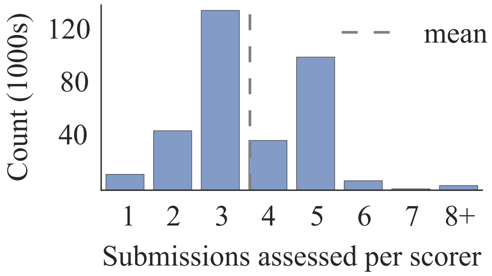


Figure 4.5: Submissions per scorer.

minimum number of peers, and a small fraction do so.

4.3.1 Probabilistic Models for Reports

We now explore the details of the assessments for different prompts, looking at the number of options (i.e., the number of possible *signals*) for different evaluation criteria, the probabilities of those signals, and the correlation structures between them.

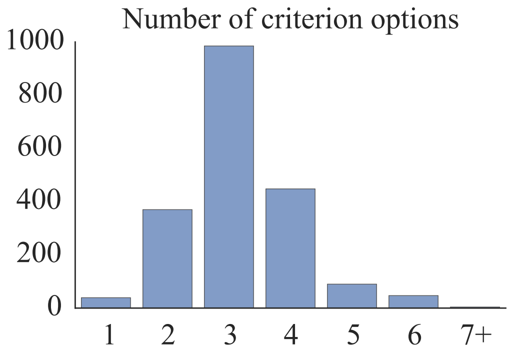


Figure 4.6: Histogram of models by number of criterion options (possible signal values).

Figure 4.6 shows the distribution of the number of distinct options per criterion. An initially surprising observation: there are several criteria where students had one option in "assessing" their peer. The explanation is creative course teams using the peer assessment tool for open ended peer feedback, without wanting numerical assessment. We will ignore these criteria in our analysis. On the other extreme, there was a course team that specified 21 score options for a criterion. Going forward, we focus on just models with two to six score options, since they account for the vast majority of the data.

Our dataset contains student reports, not the true signals observed by students. This is the best we can do without a prohibitive amount of manual grading, and we assume that the reports are a noisy approximation of the true signals. Since students are participating in a free class without much outside incentive for completion, doing the evaluation at all is indicative of exerting some effort.[7] From here on, we use *report* and *signal* interchangeably, unless explicitly distinguished.

We call a given signal distribution, corresponding to an evaluation criterion in the dataset, a *model*. Let $P(s)$ represent the prior probability of an agent seeing signal s on an arbitrary task. Let $P(s, s')$ denote the joint probability that two agents will see signals s and s'. We are also interested in what the joint distribution would be if signals were independent but with the same prior. This is the product-of-marginals distribution, written $Q(s, s') = P(s)P(s')$. Finally, we use $P(s'|s)$ to denote the signal posterior: the probability an agent observes s', conditioned on another agent observing s on the same task.

We look next at the signal priors $P(s)$, which are important to the design, applicability, and robustness of peer-prediction mechanisms. The prior for an evaluation criterion is the probability with which each score appears. Figure 4.7 shows a plot for each number of signals k, plotting all distributions of size k on one plot, along with the average values. There is significant variation,

[7] Nevertheless, as MOOCs start to provide credentials based on peer-assessed work, we believe it will become increasingly important to provide explicit credit mechanisms for peer assessment.

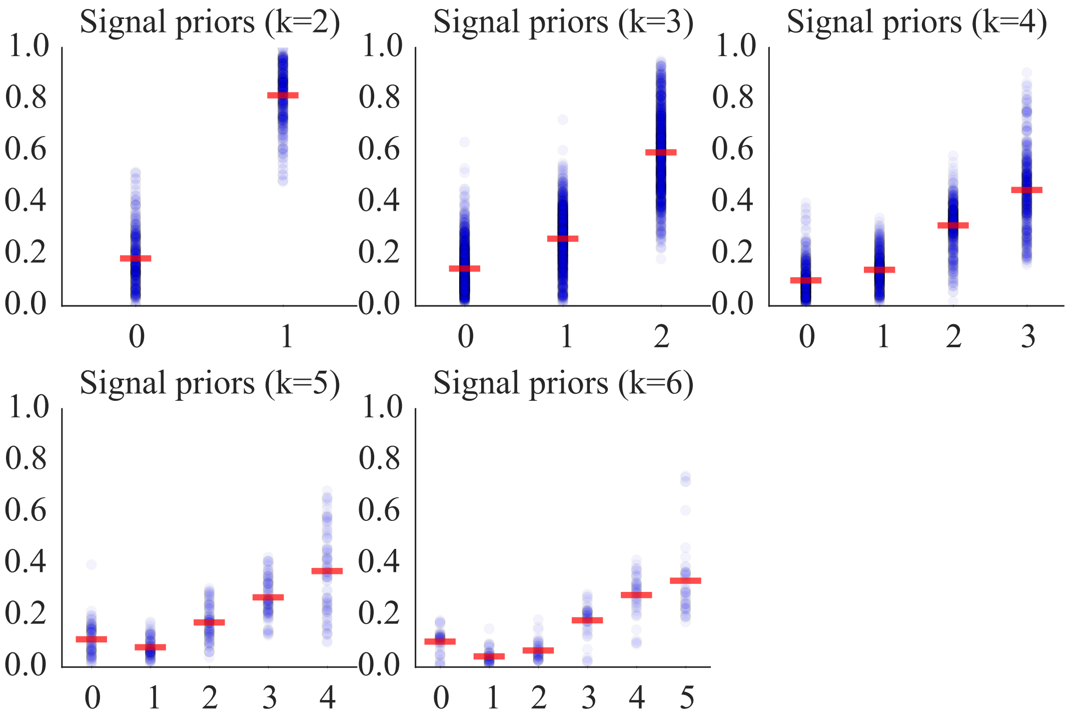


Figure 4.7: Signal priors by number of possible signal values, with all distributions for fixed number of signals on the same axis. The red line marker is the average for that signal. There is significant variation among models, with a clear trend toward higher scores.

but the priors are clearly non-uniform, with higher values more likely. An interesting secondary feature is that for $k \in \{5, 6\}$, non-zero scores below the median (1, and $\{1, 2\}$, respectively) tend to go unused. This suggests that most submissions are either very bad or incomplete, or ok-to-great, with few in between.[8]

An obvious question about a peer assessment system is whether peers usually agree. We will look at this in several ways. A summary metric is the probability that two random peers assessing the same submission will report the same assessment. This probability is 61% in our dataset. To give a baseline, the probability that two peers assessing random submissions to the same prompt

[8] It also suggests that course teams may be able to simplify their rubrics, giving fewer options without losing many meaningful distinctions.

