Modelling Techniques and Inference Procedures for Improving Quality Control in Crowdsourcing Applications

by

Yuan Jin

Thesis

submitted in partial fulfilment

of the requirements for the Degree of

Doctor of Philosophy

Supervisor: Dr. Mark Carman

Associate Supervisor: Prof. Wray Buntine

Monash University

Faculty of Information Technology

Caulfield Campus

November, 2018

Modelling Techniques and Inference Procedures for Improving Quality Control in Crowdsourcing Applications

Declaration

I declare that this thesis is my own work, and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Yuan Jin
13th March 2019 13/03/2019

Acknowledgments

I would like to express my gratitude to my two supervisors: Dr. Mark Carman and Prof. Wray Buntine for their guidance throughout my PhD journey. Without them, I would never be able to build up such an interesting and meaningful PhD project and publish four high-quality papers from it. More importantly, they help me develop a better understanding of machine learning and statistical modelling, and teach me how to become a qualified researcher who always holds high standards for his work. Furthermore, their insightful and constructive comments have significantly improved the quality of this dissertation.

I would like to thank the faculty of Information Technology of Monash and Data61 for providing me with the scholarships that helped me finish my PhD project and traveled overseas to present my papers at great conferences.

I would like to thank all the collaborators on my papers: Dr. Lexing Xie, Dr. Dongwoo kim, Dr. Ye Zhu and Dr. Lan Du who have given valuable advice to my works and spent great amounts of time and effort helping me refine the papers.

Finally, I would like to express my love and deep gratitude to my wife and parents. Without their support, I would never be able to accomplish anything in this amazing four-year journey. They inspire me to be better everyday.

Yuan Jin

Monash University

November 2018

List of Publications

Publications arising from this dissertation are listed as follows:

- Chapter 4 has been published in GamifIR 2016 workshop.

  Yuan Jin, Mark Carman, and Lexing Xie, "A Little Competition Never Hurt Anyone's Relevance Assessments," in Proceedings of the Third International Workshop on Gamification for Information Retrieval (GamifIR). CEUR Workshop Proceedings, Vol. 1642, pp. 29-36, 2016.

- Chapter 5 has been published in HCOMP 2017.

  Yuan Jin, Mark Carman, Dongwoo Kim, and Lexing Xie, "Leveraging Side Information to Improve Label Quality Control in Crowd-Sourcing," in Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP). pp. 79-88, 2017.

- Chapter 6 has been accepted as a conference paper by ACML 2018.

  Yuan Jin, Mark Carman, Ye Zhu, and Wray Buntine, "Distinguishing Question Subjectivity from Difficulty for Improved Crowdsourcing," accepted as a conference paper by the Tenth Asian Conference on Machine Learning (ACML).

- Chapter 7 has been published in PAKDD 2018.

  Yuan Jin, Lan Du, Ye Zhu, and Mark Carman, "Leveraging Label Category Relationships in Multi-class Crowdsourcing," in Proceedings of the Twenty-Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Springer, pp. 128-140, 2018.

- Chapter 2 will be submitted to the TKDE journal.

  Yuan Jin and Mark Carman. "A Survey of Statistical Modelling Techniques and Inference Procedures for Quality Control in Crowdsourcing."

Contents

Acknowledgments . . . . . **iii**

List of Publications . . . . . **iv**

List of Tables . . . . . **x**

List of Figures . . . . . **xii**

Abstract . . . . . **xvii**

1 Introduction **1**

1.1 Paid vs Unpaid Crowdsourcing 1

1.2 Paid Crowdsourcing Platforms 2

1.2.1 Quality Control for Paid Crowdsourcing 5

1.2.2 Modelling the Quality of Responses (QoR) 9

1.2.3 Correlations between QoR and Crowdsourcing Aspects 10

1.3 Notation 18

1.4 Conclusion 20

2 Literature Review **22**

2.1 Survey Outline 26

2.1.1 A Graph Visualization on Past QCC Research 26

2.1.2 Sections of this Survey 27

2.2 Modelling Worker Ability 29

2.2.1 The Dawid & Skene Model 29

2.2.2 Improvements on Parameter Estimation 32

2.2.3 Dealing with Sparsity in DS 33

2.2.4 Theoretical Bounds on Error-Rate of DS Estimation Techniques 34
2.2.5 Non-Probabilistic Worker-Ability Models 35
2.2.6 Truth-Discovery Worker-Ability Models 35
2.3 Modelling Worker Expertise and Question Difficulty 36
2.3.1 The GLAD Model 36
2.3.2 Improvements on Parameter Estimation 39
2.3.3 Multi-dimensional Worker Expertise and Side Information 40
2.3.4 Neural Network Approaches 41
2.3.5 Dealing with Ordinal Response Data 42
2.4 Modelling Worker Preferences and Question Subjectivity 43
2.5 Modelling Worker Motivation and Worker Contexts 45
2.5.1 Incentive Mechanisms Based on Monetary Payment 47
2.5.2 Monetary Payment in Task-Level Contexts 47
2.5.3 Monetary Payment in Response-Level Contexts 48
2.5.4 Incentive Mechanisms Based on Gamification 52
2.5.5 Gamification in Response-Level Contexts 52
2.5.6 Gamification in Session-Level Contexts 55
2.5.7 Gamification in Task-Level Contexts 55
2.6 Modelling Worker Expertise and Contexts 56
2.6.1 Improving Worker Expertise Using Training Mechanisms at Different Levels of Contexts 57
2.6.2 Question Allocation in Crowdsourcing 58
2.6.3 Non-Adaptive Question Allocation Based on Worker Expertise in Task-Level Contexts 59
2.6.4 Adaptive Question Allocation Based on Worker Expertise in Response-Level Contexts 60
2.7 Modelling Worker Motivation, Question Difficulty and Contexts 61
2.7.1 Monetary Payment in Response-Level Contexts 62
2.7.2 Gamification in Session-Level Contexts 62
2.8 Modelling Worker Expertise, Question Difficulty and Contexts 63
2.8.1 Modelling Interactions in Task-Level Contexts 63
2.8.2 Modelling Interactions in Session-Level Contexts 65

