Proper Scoring Rules & Peer Prediction

Jens Witkowski
University of Pennsylvania

CPS 290.4/590.4: Crowdsourcing Societal Tradeoffs

Duke University

April 3, 2015

Overview

1. Proper Scoring Rules
    - Bonus: Prediction Polls
2. Classical Peer Prediction

CS-Econ Seminar (another bonus): Robust Peer Prediction

Proper Scoring Rules

Proper Scoring Rules [Brier, 1950]

Truthfully elicit beliefs about publicly observable events.

The Good Judgment Project™

#1244 Will India and/or Brazil become a permanent member of the U.N. Security Council before March 1 ?

A: Yes
B: No

1. Agent reports belief $y \in [0, 1]$ of event occurring.
2. March 1: pay $R(y, \omega)$, where $\omega = \begin{cases} 1, & \text{if event occurs} \\ 0, & \text{if not.} \end{cases}$

Naive Approach: Linear Scoring Rule

Linear Scoring Rule

$$R_l(y,\omega) = \begin{cases} y, & \text{if } \omega = 1 \\ 1-y, & \text{if } \omega = 0 \end{cases}$$

True belief $p = 0.6$.

Expected score: $0.6 \cdot y + 0.4 \cdot \big(1-y\big) = 0.4 + 0.2y$

$\Rightarrow y = 1 \neq p$ maximizes expected score.

Linear Rule not proper!

Quadratic Scoring Rule

Quadratic Scoring Rule

$$R_q(y, \omega) = 1 - (y - \omega)^2$$

True belief: $p = 0.6$.

Expected score:

$$\begin{aligned} & p \cdot R_q(y, 1) + (1 - p) \cdot R_q(y, 0) \\ = & \ 0.6 \cdot \left(1 - (y - 1)^2\right) + 0.4 \cdot \left(1 - (y - 0)^2\right) \\ = & \ -y^2 + 1.2y + 0.4 \end{aligned}$$

Derive and set to 0: $-2y + 1.2 := 0 \Leftrightarrow y = 0.6$

Quadratic Rule is proper: $y = p$ maximizes expected score!

Bonus: Prediction Polls™

Good Judgment Project (GJP)

- Forecasting tournament for geo-political questions.
- ~10,000 active forecasters.
- ~140 questions / year.
- Prediction markets and proper scoring rules.
- (Mostly) play money (leaderboard).

Probability elicitation in real world:

- Many forecasters: aggregation?
- Not one-shot: beliefs are continuously updated.
- Not every forecaster reports on every question.
- Not every question has same duration.

How do you translate Proper Scoring Rules into the real world?

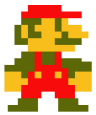
Forecast Aggregation in GJP Prediction Polls

1. Take weighted average:
   - Current score (closed questions): previous accuracy.
   - Frequency of updates: ~effort.
   - Only k most recent forecasts: robustness vs novelty.
2. Extremize:
   - If average < 0.5 $\Rightarrow$ push towards 0.
   - If average > 0.5 $\Rightarrow$ push towards 1.

Extremizing: Intuition

Probability of Heads (H) for biased  ?

- Before observing flip: $p(H) = 0.5$
- Two forecasters observe flip and report:

 $p_1(H) = 0.7$  $p_2(H) = 0.7$

- Aggregated forecast:

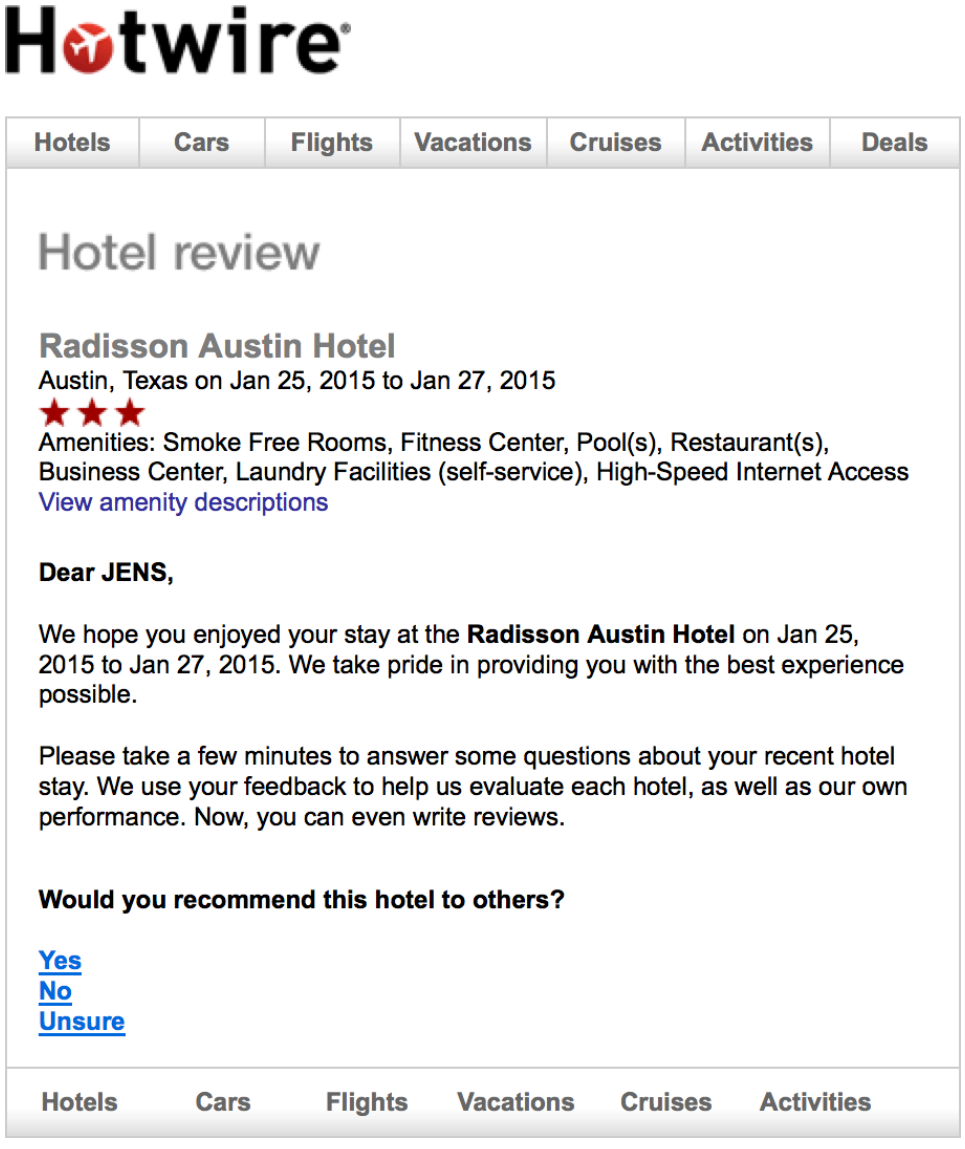
  Same coin flip $\Rightarrow p_{1,2}(H) = 0.7$

  Same coin, different flips $\Rightarrow p_{1,2}(H) > 0.7$

Less information overlap $\Rightarrow$ more extremizing!

Peer Prediction

Motivation: Information Elicitation



Motivation: Information Elicitation

The Good Judgment Project™

#1244 Will India and/or Brazil become a permanent member of the U.N. Security Council before March 1 ?

A: Yes

B: No

Comment ID: 225814, assigned: 09/01/14

Unless the Presidential winner is resolved quickly, there is no chance the Security Agreement will be signed in Wales at the September 4-5 NATO meeting. Audit results are now delayed until at least Sep. 10 and both sides have pulled their observers out of the audit process. Searches for a unity government agreement have also gone nowhere and are unlikely to be successful anyhow. Chances of signing before November 1 seem low. Efforts are under way to find a way to keep American forces in Afghanistan beyond year-end without a security agreement and that would seems to rapidly be becoming the only option short of pulling out troops.

How useful is this comment?

- 1. Not at all Useful (No use of CHAMPS KNOW)
- 2. Slightly Useful
- 3. Useful
- 4. Very Useful
- 5. Extremely Useful (Great Integration of CHAMPS KNOW)

Submit

Motivation: Information Elicitation

Does this Blog Have Any Offensive Content?