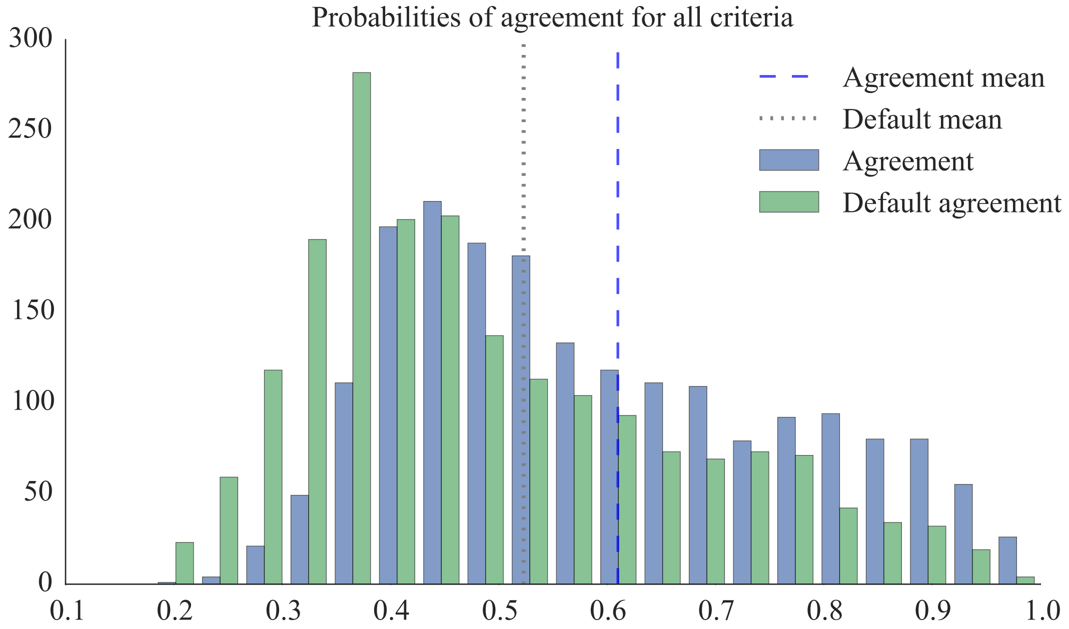


Figure 4.8: Histograms of observed and submission-independent "default" agreement between reports, per evaluation criterion. The means are weighted by the number of assessments for that criterion.

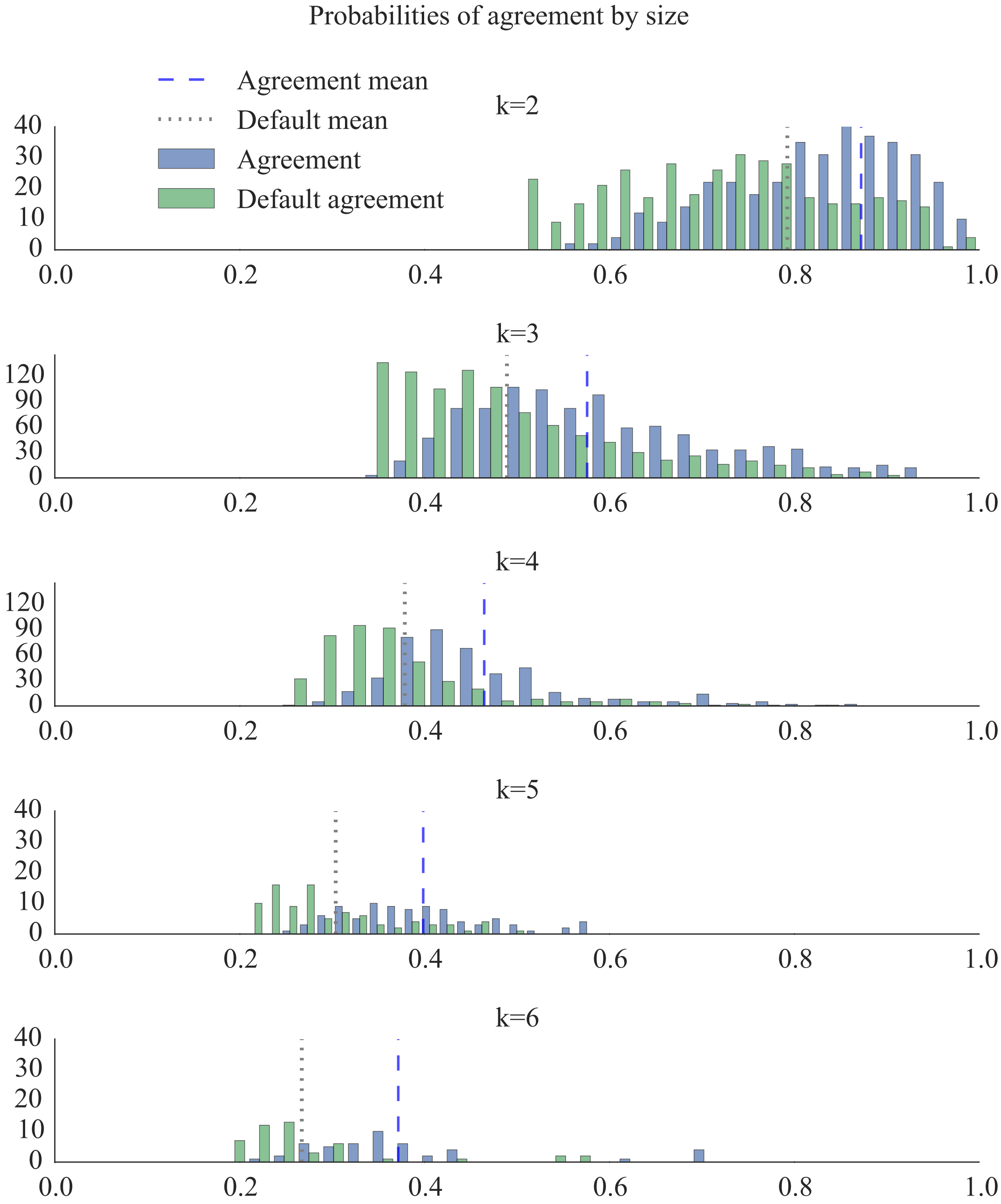


Figure 4.9: Histograms of observed and submission-independent “default” agreement between reports, per criterion, by size.

would agree is 52%. The high baseline probability makes sense in light of the non-uniform priors. Since many reports are of the highest possible signal (61% overall), two random assessments for different submissions frequently agree on that by chance. Figure 4.8 shows the distributions of observed agreement of same-submission reports and "default" agreement, if two reports for different submissions are chosen for a particular criterion. The vertical lines give the mean probabilities across the dataset, with criteria weighted by number of reports. Figure 4.9 further breaks down the distributions by size. As expected, the probability that two reports agree goes down as the number of possible reports goes up.

A natural follow-up question to the relatively low probability of agreement between reports is whether learners agree approximately, or not at all. An ad-hoc measure is the probability that two reports for a submission are within one of each other. This is a trivial condition for binary models. Per-criterion histograms for larger models, in Figure 4.10, show that approximate agreement is much more likely than exact agreement, and goes down slowly with model size.

Another way to look at agreement is to look at the correlation between pairs of signals. For this, define the *Delta matrix*[9] Δ , an $n \times n$ matrix, with entry (i, j) defined as

$$\Delta_{s,s'} = P(s, s') - P(s)P(s'), \tag{4.3}$$

or equivalently as the difference between the joint and product-of-marginals distributions:

$$\Delta = P(\cdot, \cdot) - Q(\cdot, \cdot) \tag{4.4}$$

The delta matrix encodes the correlation (positive or negative) between different realized signals. The average values in this Delta matrix, grouped by prompts with the same number of reports, are shown in Figure 4.11, along with the sign structure which gives the direction of the

[9] Repeating a bit of background from Chapter 2 here to make this chapter more self-contained.