2.8.3 Modelling Interactions in Response-Level Contexts 66
2.8.4 Leveraging Question Difficulty in Adaptive Question Allocation 66
2.9 Modelling Worker Expertise, Question Difficulty and Response Relationships . . 68
2.10 Conclusion 69
2.10.1 Payment, Gamification and Training Designs 70
2.10.2 Issues with Designing Gamified Crowdsourcing Tasks 71
2.10.3 Statistical Modelling and Estimation 71
2.10.4 A Common Issue of Attribute Estimation 71
2.10.5 Issues of Modelling and Estimation with Partially Subjective Questions . 73
2.10.6 Issues of Modelling and Estimating Response Semantic Relationships . . 73

3 Research Questions 75
3.1 Research Question 1 75
3.2 Research Question 2 76
3.3 Research Question 3 77
3.4 Research Question 4 78

4 Quality Control Designs with Leaderboards and Performance Feedback 80
4.1 Related work 81
4.2 Experiments on Gamification Designs with Performance Feedback and Leaderboards 81
4.2.1 Experiment 1: Real-time Feedback 83
4.2.2 Experiment 2: Adding a Bonus 87
4.2.3 Experiment 3: Control Questions 89
4.3 Conclusion 92

5 Leveraging Side Information for Improved Quality Control on Sparse Responses . 94
5.1 Related Work 96
5.2 Proposed Framework 96
5.2.1 Basic Framework 97
5.2.2 Incorporating Worker Information 98
5.2.3 Incorporating Question Information 99
5.2.4 Incorporating Session-Level and Response-Level Contextual Information 99
5.3 Parameter Estimation 100

5.3.1 Collapsed Gibbs Sampling for Estimating True Answer Posterior 101
5.3.2 Gradient Descent for Estimating Other Model Parameters 101
5.3.3 Estimating Worker Expertise and Question Difficulty e_i and d_j 102
5.3.4 Estimating Task-, Worker- and Session-level Regression Coefficients . . 102
5.3.5 Estimating Response-level Regression Coefficients 102
5.4 Experiments 103
5.4.1 Datasets 103
5.4.2 Feature Collection 105
5.4.3 Experiment Setup 106
5.4.4 Hyper-parameter Setup 108
5.4.5 Prediction with Subsampled Responses 109
5.5 Results 109
5.5.1 True Answer Prediction 109
5.5.2 Unseen Held-out Response Prediction 110
5.5.3 True Answer Prediction with Subsampled Responses 111
5.5.4 Unseen Held-out Response Prediction with Subsampled Responses . . . 112
5.5.5 Statistical Analysis of Feature Importance 113
5.6 Conclusion 113

6 Distinguishing Question Subjectivity from Difficulty for Improved Quality Control 116
6.1 Related Work 120
6.1.1 Latent Variable Modelling in Crowdsourcing 120
6.1.2 Latent Variable Modelling in Collaborative Filtering 121
6.2 Proposed Model 121
6.3 Estimation 124
6.3.1 Model Parameter Estimation 124
6.3.2 True Answer Estimation 125
6.3.3 Subjectivity Estimation 126
6.4 Experiments and Results 127
6.4.1 Datasets 127
6.4.2 SDR Hyper-parameter Setup 128
6.4.3 Sensitivity Analysis 129

6.4.4 Question True Answer Prediction . . . 131
6.4.5 Worker Response Prediction . . . 132
6.4.6 Subjectivity and Difficulty Estimate Consistency with Human Assessment 134
6.5 Conclusion . . . 136

7 Leveraging Response Semantic Relationships for Improved Quality Control in Multi-class Crowdsourcing . . . 139
7.1 Related Work . . . 142
7.2 Problem Formulation . . . 143
7.3 Proposed Model . . . 144
7.4 Parameter Estimation . . . 147
7.5 Experiments and Results . . . 148
7.5.1 Datasets . . . 148
7.5.2 True Answer Prediction . . . 149
7.5.3 True Answer Prediction Under Response Sparsity . . . 152
7.5.4 Consistency between Estimated Relatedness and Ground-Truth Relatedness 152
7.6 Conclusion . . . 154

8 Conclusion . . . 157
8.1 Identification of Important Crowdsourcing Aspects . . . 158
8.2 Definition of Granularities of Worker Contexts . . . 159
8.3 A Quality Control Research Graph Visualization . . . 160
8.4 Technical Taxonomies of Quality Control Papers . . . 160
8.5 Identification of Design Issues for Gamifying Crowdsourcing . . . 161
8.6 Identification of Modelling and Inference Issues for Statistical Response Aggregation 161
8.7 Empirical Study of Gamified Paid Crowdsourcing . . . 162
8.8 A Unified Scalable Framework to Leverage Side Information for Sparse Response Aggregation . . . 164
8.9 A Statistical Model Distinguishing Subjectivity from Difficulty for Partially Subjective Questions . . . 166
8.10 A Statistical Model Leveraging Semantic Relationships between Response Options 169
8.11 Future Directions . . . 171

List of Tables

1.1 List of Notation used in this thesis. 20

2.1 A summary of surveys regarding quality control for crowdsourcing. There are two types of surveys focusing on either the design of crowdsourcing applications or statistical methods. Their strengths and weaknesses have been specified in the table. 23

4.1 Results across the four groups for Experiment 1. Accuracy is given as both Micro and Macro averages, with the latter being aggregated at the *worker level* (i.e. by treating the average performance of each worker as a single observation). Crowd-workers who judged less than 50 documents were excluded from the analysis. 87

4.2 Results for Experiment 2, where a bonus was offered to participants in Control group 2 and Treatment groups 1 and 2. 89

4.3 Results for Experiment 3, where control questions were used (in both the control and the treatment groups) to vet and remove crowd-workers based on their corresponding accuracy. 91

5.1 Dataset Summary. The headers correspond to the notation introduced in Table 1.1. 103

5.2 features encoding different types of side information. 105

5.3 True label/answer predictive accuracy of the models across the three datasets. We denote the side information about the workers, the question items, the sessions and the responses respectively with capital letters “L", “I", “S" and “R". 110

5.4 Unseen (held-out) response prediction error of the models across 30% held-out response data from the three datasets. 111