Prohibited Sexual Material or Nudity:

- sexually explicit or overtly suggestive content
- nudity (frontal, back or side)
- nudity (particularly of the genitals) covered by a towel, hat or other means
- grabbing, holding or touching genitals or genital area
- transparent/sheer or wet material below the waist or covering women's nipples/breasts
- erections or outline of genitals through clothing
- bare skin one inch directly above the pubic area
- shirtless body shots indoors. shirtless body shots are only allowed in natural settings (e.g. beach or swimming pool)
- cleavage shots without a face
- pubic hair
- underwear, including underwear waistband showing above pants
- body/torso shots without a head/face

CLICK HERE! to visit the blog.

Did you find content on the website that is deemed offensive based on the criteria provided above?

- Yes
- No

Research Questions

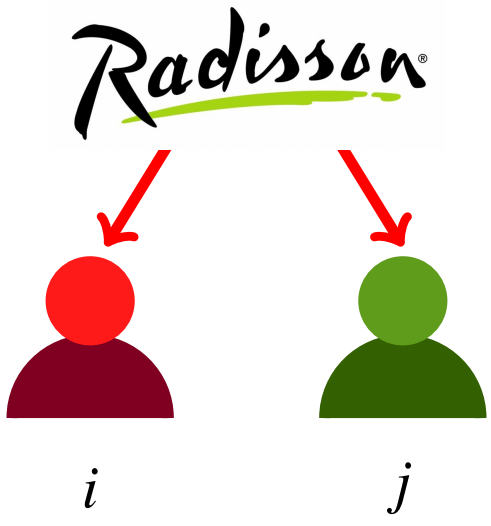
1. How can opinions or experiences be elicited truthfully?

2. How can we incentivize effort for information acquisition?

Basic Setup

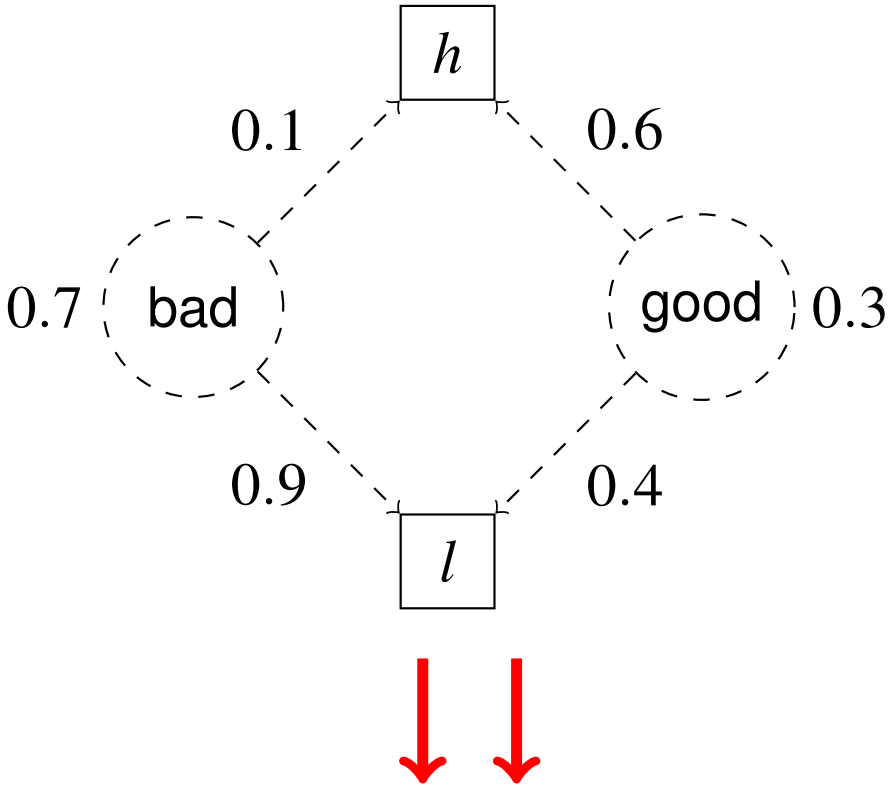
- Elicit informative signal (e.g. “high” or “low” experience).
- Ground truth never observed (e.g. true quality of hotel).
- Allow for payments.