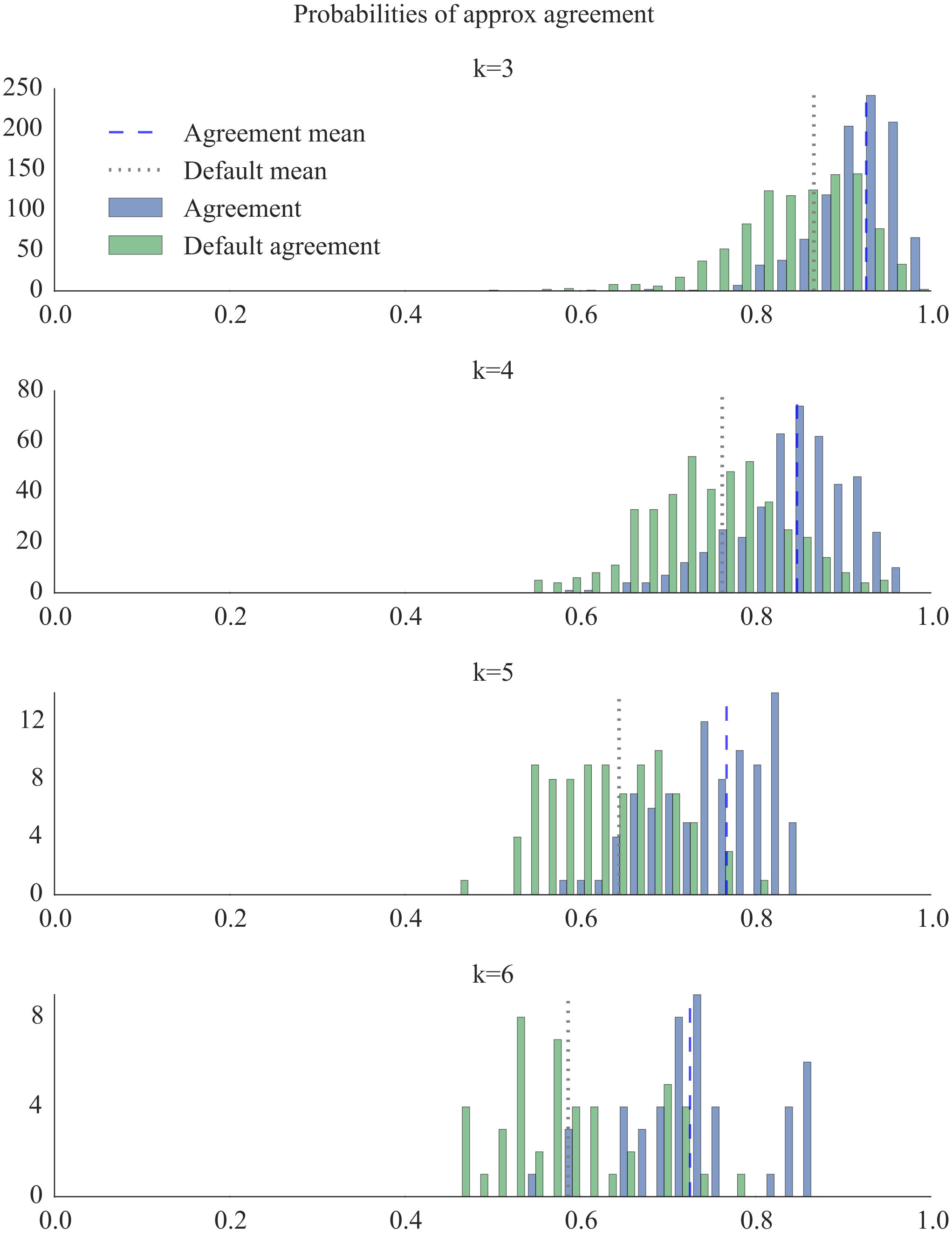


Figure 4.10: Probabilities of approximate agreement—two reports within 1 point of each other.

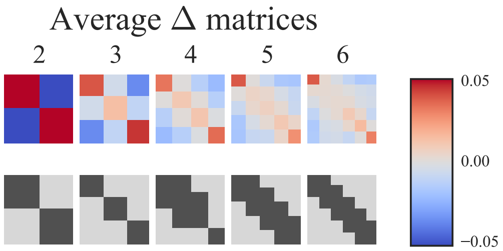


Figure 4.11: Average delta matrices and their sign structure. The positive areas along the diagonal correspond to the ordinal structure of peer assessment—nearby signals are likely to be positively correlated.

correlation. As expected in a setting where the signal values are ordered,[10] the correlations are positive along the diagonal—if one student thinks the right score is 3/5, it increases the likelihood that their peer will say 2, 3, or 4.

4.4 Analysis I: Appropriateness

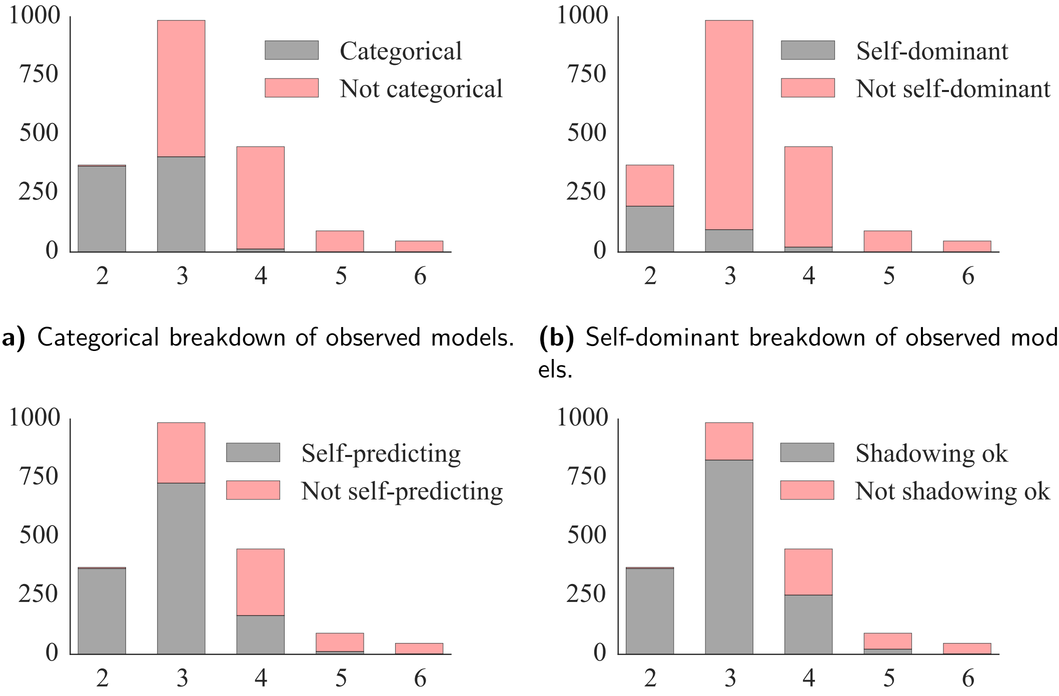
As a first analysis step, we look at how often the technical conditions required for the validity of different peer prediction mechanisms hold in our dataset. We start with the *categorical* condition needed for DGMS (See Chapter 4) to be strong truthful:

$$P(s'|s) < P(s') \quad \forall s' \neq s. \tag{4.5}$$

Here, seeing a signal makes all other signals less likely than their prior. Figure 4.12a shows the breakdown, by model size. As fits the ordinal setting in peer assessment, most non-binary models are not categorical: seeing a particular signal increases the probability of adjacent ones.