5.5 Comparison between top 5 most predictive features for supervised and unsupervised (actual) setting. 113

6.1 The objective and the partially subjective datasets used in this paper. The headers correspond to the notation introduced in Table 1.1. 127
6.2 Average accuracy of our model with 1 and 2 latent preferences on predicting the held-out validation response of each worker over 4 objective tasks. 130
6.3 Average accuracy of our model with 1, 2 and 3 latent preferences on the held-out validation prediction over 10 partially subjective tasks the first 5 of which are sub-tasks of the *Image* task. 130
6.4 Accuracy of all the models on predicting the true answers of the four partially subjective datasets (the results for the *Image* task are not included as the number of items in this task is too small to show any difference in the performance of different models). 131
6.5 Average accuracy of all the models on predicting the unseen held-out test response of each worker across all the partially subjective datasets. 132

7.1 Example 1 of equation 7.2 146
7.2 Example 2 of equation 7.2 146
7.3 Dataset Summary. The headers correspond to the notation introduced in Table 1.1. 149
7.4 The accuracy of different models on inferring the true (answer) responses of the (question) items across the four datasets. 150

List of Figures

1.1 An example of how majority vote aggregation can be improved by correlating the quality of each answer with their respective workers. 8

1.2 Aspects of crowdsourcing applications that have been studied by past research for better controlling the quality of responses. Each ellipse node denotes a particular crowdsourcing aspect. Four major aspects at level 1 are worker, question, response option and context. Question and context have finer definitions at level 2. Rectangle nodes denote prominent attributes of aspects that have been modelled or utilized in the modelling by the QCC methods. 11

1.3 The green circles represent the four supporting systems of a crowdsourcing platform: the question allocation system, the reward system, the training system and the quality assurance system. They provide components of worker contexts, illustrated as the intersections with the red circle denoting the contexts. Different settings of these components (e.g. by mechanisms deployed in the systems) change the contexts which further influences workers' behaviour. Note that both the training system and the quality assurance system embed quizzes into the contexts. 14

2.1 A diagram that shows the respective parts played by designs and statistical models in quality control for crowdsourcing. The red box shows the part of QCC designs which aim to intervene in crowdsourcing to improve response quality. The blue box shows the part of statistical models which yield estimates of true answers and other attributes. Each part has its dedicated surveys but there lacks a current survey that links the statistical models with the designs. 24

2.2 A graph visualization of the past QCC research. Vertices (except the root) denote either crowdsourcing aspects, key attributes or their combinations. Paths indicate different lines of QCC research that considered different (combinations of) aspects and attributes. Tags on the edges correspond to the types of QCC methods studied by the particular lines of research. 26
2.3 A taxonomy of QCC papers that only considered worker ability/accuracy/expertise to account for response quality. These papers focused on statistical modelling and inference. 30
2.4 A taxonomy of QCC papers that considered both worker expertise and question difficulty to account for the response quality. These papers also focused on statistical modelling and inference. 37
2.5 A taxonomy of QCC papers that considered worker preferences and question subjectivity. These papers focused on statistical modelling and inference. 43
2.6 A diagram shows payment and gamification mechanisms control worker contexts to affect worker motivation, which further influence worker effort and truthfulness, and eventually the correctness of responses. 45
2.7 A taxonomy of QCC papers that considered the interaction between worker context and (extrinsic) motivation. These papers focused on designing monetary payment mechanisms which rely on statistical modelling (and possibly estimation) of worker attributes. 46
2.8 A taxonomy of QCC papers that considered the interaction between worker context and (intrinsic) motivation. These papers focused on designing gamification mechanisms which, in most cases, rely on statistical estimation of worker accuracy/expertise. 53
2.9 A diagram shows training mechanisms control worker contexts to improve worker expertise, which further improves the correctness their responses. 57
2.10 A taxonomy of QCC papers that considered the interaction between worker expertise and context. These papers focus on designing either training mechanisms that alter contexts to improve worker expertise or question allocation mechanisms that use worker expertise to determine questions to be answered in the contexts. . . . 58

2.11 A taxonomy of QCC papers that considered worker motivation, context and question difficulty. These papers focused on designing either monetary payment mechanisms which additionally modelled question difficulty or gamification mechanisms which increase question difficulty to challenge workers. 61

2.12 A taxonomy of QCC papers that considered worker expertise, context and question difficulty. These papers focused on statistical modelling of how worker expertise varies with contexts at specific levels, and the interaction between the context-aware expertise and the difficulty. 64

2.13 A taxonomy of QCC papers that considered worker expertise, question difficulty and semantic relationships between responses. These papers focused on statistical modelling which leveraged pre-computed response similarities from external knowledge to account for response biases. 68

2.14 A diagram shows how worker expertise, question difficulty and response semantic relationships contribute to the correlations within responses. 69

4.1 Architecture of experiment environment. Crowd-workers are randomly assigned to a particular group after joining the task on CrowdFlower, and see the same version of the interface for each subsequent interaction from their browsers. The server updates their performance statistics (or rank positions) once each page of judgements was completed.
. 83

4.2 A screenshot of the interface shown to crowd-workers assigned to Treatment group 2 for Experiment 1, containing a leaderboard ranking contributors based on their labelling accuracy in percentage.
. 84

4.3 A screenshot of the interface shown to crowd-workers assigned to the control group for Experiment 1, which contained no additional information about the workers' performance.
. 85

4.4 A boxplot showing the distribution of Accuracy values across the crowd-workers for different groups in Experiment 1.
. 86

4.5 A screenshot of the interface shown to crowd-workers assigned to Treatment Group 2 in Experiment 2. 88
4.6 A boxplot of Accuracy across the different groups in Experiment 2 where a $1 bonus was paid to the top 10 contributors in Control group 2 and Treatment groups 1 and 2 at the end of the task. 90
4.7 A boxplot of Accuracy across the different groups in Experiment 3 where control questions were used to guarantee a certain level of relevance judging accuracy. . . . 91

5.1 Models we have developed for extending the *GLAD model* (a) with *worker features* (b), *question features* (c), *worker session features* (d), and *response features* (e). . 97
5.2 a question for relevance judgement 104
5.3 a question for Stack Overflow post status judgement 104
5.4 Changes of the true label/answer predictive accuracy by varying the number of responses subsampled from each worker across the three datasets. 112
5.5 Changes of the unseen (held-out) response predictive accuracy by varying the number of responses subsampled from each worker across the three datasets. . . 112