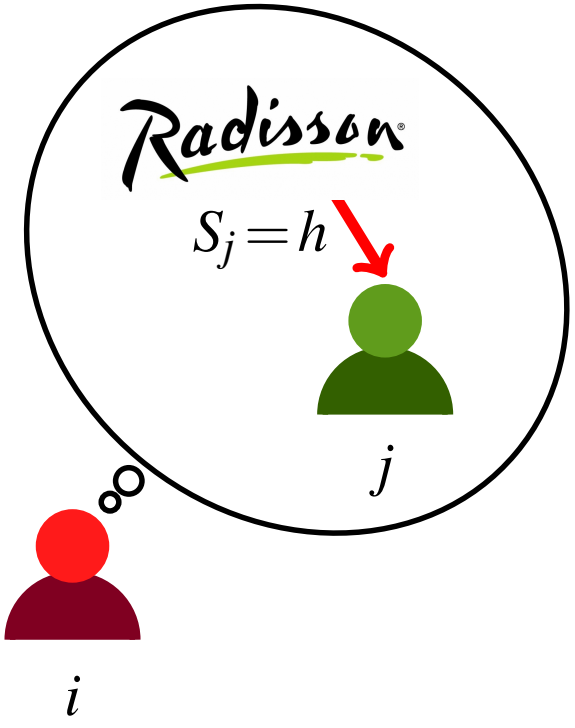
Agents experience same environment:



Key assumption: signals are correlated!

Belief Model (Common Knowledge)





Agent i's belief that agent j observes h:

$p(h|l) = 0.18$

$p(h|h) = 0.46$

Minority Opinions

Agent i's belief that agent j observes h:

$p(h|l) = 0.18$

$p(h|h) = 0.46 \leftarrow$ minority opinion: $p(h|h) < p(l|h)$

Is Chicago capital of Illinois? [Prelec and Seung, 2006]

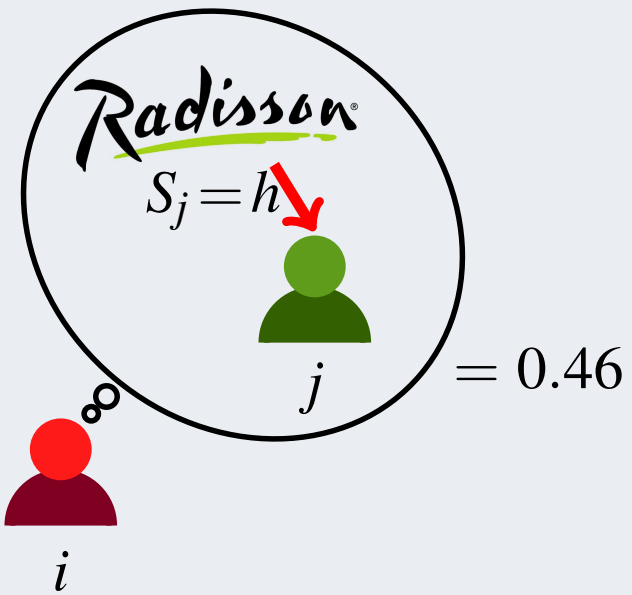
People who know it's not, still believe they're in the minority.

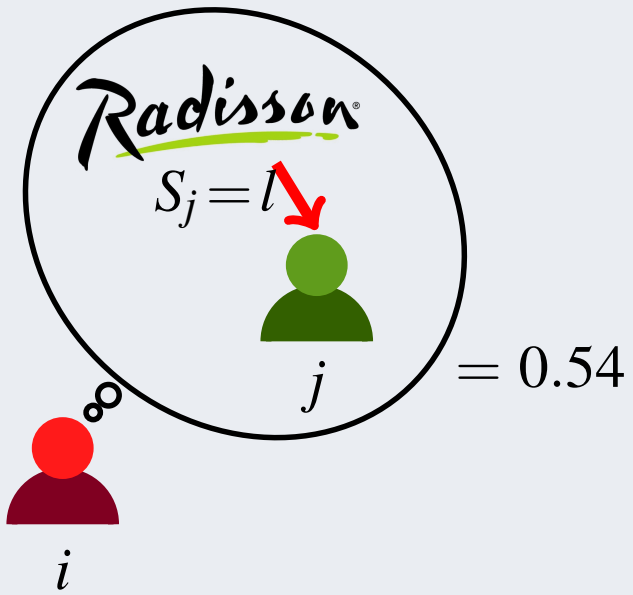
Peer prediction mechanisms elicit minority opinions truthfully!