The next condition is the *self-dominant* condition, which is required for OA to have a truthful

[10] As opposed to an unordered classification setting like labeling images as cars, animals, or people.



(a) Categorical breakdown of observed models.

(b) Self-dominant breakdown of observed models.

(c) Self-predicting breakdown of observed models.

(d) Shadowing condition breakdown of observed models.

Figure 4.12: Breakdowns of models by the categorical condition, self-dominant and self-predicting conditions, needed to achieve truthfulness in DGMS, OA and RPTS, respectively.

	¬ SD	SD
¬ SP	656	11
SP	800	107

	¬ SD	SD
¬ S	465	7
S	991	111

	¬ SP	SP
¬ S	445	27
S	222	880

Figure 4.13: Contingency tables for pairs of conditions. This confirms that in our data, self-dominant (SD) is more restrictive than self-predicting (SP), which is more restrictive than the shadowing condition (S).

equilibrium:

$$P(s|s) > P(s'|s) \quad \forall s' \neq s. \tag{4.6}$$

A model is self-dominant if seeing a signal makes this signal the most likely signal for a peer. Figure 4.12b shows the breakdown. Most models do not satisfy this condition. An interesting observation is that it does not always hold even for binary models. This happens when one signal is much more likely than another: if an agent observes a very unlikely signal, she may still expect a peer to observe the more likely signal with probability more than 0.5.

Another condition is *self-predicting*, and is needed for the RPTS and related 1/prior mechanisms to have their intended properties:

$$P(s|s) > P(s|s') \quad \forall s' \neq s. \tag{4.7}$$

In words, an agent's peer is more likely to see a particular signal if the agent also sees that signal. Figure 4.12c shows the breakdown. This condition is weaker than the previous two, and holds for the majority of size three models, though not for most larger ones. This means that RPTS is manipulable in peer assessments with many options, though experiments would be needed to see whether students find the manipulations in practice.

For completeness, we also include a breakdown for an additive analog to self-predicting, called the *peer shadowing* condition, from Witkowski & Parkes (2012c):

$$P(s'|s) - P(s') < P(s|s) - P(s), \tag{4.8}$$

which says that when an agent sees signal s, the likelihood of a peer getting signal s' cannot go up more than the likelihood of the peer getting signal s. This appears to be the least restrictive condition for our dataset (Figure 4.12d), though it too does not hold for most larger models.

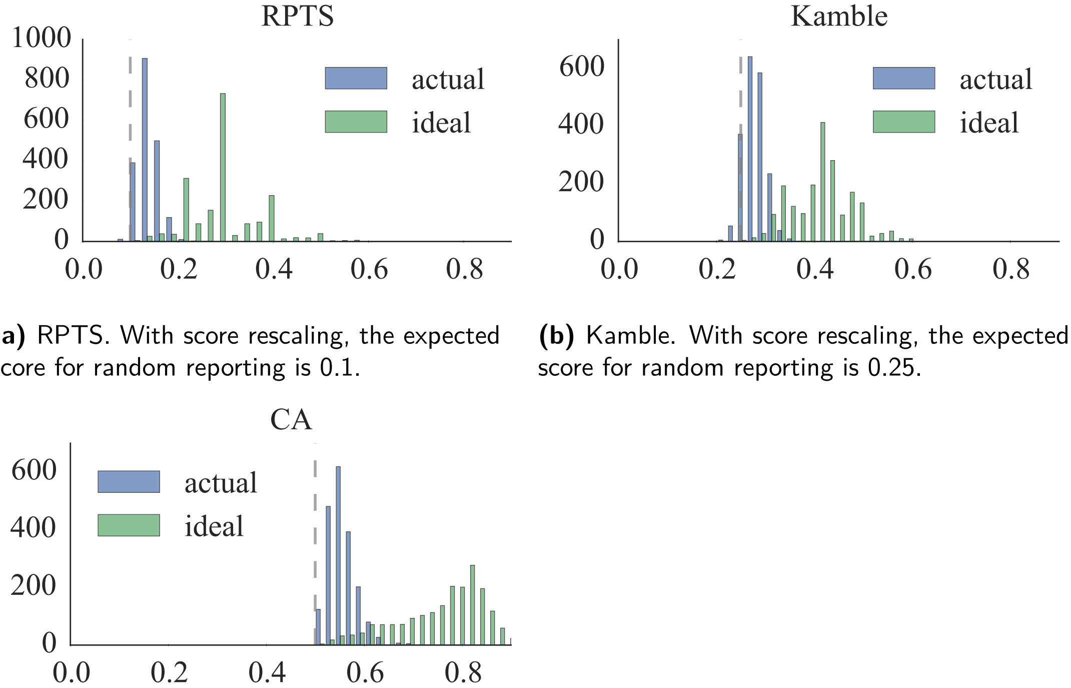
Figure 4.13 shows the relationships between pairs of conditions, confirming that in our dataset, if a model is self-dominant, that usually implies that it is self-predicting, which implies it likely satisfies the shadowing condition. The overall conclusion is that to guarantee truthfulness and strong truthfulness in this domain, we need mechanisms that do not place restrictions on the probabilistic signal models.

4.5 Analysis II: Expected Score

As a second analysis step, we examine the incentives for investing effort in doing a careful assessment of a peer's submission. For example, if a student can expect 50 out of 100 points by reporting randomly, and only 55 by carefully reviewing their peer, she may decide that the effort needed for careful review is not worth it. We omit OA from this analysis, because as discussed above, it is not strong truthful, and if enough students are willing to misreport, constant reporting will actually increase their scores.

We look at this numerically in Figure 4.14: the "actual" histograms show the expected scores for truthful reporting for all criteria. For all mechanisms, less than half the criteria have expected scores that are more than 0.05 above random reporting. The benefit to being honest is fairly small because of the noise in peer assessments.

To understand whether the small benefit from truthfulness relative to the working range of the scores is inherent (i.e., due to the non-uniform marginal distribution on reports) or due to the relatively low agreement between peers, we also plot the "ideal" expected scores that students would get in each mechanism, with the same signal prior but perfect agreement on reports. These are much higher, suggesting that there is an opportunity to address this problem by training students to peer assess more consistently; e.g., through better assessment rubrics, encouraging effort through schemes such as peer prediction, and through non-incentive-based methods (e.g.



(a) RPTS. With score rescaling, the expected score for random reporting is 0.1.

(b) Kamble. With score rescaling, the expected score for random reporting is 0.25.

(c) CA. The expected score for random reporting is 0.5.

Figure 4.14: Histograms for expected scores vs. ideal scores with perfect agreement on reports, per mechanism. For all mechanisms, the relatively low agreement between student reports makes expected scores for truthfulness only slightly better than for random reporting, when compared to the overal range of possible scores. "Ideal" scores, if peers agreed perfectly, are much higher, so there is a need to improve agreement.

adding “Please do a good job. Your peers depend on it!” to the instructions), and compensating for student bias, for example using the methods described in Piech et al. (2013). Another pragmatic workaround may be to clamp scores more severely in order to expand the working range of scores.[11]

Finally, some peer assessment exercises are simply not appropriate for peer prediction: if submissions are judged very subjectively (e.g. “do you like this art by your peer?”), it would be better to ask for peer feedback and reward participation rather than trying to reward accuracy.