6.1 Heatmaps showing inter-worker response similarity (% of response agreement) for two different tasks: (a) a *relatively objective* product matching task and (b) a *more subjective* fashion judging task, both involving binary worker responses. Hierarchical clustering was performed to order workers such that similar workers are close together. The three yellow blocks in (b) indicate three groups of response behaviour and higher subjectivity for task (b). 118
6.2 (a) shows GLAD with a latent variable l_j for each objective truth, (b) shows a collaborative filtering model without objective truths, and (c) is the proposed *subjectivity-and-difficulty response* (SDR) model for partially-subjective questions that is able to distinguish question difficulty from subjectivity. 119
6.3 (a) shows the 3 worker clusters on identifying sky from images and (b) shows the 4 worker clusters on judging beautiful images. 131

6.4 (a) and (b) show the correlations of both the difficulty and the subjectivity estimates with the corresponding rankings judged by human assessors, while (c) and (d) show the correlations respectively with the levels to which the images were categorized by the assessors. (e) shows the images as points with coordinates being the difficulty and the subjectivity estimates, and highlights some points, while (f) shows their corresponding images. 133

7.1 (a) Worker response accuracy versus category relatedness. (b) question difficulty (in terms of response error) versus category relatedness. 142

7.2 The DELRA model with and without encoding observed knowledge matrix $\boldsymbol{X}$ specifying relationships between categories are shown in Figure 7.2a and 7.2b. . 144

7.3 The accuracy of different models on inferring the true answers of the items from 10% to 50% of the total responses across the four datasets. Note that x-axis and y-axis in each figure are respectively the sampling proportions of responses and the average true answer prediction accuracy over 10 runs. 151

7.4 Average Pearson correlation coefficients between the Top-N most related category rank yielded by different methods, and the ground-truth Top-N rank yielded by the pre-computed related scores based on equation 7.1. Note that x-axis and y-axis in each figure are N and average correlation, respectively. 153

ABSTRACT

Online crowdsourcing provides a more scalable and cost-effective path to collecting knowledge (e.g. labels) about various types of data items (e.g. text, audio, video) as compared to the traditional in-house counterpart of employing professional labellers. Crowdsourcing is known however to result in larger variance in the quality of recorded answers, preventing them from being used directly for training machine learning systems. To alleviate this problem, multiple crowd-workers are asked to answer the same question with the consensus opinion providing a more reliable final answer. However, due to tight budgets, the number of answers collected per question is typically too low to allow for naive aggregation based on the majority-vote to be effective. To resolve this issue, more nuanced wisdom-of-the-crowd approaches have been developed over the years. The effectiveness of these approaches can be credited to their ability to estimate individual worker accuracies. Despite the success of these methods, their modelling of worker ability still needs further refinement. In particular, there exist underlying correlations between worker ability and worker context that have been overlooked and need to be identified and properly modelled.

In this dissertation, I endeavour to address the existing research gaps in quality control for crowdsourcing in the following ways. First, I present a comprehensive literature review on quality control for crowdsourcing. The review specifies which types of correlations have already been identified and modelled. Then, I proceed to investigate more effective modelling of the important interactions. This involves modelling the correlation between answer quality and (i) worker motivation, as influenced by gamified incentivizing mechanisms combining leader-boards, bonuses, etc., and (ii) workers' abilities, as influenced by their labelling context. I then proceed to identify and efficiently model other sources of variance in worker responses, such as (iii) question subjectivity, modelling it affects the variance of crowd-workers' responses to the same questions, and (iv) frequent confusions, caused by semantic relationships that exist between the different response options. In summary, this dissertation develops a suite of response modelling techniques and inference procedures for improving quality control in crowdsourcing applications.

Chapter 1

Introduction

With the advent of Web 2.0 functionality, users of the Web gained the ability to submit questions online and get answers from other users. Crowdsourcing provides a mechanism by which submitted questions are distributed and solved by generally large and anonymous online crowds. When answering questions, the online crowds are characterized by different and variable *motivation* (e.g. being money-driven or enjoyment-driven), and different and varying degrees of *expertise*. As a result, they exhibit more diverse and in general less accurate question-answering behaviour as compared to *in-house workers* who are trained to work more professionally and specifically on internal platforms for tasks of particular companies [1]. On the other hand, online crowds are more readily accessible and usually less expensive than in-house workers [2, 3, 4].

1.1 Paid vs Unpaid Crowdsourcing

Crowdsourcing can be broadly categorized into being either *unpaid* or *paid*. Unpaid crowdsourcing relies on online platforms accommodating large and diverse communities of *volunteers*. Volunteers are self-motivated to participate and respond accurately due to their pursuit of intangible goals such as fun and enjoyment, skill acquisition and knowledge development, sense of belonging to a community, and altruism [5]; rather than those tangible goals, most notably, money. There are countless varieties of unpaid crowdsourcing platforms online most of which are *open source projects* advertised online to attract public contributions. Two particularly famous varieties are:

- *Wikipedia*[1] is the most famous public open-source unpaid crowdsourcing platform. It relies on volunteers to process unstructured and organic knowledge into well-indexed dedicated Webpages that describe known entities.

- *Citizen science* is another fast-growing venue for unpaid crowdsourcing, in which scientific research projects are (partially) opened to the online public for their participation and contributions. A notable citizen science project is *Galaxy Zoo* [6] which calls upon amateur volunteers to annotate/classify galaxy images in terms of their shapes. This project has been studied widely by crowdsourcing research communities to understand the motivations and behaviour of those volunteers [7, 8, 9, 10].

Even though the volunteers very often perform tedious tasks such as labelling items, there appears to be little compromise in the quality of answers as compared to using paid crowdsourcing [11]. However, this requires that the requesters much more carefully design their tasks (or even the entire projects) to make them attractive enough for the online crowds to participate in and constantly reignite the crowd's internal motivations to stay connected and perform reasonably well. As a result, such a design process usually induces more overhead than the corresponding process in paid crowdsourcing where financial rewards are often sufficient to incentivize the crowds.