Output Agreement

Compare two agents' reports and pay: $\begin{cases} \$2 & \text{if reports agree,} \\ \$0 & \text{otherwise.} \end{cases}$

Example with $S_i = h$



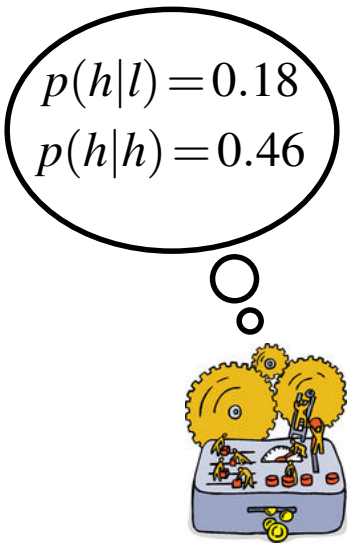


$\mathbf{E}[\text{payment}]$ reporting h: \$0.92 $\mathbf{E}[\text{payment}]$ reporting l: \$1.08

Output Agreement not truthful!

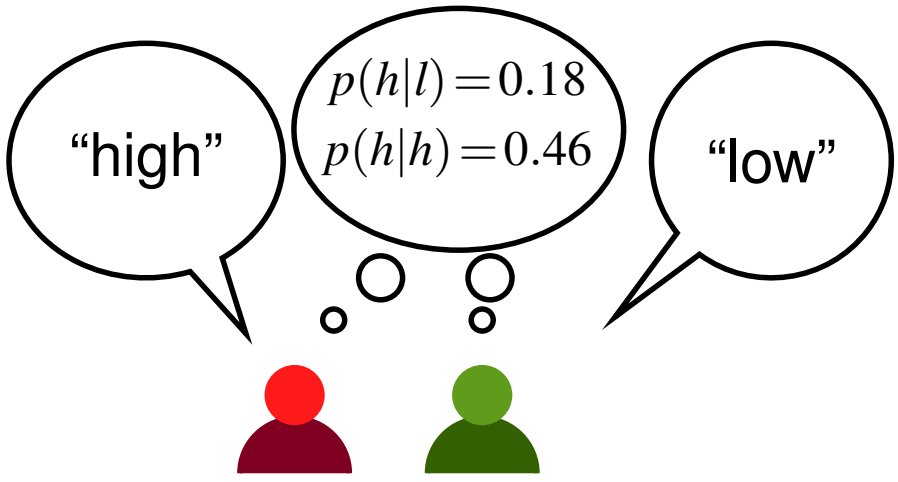
Classical Peer Prediction [Miller et al., 2005]