4.6 Analysis III: Variability

Most of the theoretical analysis in the peer prediction literature focuses on expected value, and says that agents prefer one strategy to another if the former has higher expected payoff.[12] However, variability in scores is also likely to be important for several reasons. First, fairness is important, especially in education: two students who do work of equal quality should get the same score. A second reason is risk aversion: a student whose expected score is 5 points will be happier to always get 5 points rather than a 25% chance of 20 points, and might prefer a more certain strategy with lower expected score. Finally, students are likely to learn better with consistent feedback.

Risk aversion is concerned with overall score variability, while for fairness, variance in scores *ex ante*, before seeing task, is ok—different types of tasks may reasonably give different expected scores.[13] We are more concerned about variance in score given a signal—as a student, if I assess

[11] However, this should be done with caution because it would break the incentive guarantees. For example, with RPTS, if we increase the minimal allowed prior to 0.25, then when a student got a signal that was less likely than 0.25, she could want to misreport, giving a more common response instead.

[12] One partial exception is in Chapter 3, where we use replicator dynamics to model population learning rather than assuming equilibrium play based on expected rewards. The replicator dynamics evolve based on expected scores in a continuous population, so the core focus on expected score is still present.

[13] For example, in RPTS, unlikely reports have a higher expected score, so a student who gets a rare bad submission would expect more points than a students who gets a good submission.

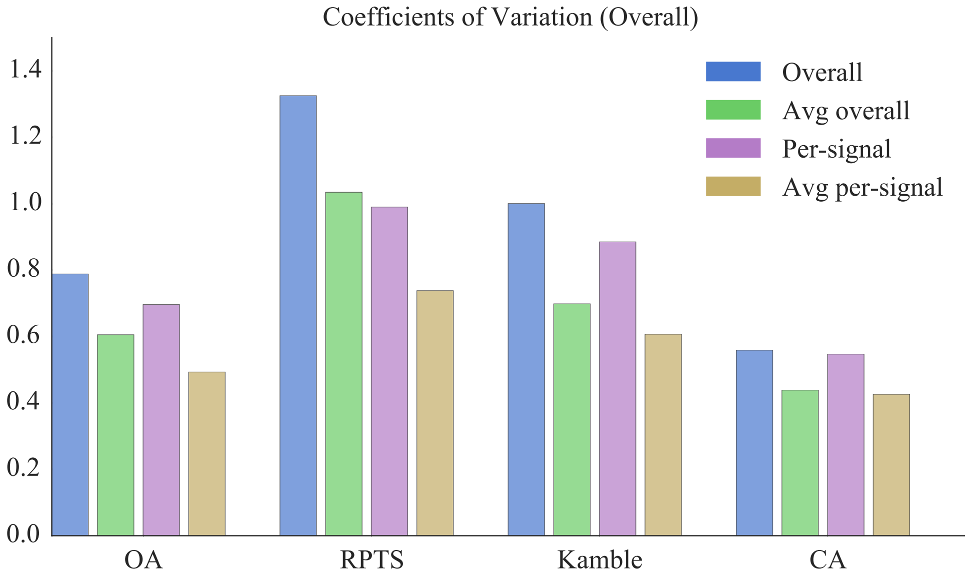


Figure 4.15: Overall coefficient of variation and expected per-report coefficient of variation of scores of different mechanisms, with and without averaging scores across all peers.

two very similar submissions, give each the same score, but get very different feedback, I may feel cheated.

Since the mechanisms that we study have different effective score ranges, variance is not a good metric for comparison. Instead, we use the *coefficient of variation*; i.e., the standard deviation divided by the mean. This is a standard way of comparing distributions with different scales.

Figure 4.15 shows the coefficient of variation for each mechanism, both overall and conditioned by signal. The signal-conditional value is the average of the individual coefficients of variation for each signal, weighted by the number of reports of that signal. In other words, it is the *a priori* expected coefficient of variation, before receiving a signal. Averaging scores for all peers always reduces the coefficient of variation, and the CA mechanism has the lowest overall variation.[14]

[14] RPTS has a bigger drop in variation from overall to per-signal, which makes sense because it uses different score ranges for different signals, so the overall distribution has high variance.

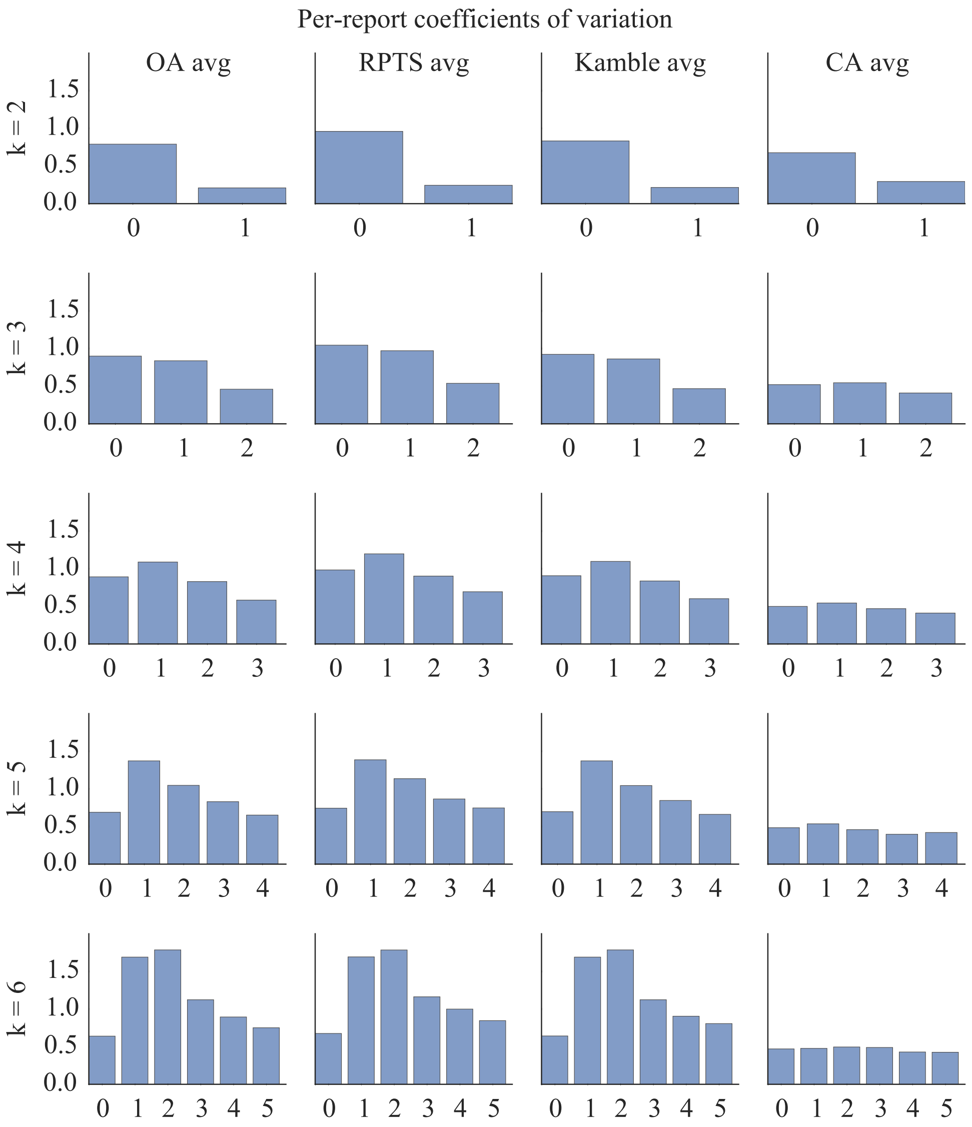


Figure 4.16: Per-report coefficients of variation for all mechanisms, averaging scores for all peers.