This dissertation concentrates on the paid crowdsourcing as we focus on low-overhead crowdsourcing applications which are simple to design, straightforward to implement and easy to manage and collect answers from the online crowds, but which still preserve all the internal characteristics of the crowds in unpaid crowdsourcing.

1.2 Paid Crowdsourcing Platforms

Over the past decade, online human intelligence marketplaces have been thriving, providing organized and billed crowdsourcing services to requesters all over the world. The requesters post their *tasks* on the marketplaces each of which typically contains the following elements:

- *Requirements* which specify what a requester attempts to collect via crowdsourcing (e.g. labels for images, judgements for sentiments of tweets and etc.), and any constraint that the requester imposes on the collection process. The constraints can include (i) the limit on the financial budget, (ii) the minimum qualification for the workers to participate in the task,

[1] `https://en.wikipedia.org/wiki/Main_Page`

such as education levels, languages, minimum accuracy for passing a qualification quiz, etc, and (iii) the maximum workload and working hours for the qualified crowd-workers.

- *Instructions* which generally articulate how a crowd-worker can satisfy the requirements of the requester. The instructions can consist of guidelines regarding the fulfilment of the requirements and rules that regulate the behaviour of the worker.

- *Questions* which need to be answered by crowd-workers either for the qualification purpose when the questions are situated in the quiz or for the data collection purpose when they appear on each of the Web-pages shown to the workers. We will discuss the characteristics of the questions in both cases with more details in later sections of this chapter.

Online crowds are registered with the marketplaces as *crowd-workers* who are autonomous and generally receive a small payment, typically a few cents [12], for finishing each question. Two popular crowdsourcing marketplaces are Amazon Merchanical Turk[2] (AMT) and CrowdFlower[3]. Both of them provide requesters with facilities including templates for designing their own tasks, qualification filters on the participants (e.g. English-speaking only), payment settings on base rates and bonuses, training and vetting mechanisms, launching and monitoring panels (e.g. showing real-time statistics of the participating workers and feedback from them) and message systems that enable the requesters to reach out to the workers directly. The above facilities enable the requesters to be both time and cost efficient in collecting answers. They also provide moderate quality guarantees on the answers through training and vetting of crowd-workers.

In recent years, AMT and CrowdFlower have become very successful in providing human intelligence support to the machine learning and data mining communities. They allow the communities to perform large-scale label collection for data items used to train all kinds of machine learning systems such as learning-to-rank systems in Information Retrieval [13], machine translation systems [14] and general supervised learning systems [15, 16, 17]. As these crowdsourcing marketplaces continue to grow by attracting more answer collection tasks with ever larger sizes, their crowd-workers start to face more choices of tasks to participate in at any point of their *lifetime* on the platform. In this case, many of them start to adjust their *motivations* to maximize their monetary income by accelerating their work in consecutive tasks or by working on concurrent tasks as their total payments received at the end are proportional to their respective question completion rates. Increasing such rates usually results in a *deterioration* in the *quality of workers' answers* to

[2] https://www.mturk.com/
[3] https://www.crowdflower.com/

various extents depending on how well the workers cope with more intensive workloads. To make matters worse, the motivation to maximize the monetary income has spawned varieties of cheating behaviour that is now prevalent on crowdsourcing platforms [18].

Moreover, as the sizes of tasks grow larger on the platforms, hundreds or even thousands of crowd-workers are often required to participate in them. In this case, the difference in their respective *expertise* (e.g. some workers excel at judging the relevance of documents regarding sports related queries, while others are skilled at judging the relevance of science related documents), and the changes of the levels of their expertise over time (as they interact further with the tasks) become significant in determining the quality of their answers.

Furthermore, larger tasks might also lead to the questions being more diverse in terms of their *topical* areas, their level of *difficulty* or even their level of *subjectivity*. For tasks involving questions on varied topics, good matching between the (topical) expertise of the workers and the topics of the questions can have positive influence on the quality of the workers' answers. Likewise, some questions in a task can be much harder to answer than the others. For example, consider the task of counting the exact number of people in a photo. The question is easy if the photo contains only a few individuals and hard if it contains large crowds. Therefore, the answers to the more difficult questions are more prone to errors. In addition, a task can also contain a mix of subjective and objective questions. For example, considering a task which asks crowd-workers to judge whether the people in images are wearing fashionable clothes or not. Since the questions for this task refer to inherently subjective concept (of current fashion), we should expect to see more variation in the responses across crowd-workers. Note that some questions for this subjective task may still be considered entirely objective (e.g. because the corresponding images do not contain any people). Since the requester may be unaware that some of their posted questions are subjective, being able to properly distinguish subjective answers from incorrect answers can become crucial.

Under different *contexts*, a crowd-worker might yield different answers (with various levels of quality) even to the same question. For example, a worker who is currently on a crowded bus using her mobile phone to answer the question is more likely to be knocked and click on the wrong answer than she would be when working more comfortably from home using her laptop. Finally, when certain *response options* have close *semantic relationship* with one another, crowd-workers often tend to confuse similar responses. For example, crowd-workers judging dog breeds may mistake a Labrador for a Golden Retriever but not a Poodle. The bias towards certain similar (but incorrect) answers tends to fade away for less capable crowd-workers who provide essentially more

random responses. The level of bias can depend on question difficulty with easier questions tending to see more responses that are similar to their true answers, while difficult questions tending to see more random responses.

In this dissertation, our goal is to develop techniques that are able to effectively *control* the *quality* of answers from crowd-workers to questions in crowdsourcing. This is done through the efficient *modelling* of the *correlation* of answer quality with the various *aspects* of crowdsourcing such as crowd-worker accuracy, question difficulty, worker contexts and semantic similarity between responses.