Mechanism



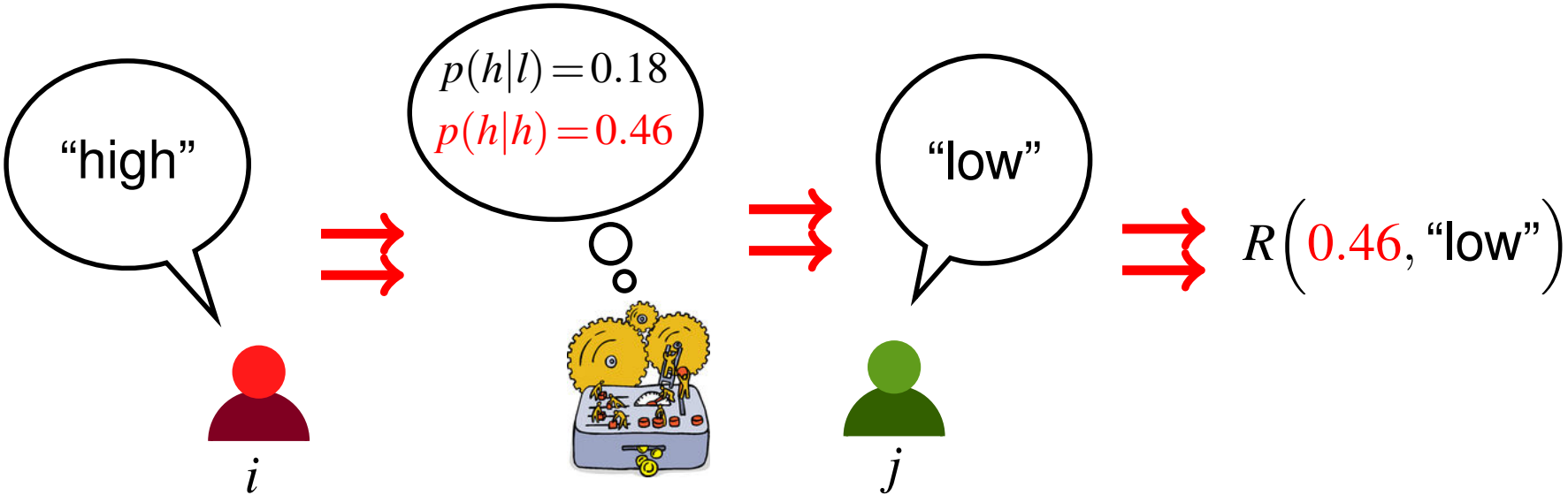
- Knows belief model.

Agents



- Share same belief model.
- Report: Signal.

Classical Peer Prediction: Mechanism



Intuition

1. Define agent j's signal report as event.
2. Restrict possible belief reports to possible posteriors.

Crucial: mechanism knows how to transform signal to belief!

Subsequent Work in Peer Prediction

- Linear Programming formulation [Jurca and Faltings, 2006]
- Collusion-inhibiting mechanisms [Jurca and Faltings, 2009]
- Multiple equilibria unavoidable
  [Waggoner and Chen, 2014]
- Mechanisms not needing to know belief model
  [Prelec, 2004, W. and Parkes, 2012a, Radanovic and Faltings, 2013]
- Mechanisms for subjective prior beliefs
  [W. and Parkes, 2012b, 2013]
- Effort incentives [W. et al, 2013]

Summary

1. Proper Scoring Rules: elicit probabilistic forecasts.
2. Prediction Polls: aggregate forecasts in real-world system.
3. Peer Prediction: elicit opinions, experiences, or ratings.

CS-Econ: Peer Prediction with relaxed common knowledge!

References I

Jurca, R. and Faltings, B. (2006).

Minimum Payments that Reward Honest Reputation Feedback.

In Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06), pages 190–199.

Jurca, R. and Faltings, B. (2009).

Mechanisms for Making Crowds Truthful.

Journal of Artificial Intelligence Research (JAIR), 34:209–253.

Miller, N., Resnick, P., and Zeckhauser, R. (2005).

Eliciting Informative Feedback: The Peer-Prediction Method.

Management Science, 51(9):1359–1373.

References II

Prelec, D. (2004).

A Bayesian Truth Serum for Subjective Data.

Science, 306(5695):462–466.

Prelec, D. and Seung, S. (2006).

An algorithm that finds truth even if most people are wrong.

Working Paper.

Radanovic, G. and Faltings, B. (2013).

A Robust Bayesian Truth Serum for Non-binary Signals.

In Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13), pages 833–839.

References III

Waggoner, B. and Chen, Y. (2014).

Output Agreement Mechanisms and Common Knowledge.

In Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP'14).

Witkowski, J., Bachrach, Y., Key, P., and Parkes, D. C. (2013).

Dwelling on the Negative: Incentivizing Effort in Peer Prediction.

In Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13), pages 190–197.

Witkowski, J. and Parkes, D. C. (2012a).

A Robust Bayesian Truth Serum for Small Populations.

In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12), pages 1492–1498.

References IV

Witkowski, J. and Parkes, D. C. (2012b).

Peer Prediction Without a Common Prior.

In Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12), pages 964–981.

Witkowski, J. and Parkes, D. C. (2013).

Learning the Prior in Minimal Peer Prediction.

In Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC'13).