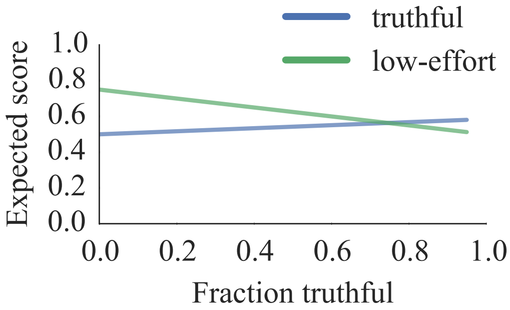


Figure 4.17: Example truthful vs low-effort scores using the CA mechanism, for a single criterion.

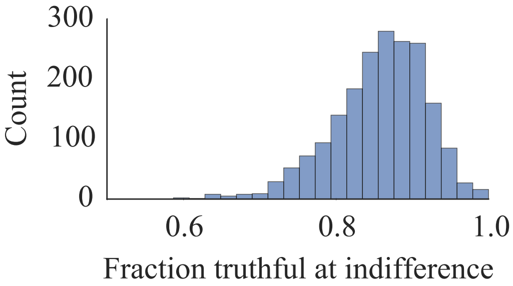


Figure 4.18: Fraction truthful at indifference across evaluation criteria.

Figure 4.16 shows the details behind the averaged, per-signal bars in Figure 4.15. OA, RPTS, and Kamble all show similar patterns, because they are all based on output agreement and only adjust the relative payoff for each signal. The coefficient of variation for each *k* and signal is roughly inverse to the frequency of that report (recall Figure 4.7)— unlikely reports match less frequently, so get more varied scores. The coefficients of variation for CA are much more uniform across signals, because it rewards agreement between correlated signals as well as exact agreement.

Overall, CA appears to be better than the others candidate mechanisms in terms of variance. Unlike CA, RPTS, Kamble, and OA all rely on exact agreement, and so effectively work based on the difference between the diagonals of the joint and product-of-marginals distributions. As the number of signals goes up, there is less and less probability of two signals being exactly equal, making these mechanisms more fragile, relying on rare rewards for their guarantees, and increasing variance.

4.7 Analysis IV: Risk of Collusion

We now look at another potential problem with peer prediction. As Gao et al. (2016) point out, students can potentially correlate on a *low-effort signal*, based on unintended and easy-to-observe properties of a submission such as length, id number, title, and so forth, thus matching without exerting effort. The suggested solution in Gao et al. (2016) is to give up on peer prediction entirely, and use trusted TAs to spot check student evaluations. While that is certainly effective when TAs are available, we are working in a model without many TAs, and look instead at the limits of what is possible under this kind of collusion. In particular, assuming that peer assignment is done randomly, we examine what fraction of the students needs to collude to benefit. It is likely in a large class that agreeing on such a correlation scheme would only be done by a fraction of the students.[15]

We focus on the CA mechanism here as the most promising candidate given the reward variance results above,[16] and assume a uniform distribution of low-effort signals, as in Gao et al. (2016). Figure 4.17 shows the expected CA scores for a particular evaluation criterion chosen as an example, as the fraction of the population that is truthful varies, with the rest assumed to collude on a perfectly correlated low-effort signal. As expected given the ideal vs. actual score histograms in Figure 4.14c above, scores for the perfectly correlated low-effort signal are much higher than truthful scores. A large fraction of the population must be truthful to get better scores than a subpopulation with perfect correlation. The intersection of the lines is the indifference point.

Figure 4.18 shows a histogram of the points of indifference for all the evaluation criteria. The

[15] We also note that there are reasons to expect collusion to be difficult in practice: students typically submit written feedback, not just score, and so still have to look at submissions. Students can complain if they get unfair evaluations, and students who are obviously cheating can be punished. Similarly, even a low percentage of spot checking can discourage cheating if the punishment is substantial.

[16] The results for Kamble and RPTS are similar.

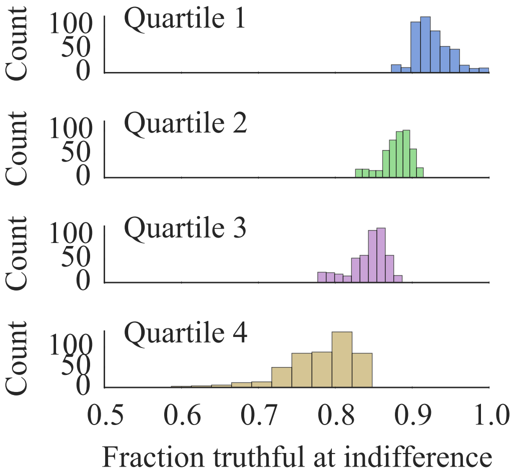


Figure 4.19: Histograms of the fraction truthful at indifference across evaluation criteria, with all criteria split into quartiles, sorted by amount of correlation. As correlation increases, the necessary fraction truthful goes down.

values are quite high—80% or more of students have to be truthful for that to be the best strategy, given that the rest agree on a single perfect low-effort signal with uniform prior. The pattern is not very sensitive to how uniform the prior for the low-effort signal is, as long as it is not too extreme. It is quite sensitive to the score for truthful reporting, and to the assumption that all colluding students agree on a particular correlation method.[17]

The best solutions are to improve the likelihood of agreement for truthful reporting, which would make truthful scores go up and bring the indifference point lower, as well as spot checking and allowing complaints for low scores, as in *Mechanical TA* (Wright & Leyton-Brown, 2015). To see the effects of improved agreement on the intended signal, we sort the assessment criteria by amount of correlation, as measured by total variation distance between the joint and product-of-marginals distributions (equivalently, the expected score of the CA mechanism), and plot a separate histogram of the points of indifference for each quartile, in Figure 4.19. The increased

[17] It seems difficult to agree on such a method in practice in a large online course, without tipping off the course team by discussing it in some public forum.

correlation in higher quartiles means a significantly lower point of indifference.

4.8 Conclusion

We examined patterns of reports in a MOOC peer assessment system, and simulated four peer prediction mechanisms applied to these reports. We found that agreement between peer reports is low overall, and argue that this should be addressed with incentive mechanisms such as peer prediction, for instance through better student training, encouragement, as well as bias-reduction techniques based on machine learning.

We argued that reward variance is an important consideration in mechanism design alongside expected score, and find that the CA mechanism is better in this regard than mechanisms that only reward students based on exact agreement. An experimental follow-up question is whether the variability is low enough to be used in practice. A theoretical question is whether mechanisms with even lower variability can be designed.

We showed that collusion on unintended properties of submissions could be profitable with a small colluding sub-population, given the low base agreement, and suggest that improving agreement between peers and monitoring by the course team will both help deter this behavior.

There are many directions for further research. In peer prediction, this includes exploring more mechanisms, perhaps using the information-theory based framework from Kong & Schoenebeck (2016), that provides a general way to design strong truthful mechanisms in a variety of settings. Another important direction is to find ways to handle user heterogeneity directly in peer prediction methods. A practical system would need to incorporate other approaches to incentive alignment with peer prediction, and there is perhaps no better way to see what works than to try. The mechanisms are well enough understood that experiments and real deployments are both feasible and necessary to complement the theory. Ideally, the incentives for effort would ultimately lead to better student learning.