1.2.1 Quality Control for Paid Crowdsourcing

As mentioned earlier, crowdsourced responses generally vary in quality. Intuitively, the *quality of a response* from a crowd-worker to a particular question determines how likely the answer is to be correct for the question. In the literature on crowdsourcing, this quality can also be interpreted as the *probability* of the response being correct. A high-quality response means it is less likely to be an error and conversely, a low-quality response means it is more likely to be an error. When the response is indeed incorrect, this error can be either *random* or *systematic* in nature. The random error is viewed as an arbitrary guess by the worker as to the correct answer for the question, while the systematic error is a systematic tendency of that worker to choose a certain response other than the correct answer for the question [19]. A random error is also referred to as *noise* from which the corresponding true answer is *unrecoverable*, while a systematic error results from *bias* from which the corresponding correct answer is *recoverable* [20, 21]. To be more specific, when a worker's response to a question is due to a random error, the *uncertainty* about the correct answer of the question is not reduced by it. In other words, knowing that response provides no information to the posterior knowledge of the correct answer (needed for its recovery). When the response is a systematic error, the uncertainty about the correct answer is essentially reduced by the extra information this response provides: the correct answer being one of the other options. Therefore, a desirable quality control mechanism should be able to separate the random errors/noise from the systematic errors/biases in the responses and then reverse the bias information to recover the true answers of the questions. The conventional quality control mechanism applied in crowdsourcing marketplaces is to vet and *filter* the workers using *gold standard* (*control*) questions for which the true answers are already known by the requesters ahead of time.

Worker-Filtering

If each question in a crowdsourcing task is given to only one crowd-worker to answer, then there is a significant chance that the response from that worker could be incorrect. A basic remedy provided by the crowdsourcing marketplaces for the above issue is *worker-filtering*. This process removes two types of workers: *unqualified* workers and *low-performing* workers. The *unqualified* workers are removed using a *quiz* before the task commences. The quiz contains only gold standard control questions (for which the true answer is known) and each worker must achieve a certain accuracy on the control questions to be admitted to the task. The *low-performing* workers are removed from the worker pool during their participation in the task using unseen control questions embedded amongst the target questions in each task page. Since the filtering mechanism typically removes all of the responses by a crowd-worker as soon as is deemed to be too low-performing, any possibly correct answer that this worker has given up to that point will be lost. This can result in the budget and time wasted on recruiting those low-performing workers who were eventually removed. Moreover, other qualified crowd-workers will need to redo all the questions for which the removed workers had already answered correctly. To save both the budget and time, an improved quality control mechanism is needed that allows for the retention of all the correct responses from those low-performing workers while minimizing the impact of their incorrect responses.

The Wisdom of the Crowd

An alternative to the filtering that keeps both the correct and the incorrect responses while "smoothing out" the influence of the incorrect ones is *the wisdom of the crowd* (WoC) [22]. It centres on the observation that proper *aggregation* of *multiple answers* given by crowd-workers to the same question is able to yield an overall answer that is close to the correct answer of that question. Such a phenomenon has been observed widely in practice. For example, websites, such as Rotten Tomatoes[4], Netflix[5] and Last.fm[6], utilize the wisdom of the crowd (i.e. aggregation of reviews from their respective users) to provide summary reviews and overall recommendation scores about products in specific domains (e.g. film and television, music) to influence the behaviour of the larger online public. These summary reviews and scores have turned out to be very accurate in depicting the quality of the products. The wisdom of crowds happens in a slightly more subtle way in many question-answering Websites (e.g. Stack Overflow, Quora) which do not directly provide

[4]https://www.rottentomatoes.com/
[5]https://www.netflix.com/
[6]https://www.last.fm/

single overall answers to questions. Instead, a question keeps gaining new answers from users which complement one another. From these answers, the person posting the question can elicit their favourite (possibly compound) final answer.

The general wisdom of the crowd approach relies on two crucial aspects to account for its efficacy. The first aspect is the *redundancy* of the responses to the same question. More specifically, since one crowd-worker is usually not capable of consistently providing the correct answer, naturally more workers are needed to work on the same question to produce multiple responses. The second aspect is the *aggregation* of the redundant responses for eliciting an accurate overall answer for the question. Two prerequisites need to be satisfied for the aggregation to work properly. First, the majority of the crowd are reliable in deriving their respective answers by utilizing their expertise and exerting adequate amounts of effort, and in reporting their answers truthfully for the aggregation. Second, there were sufficient responses collected for each question.

In summary, the wisdom of the crowd ensures the *high quality* (or equivalently, low probability of error) of the overall answer, aggregated from multiple worker responses, to each question while also satisfying any other requirements of the requester. The main advantage of the wisdom-of-the-crowd approach over filtering is that it tends to preserve all the worker responses and meanwhile uses aggregation to smooth out the noisy responses and reverse the biased responses for eliciting better final answers to questions to the degrees depending on how sophisticated the aggregation is. Applying either worker-filtering or the wisdom-of-the-crowd aggregation or their combination to crowdsourcing is broadly called *Quality Control for Crowdsourcing* (QCC). The aim of QCC is to control the quality of crowd-workers' responses to the point where their influence on determining the corresponding final answers to the questions should be consistent with their respective quality. Worker-filtering is typically applied prior to or during the crowdsourcing, while wisdom-of-the-crowd aggregation is typically applied during or after the crowdsourcing. As for combining the approaches, the techniques can become sophisticated when the worker-filtering and the wisdom-of-the-crowd aggregation are both applied during the crowdsourcing (e.g. quality control for active learning based crowdsourcing [23]). In this dissertation, we will focus on studying and developing QCC approaches using sophisticated wisdom-of-the-crowd aggregation approaches.

The basic wisdom of the crowd: Majority Vote

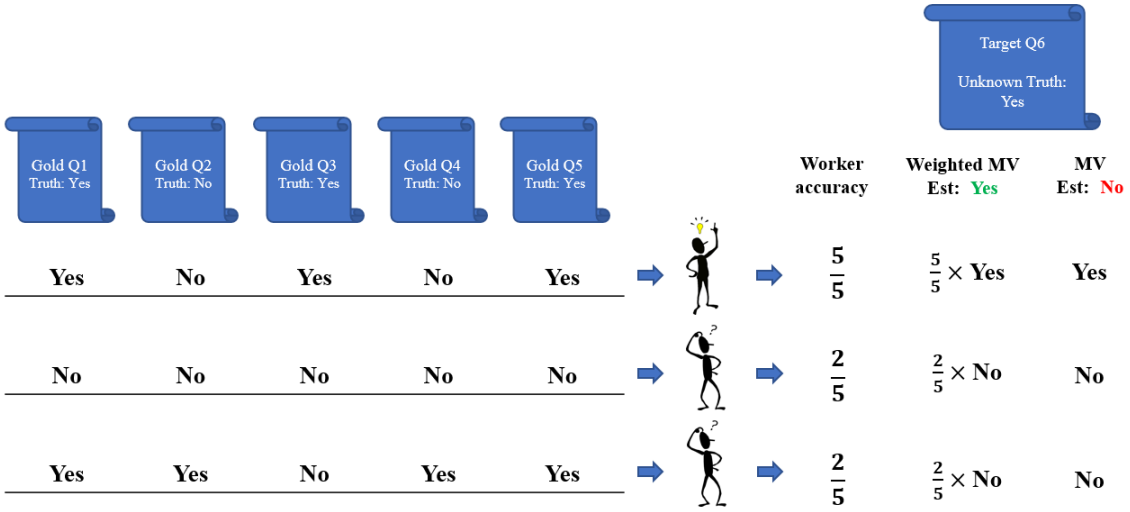


Figure 1.1: An example of how majority vote aggregation can be improved by correlating the quality of each answer with their respective workers.

Based on the wisdom-of-the-crowd aggregation, the simplest and most widely used QCC approach is the *majority vote* (MV). It considers the most likely correct answer to a question to be the answer agreed upon by the majority of the crowd-workers who responded to that question. Provided that the majority of the crowd-workers are reliable, the majority vote response approaches the correct answer to the question as the redundancy of the responses to that question becomes larger.

The majority vote is, however, not robust when only a small number of responses are collected for each question in crowdsourcing. In this case, many of the questions can by chance have the same number (or more) of incorrect responses as correct responses. For instance, if for each question, only two answers are collected, then there can be many unnecessary ties that the majority vote is unable to break. If each question receives three responses, there can be some questions that randomly collect two incorrect and one correct answer, even if the number of correct responses would dominate the number of incorrect ones were the response count large. The vulnerability of the majority vote to small numbers of answers is exacerbated by the fact that it treats the responses to be independent of one another with no consideration of individual response quality. As a result, for each question, the majority vote aggregation only uses the counts of the different responses as their fitness for being the correct response to the question, thereby easily succumbing to the problem of scarce responses per question.

1.2.2 Modelling the Quality of Responses (QoR)

Crowdsourced responses are fundamentally not independent and as mentioned earlier, vary in quality. The former suggests that the quality of a response can be indicated by the quality of other responses to which it is related (e.g. by coming from the same crowd-worker, or same worker context, etc.), while the latter implies that responses with higher quality should be modelled to have greater influence in the wisdom-of-the-crowd aggregation and vice-versa.

Figure 1.1 provides an example that illustrates the above two points. In the example, three crowd-workers each have answered first five control questions ($Q1$ to $Q5$), and now are about to answer the first target question ($Q6$) whose correct answer is "Yes" but is unknown to the requester who therefore turns to the three workers for their responses. In this case, two of the workers respond with the answers "No" while the remaining one answers "Yes". When the majority vote is applied to predict the correct answer for $Q6$, it yields "No" as the prediction which however is incorrect. To obtain the correct answer for $Q6$, we need to have more information than just the counts of the different responses. In this case, since all the workers have answered the control questions from $Q1$ to $Q5$, we can readily obtain additional useful information, specific to each worker, regarding their accuracy on the control questions.

Based on such worker-specific information, we can better determine the likely quality of the questionable responses to the target question $Q6$. A common statistic to calculate is the *correct answer ratio* for each of the workers, which is interpreted as the *accuracy/ability* of a worker. We can simply calculate the ratio for a worker as: $\frac{\#\text{control questions answered correctly}}{\#\text{control questions answered}}$. Figure 1.1 shows the ratio-based accuracy of the three workers (i.e. $\frac{5}{5}$, $\frac{2}{5}$, $\frac{2}{5}$), which indicates that one of them is an expert and the other two are novices. Now we can modify the uniform aggregation employed by the majority vote to take into account the accuracy values of the three workers for weighing their respective answers to $Q6$. As a result, the single "Yes" response from the expert worker still weighs 1, but the two "No" responses from the novices are now both weighed down to 0.4. Then, by comparing the respective sums of the weights on different response options and choosing the option with the largest total weight, the final predicted answer to $Q6$ is now "Yes" which turns out to be the correct answer.

The above example shows the importance of modelling the quality of individual responses as unknown quantities to be estimated. In the example, the quality of the responses to a question were modelled to be the correct answer ratios for their respective workers which were estimated from their remaining responses to the other questions. The ratios are then introduced as the weights of

the corresponding answers into the majority vote aggregation for the correct answer for the question. In this way, the aggregation alleviates the negative effect of *sparse answers* on the question and improves its prediction accuracy for the correct answer to the question.

1.2.3 Correlations between QoR and Crowdsourcing Aspects

In the above example, the statistic used to predict the quality of a response given by a crowd-worker was calculated by aggregating her accuracy over all her remaining responses. In other words, the statistic was obtained by *correlating* the quality of a particular response with the accuracy of the corresponding crowd-worker on the other questions. This correlation allowed the quality of the response to be estimated properly. In general, the quality of responses can exhibit various correlations with different *aspects* of paid crowdsourcing including the crowd-workers themselves, their current contexts, the questions, etc. Encoding these correlations explicitly in the statistical models can potentially refine the estimation of the response quality. Such improved quality estimates can further enhance the wisdom-of-the-crowd aggregation to produce more accurate predictions of question true answers.

Aspects of paid crowdsourcing. To provide a better idea about what are the various correlations that be leveraged to improve true answer prediction, we need to formally model a typical crowdsourcing platform by extracting its prominent aspects. Past research on quality control for paid crowdsourcing model the following four principal aspects of crowdsourcing platforms (as shown by Figure 1.2):

Items/Questions. A question is the smallest unit in a crowdsourcing task which crowd-workers need to respond to and answer. In the literature of crowdsourcing, the term question is usually used interchangeably with *data item* or simply *item*, especially in the context of *labelling/annotating/rating/judging* items using crowdsourcing [24]. In this case, the correct answer to a question is essentially the *true label* of a data item. As for the types of questions that might appear in a crowdsourcing task, there exist the following three types:

- *Objective* question: The question has a single correct answer, that is the single true label for the data item described by the question. For instance, one may label whether the bird in an image is a *duck*, is *not-a-duck* or there is *no-bird-in-the-image* [25].