A concrete suggestion for a low-risk first implementation is to use peer prediction to give students feedback on how well they are assessing each other, without factoring the results into student grades. This should avoid strategic incentive issues and allow comparisons between mechanisms and the incentive effects of ungraded feedback.

5
Conclusion

Congratulations, you've made it to the conclusion! I will summarize what I did, point out a few things I might have done differently, and then discuss opportunities for future work.

While the first peer prediction mechanisms were proposed almost fifteen years ago, many people have been skeptical about whether peer prediction can be made practical. Objections included the need for complex prediction reports in some mechanisms, concerns about high-reward uninformative equilibria, and more recent questions about the validity of the underlying one-signal-per-task model (Gao et al., 2016).

My work in Chapter 2 is one of several new minimal mechanisms that solves the issue of uninformative equilibria, and the first such mechanism that works for small numbers of reports.

As far as I know, the work in Chapter 3 is the first to study peer prediction with learning agents, complementing equilibrium-only analysis and lab and real-world experiments. Chapter 4 has the first empirical study of the potential effects of low-effort signals, confirming that such collusive behavior may be a problem, and requires further study. Chapter 4 also introduces score variability as a concern in educational use of peer prediction, and presents initial results that suggest that mechanisms that reward more than just exact agreement are likely to have lower variability.

There were a number of things I wanted to do that did not happen. The first is theoretical: soon after I started to look for strong truthful mechanisms, I had the sense that there should be an information theoretic way to design and analyze mechanisms that only reward agents for the "extra" information their signal gives them beyond the prior, along the lines of the data processing inequality–informally, the information theoretical idea that "post-processing cannot increase information". I did not end up developing a general framework for this, but luckily, Kong & Schoenebeck (2016) did.

The second obvious gap is experimental. Experiments in real applications would be a great complement to equilibrium analysis and learning dynamics. So far, most experiments have been in toy settings, and while those are informative, I think the details of the application, user population, and how the mechanism is presented are likely to make a big difference. This is an important area for future work.

There are other theoretical and applied areas for future work as well. On the theoretical side, the Kong & Schoenebeck (2016) framework is a great starting place—their framework gives a general way to build a family of peer prediction mechanisms for a domain. This makes systematic optimization possible: for example, what mechanism would minimize score variance, or vulnerability to certain kinds of collusion?

Another theoretical question concerns the notion of informed truthfulness from Chapter 2. It allows for ambivalence among a set of reports, with the truthful report being just one among them. In the binary effort model in the chapter, this works—if agents observe their signal exactly, there is no reason for them to misreport. However, the current definition is likely to break down in a progressive effort model, where agents get more accurate signals as they invest more effort. Is there an expanded notion that captures effort more realistically, while still being achievable for all world models?

A different direction is to extend the Correlated Agreement mechanism to handle heterogeneity among agents without having to learn a separate joint distribution for each pair. One possible approach is to separate the problem into calibration and incentive-alignment—have the system learn the biases of each agent individually, and adjust for them before input to a peer prediction mechanism. The challenge is that this makes the bias calibration algorithm part of the mechanism, and may create new incentive concerns.

In geographical applications—gathering info to improve maps or citizen science projects to track birds or whales—the assumption that agents are assigned tasks randomly does not hold: agents choose where they go. What mechanisms give strong guarantees in such applications?

In peer assessment and business evaluation, users typically answer several questions about the same object—e.g. evaluate the grammar, structure, and style of this essay, Does this restaurant have good food? Is it loud? How expensive it is? My dissertation and other peer prediction work treats each of these questions completely independently, but there may be value in studying the vectors of reports directly, since answers are likely to be correlated. Relatedly, the definition of the task "universe" could be examined more carefully in some settings—should all images in an image labeling application be thrown into the same pool, or segmented somehow? Should businesses in New York, Tokyo, and Mumbai be part of the same pool, treated independently, or further subdivided into neighborhoods? The correlations between reports and expected scores are likely to depend on these decisions.

There are many ways to expand the use of learning dynamics to study peer prediction. Another learning model, other than replicator dynamics, may be a better proxy for reality. Replicator dynamics could be used to study what happens in mechanisms that have signal model assumptions when those restrictions are not met. And there is a need to understand when and how well such dynamics models match real behavior.

There are many applied questions to explore as well. One is how to combine peer prediction with other incentive mechanisms—for example, a system could use peer prediction as the primary mechanism, run a suspicious behavior detector in the background, and give users that trigger the detector a few known-answer tasks.

Succesful use of peer prediction will require going beyond the basic models to also utilize lessons from behavioral economics, psychology and user experience design. In some applications, including education, citizen science, and crowdsourced invention, community building and other social aspects are likely at least as important as the "official" incentives. One example along these lines is Law et al. (2016), which shows that curiosity about a crowdsourcing task is a strong motivating factor in successful completion. The question is how to know which insights from all these fields are most relevant in a given setting, and how best to combine them.

Following my own advice, I want to discuss next steps in peer assessment and education. I already mentioned the need to try peer prediction in practice. Going up a level, recall that the goal for peer assessment is not just accurate multiple choice rubric-based evaluation, but also useful open-ended comments. How can we automatically evaluate the quality of such comments? One approach is to find ways to measure how much peers help each other learn, and how their skill at assessment improves over time.

Ultimately, we need to remember that peer assessment is just a small part of the challenge of open, scalable education. The real question is how to combine many different techniques to enable effective education that is accessible and affordable to all, and this is likely to keep researchers and practitioners busy indefinitely.

References

Arora, S., Hazan, E., & Kale, S. (2012). The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1), 121–164.

Bloembergen, D., Tuyls, K., Hennes, D., & Kaisers, M. (2015). Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53, 659–697.

Busemeyer, J. R. & Townsend, J. T. (1993). Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. *Psychological review*, 100(3), 432–459.

Cai, Y., Daskalakis, C., & Papadimitriou, C. (2015). Optimum statistical estimation with strategic data sources. In *Proceedings of The 28th Conference on Learning Theory* (pp. 280–296).

Chen, Y., Fortnow, L., Nikolova, E., & Pennock, D. M. (2007). Betting on permutations. In *Proceedings of the 8th ACM Conference on Electronic Commerce* (pp. 326–335).

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98(4), 891–901.

Dasgupta, A. & Ghosh, A. (2013). Crowdsourced Judgement Elicitation with Endogenous Proficiency. In *WWW13* (pp. 1–17).

Devroye, L. & Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer New York.

Easley, D. & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Erev, I. & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, 112(4), 912–931.

Erev, I. & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*, 88(4), 848–881.

Erev, I. & Roth, A. E. (2014). Maximization, learning, and economic behavior. *PNAS*, 111.

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education & Training International*, 32(2), 175–187.

Faltings, B., Pu, P., & Tran, B. D. (2014). Incentives to Counter Bias in Human Computation. In *The Second AAAI Conference on Human Computation & Crowdsourcing (HCOMP 2014)* (pp. 59–66).

Frongillo, R. & Witkowski, J. (2016). A geometric method to construct minimal peer prediction mechanisms. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*.

Fudenberg, D. & Tirole, J. (1991). *Game Theory*. Cambridge, MA: MIT Press.

Gao, X. A., Mao, A., Chen, Y., & Adams, R. P. (2014). Trick or Treat : Putting Peer Prediction to the Test. In *The Fifteenth ACM Conference on Economics and Computation (EC'14)*.

Gao, X. A., Wright, R. J., & Leyton-Brown, K. (2016). Incentivizing Evaluation via Limited Access to Ground Truth : Peer Prediction Makes Things Worse. *2nd Workshop on Algorithmic Game Theory and Data Science at EC 2016*.

Gintis, H. (2009). Evolutionary Dynamics. In *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction* chapter 12, (pp. 271–297).

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Goldfinch, J. & Raeside, R. (1990). Development of a peer assessment technique for obtaining individual marks on a group project. *Assessment & Evaluation in Higher Education*, 15(3), 210–231.

Hanson, R. (2003). Combinatorial Information Market Design. *Information Systems Frontiers*, 5(1), 107–119.

Hofbauer, J., Sorin, S., & Viossat, Y. (2009). Time average replicator and best-reply dynamics. *Mathematics of Operations Research*, 34(2), 263–269.

Jain, S. & Parkes, D. C. (2013). A Game-Theoretic Analysis of the ESP Game. *ACM Transactions on Economics and Computation*, 1(1), 3:1–3:35.

Jurca, R. & Faltings, B. (2005). Enforcing truthful strategies in incentive compatible reputation mechanisms. In *Proceedings of the 1st International Workshop on Internet and Network Economics (WINE'05),*, volume 3828 LNCS (pp. 268–277).

Jurca, R. & Faltings, B. (2009). Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34(1), 209–253.

Jurca, R. & Faltings, B. (2011). Incentives for Answering Hypothetical Questions. In *Workshop on Social Computing and User Generated Content, EC-11*.

Kamble, V., Shah, N., Marn, D., Parekh, A., & Ramachandran, K. (2015). Truth Serums for Massively Crowdsourced Evaluation Tasks. *arXiv:1507.07045*.

Kleinberg, R., Piliouras, G., & Tardos, É. (2009). Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing* (pp. 533–542).

Kleinberg, R. D., Ligett, K., Piliouras, G., & Tardos, É. (2011). Beyond the Nash equilibrium barrier. In *ICS* (pp. 125–140).

Kong, Y. & Schoenebeck, G. (2016). A Framework For Designing Information Elicitation Mechanism That Rewards Truth-telling. *arXiv:1605.01021*.

Kong, Y., Schoenebeck, G., & Ligett, K. (2016). Putting peer prediction under the micro(economic)scope and making truth-telling focal. *CoRR*, abs/1603.07319.

Kulkarni, C., Bernstein, M., & Klemmer, S. (2015). PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. *The Proceedings of the Second ACM Conference on Learning @ Scale*.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S. R. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, 20(6), 1–31.

Law, E., Yin, M., Goh, J., Chen, K., Terry, M., & Gajos, K. Z. (2016). Curiosity Killed the Cat , but Makes Crowdwork Better. *CHI 2016*.

Leyton-Brown, K. & Shoham, Y. (2008). *Essentials of Game Theory: A Concise, Multidisciplinary Introduction*. Morgan & Claypool Publishers.

Miller, N., Resnick, P., & Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, 51, 1359–1373.

Orsmond, P. & Merry, S. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, 21.

Panageas, I. & Piliouras, G. (2014). From pointwise convergence of evolutionary dynamics to average case analysis of decentralized algorithms. *arXiv:1403.3885*.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned Models of Peer Assessment in MOOCs. *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee*.

Prelec, D. (2004). A Bayesian Truth Serum For Subjective Data. *Science*, 306(5695), 462.

Radanovic, G. & Faltings, B. (2013). A Robust Bayesian Truth Serum for Non-Binary Signals. In *AAAI13* (pp. 833–839).

Radanovic, G. & Faltings, B. (2014). Incentives for Truthful Information Elicitation of Continuous Signals. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14)* (pp. 770–776).

Radanovic, G. & Faltings, B. (2015a). Incentive Schemes for Participatory Sensing. In *AAMAS 2015*.

Radanovic, G. & Faltings, B. (2015b). Incentives for Subjective Evaluations with Private Beliefs. *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*, (pp. 1014–1020).

Radanovic, G., Faltings, B., & Jurca, R. (2016). Incentives for Effort in Crowdsourcing using the Peer Truth Serum. *ACM Transactions on Intelligent Systems and Technology*, (January).

Reeves, D. M., Wellman, M. P., MacKie-Mason, J. K., & Osepayshvili, A. (2005). Exploring bidding strategies for market-based scheduling. *Decision Support Systems*, 39(1), 67–85.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, (pp. 1–42).

Russell, A. A. (2004). Calibrated peer review-a writing and critical-thinking instructional tool. *Teaching Tips: Innovations in Undergraduate Science Instruction*, 54.

Sandholm, W. H. (2009). Evolutionary game theory. In *Encyclopedia of Complexity and Systems Science* (pp. 3176–3205). Springer.

Smith, J. M. (1972). Game theory and the evolution of fighting. *On evolution*, (pp. 8–28).

von Ahn, L. & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04 (pp. 319–326). New York, NY, USA: ACM.

Waggoner, B. & Chen, Y. (2014). Output Agreement Mechanisms and Common Knowledge. In *Second AAAI Conference on Human Computation*.

Witkowski, J. (2014). *Robust Peer Prediction Mechanisms*. PhD thesis, Albert-Ludwigs-Universitat Freiburg Institut Fur Informatik.

Witkowski, J. & Parkes, D. C. (2012a). A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*.

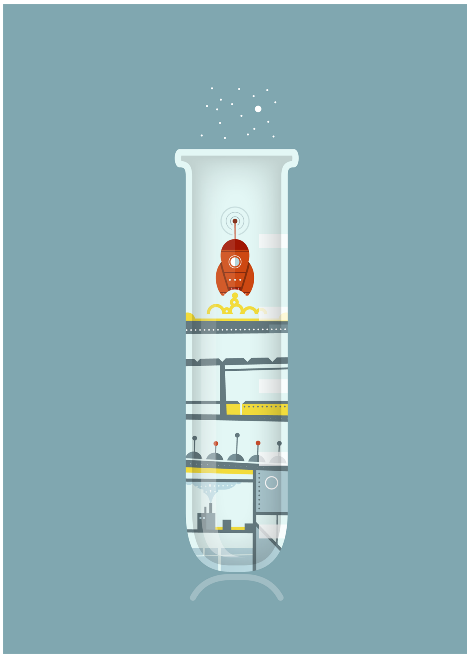
Witkowski, J. & Parkes, D. C. (2012b). Peer Prediction Without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC2012)*.

Witkowski, J. & Parkes, D. C. (2012c). A Robust Bayesian Truth Serum for Small Populations. *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*.

Witkowski, J. & Parkes, D. C. (2013). Learning the Prior in Minimal Peer Prediction. In *The Fourteenth ACM Conference on Economics and Computation (EC'13)*.

Wright, J. R. & Leyton-Brown, K. (2015). Mechanical TA : Partially Automated High-Stakes Peer Grading. In *SIGSCE'15*.

Wu, W., Daskalakis, C., Kaashoek, N., Tzamos, C., & Weinberg, M. (2015). Game theory based peer grading mechanisms for moocs. In *Proceedings of the Second ACM Conference on Learning @ Scale* (pp. 281–286). New York, NY, USA.



THIS THESIS WAS TYPESET using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 12 point Crimson. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with a similar look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.