- *(Purely) subjective* question: All of the available response options for the question are correct and crowd-workers choose the correct answer based solely on their personal preferences,

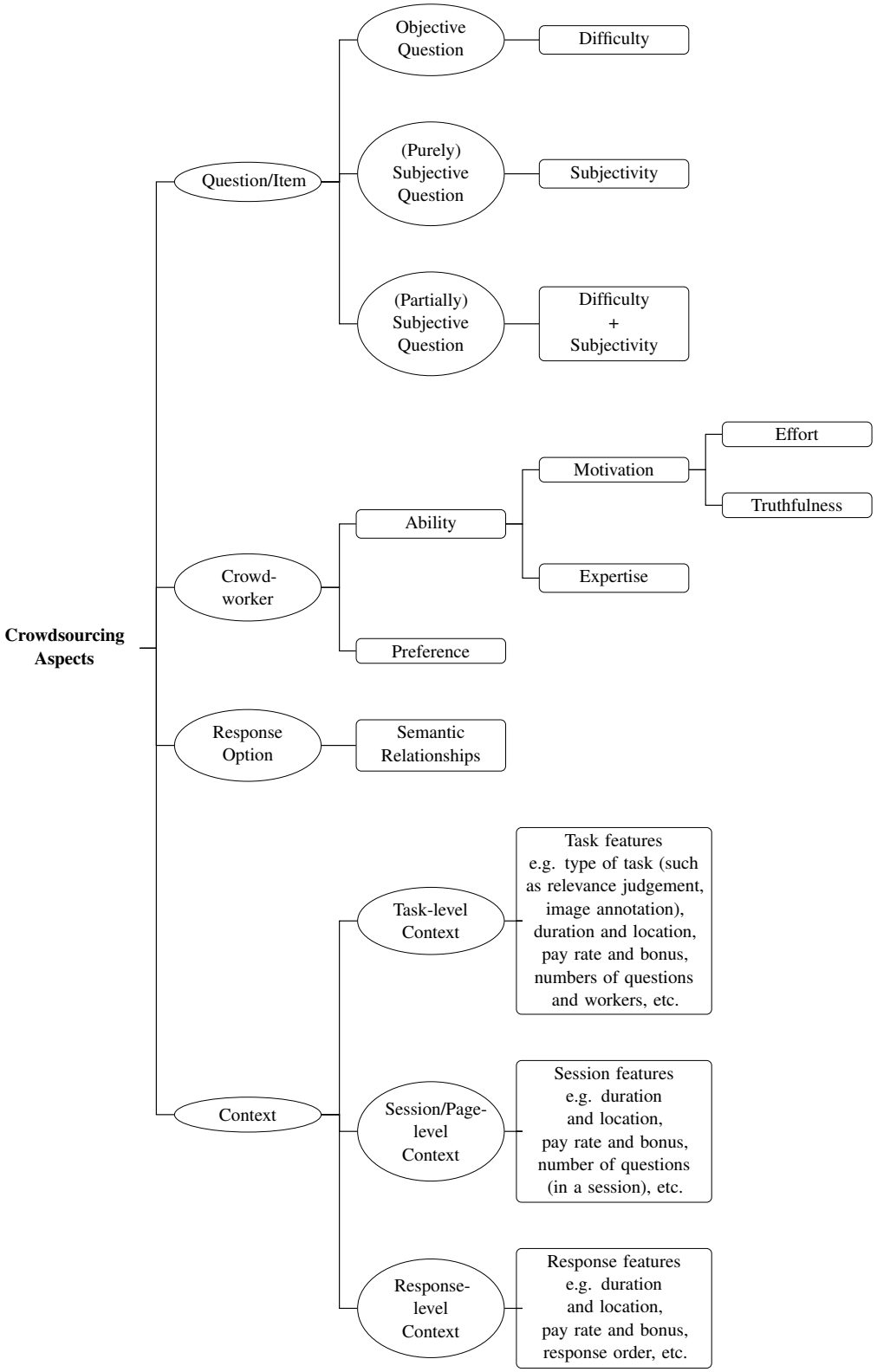


Figure 1.2: Aspects of crowdsourcing applications that have been studied by past research for better controlling the quality of responses. Each ellipse node denotes a particular crowdsourcing aspect. Four major aspects at level 1 are worker, question, response option and context. Question and context have finer definitions at level 2. Rectangle nodes denote prominent attributes of aspects that have been modelled or utilized in the modelling by the QCC methods.

opinions and tastes rather than their expertise (and without the exertion of effort). For instance, asking a crowd-worker whether she prefers comedies or drama films is a purely subjective question.

- *Partially subjective* question: Intuitively, the question has more than one correct answer but not all response options are correct answers [26]. This means there is at least one incorrect response option for the question. For instance, consider the task to judge whether an image contains a person wearing fashionable clothes or not. The possible responses are *yes*, *no* and *no-person-in-the-image*. If there is indeed a person shown in a particular image, then the option *no-person-in-the-image* would be an objectively incorrect answer. Whether the clothes worn by the person in the image are considered fashionable depends on the subjective tastes of the individual crowd-workers.

  As for questions with binary response options, it is not immediately clear as to how they can be partially subjective according to the above definition. However, using intuitive pairwise comparisons, we can judge which question is more subjective. As an example, consider the task of judging whether a Web-page is evergreen or not (i.e. always remaining informative and relevant for average online users in the future or only being of transient interest). Under comparison, a celebrity gossip Web-page is less subjective to judge than a food recipe Web-page. This is because the degrees of appreciation for the recipe from the online public tend to vary much more greatly according to their subjective tastes and familiarity with the recipe.

Both objective questions and partially subjective questions possess certain degrees of *difficulty* which obscure their correct answers from crowd-workers to various extents. According to the literature, a difficult question leads to more variation in the responses across crowd-workers. (For extremely difficult questions, crowd-workers may have to resort to random guessing). A question is *deceptive* when it is difficult enough to bias the majority of the workers towards answers that are not the correct answer. In other words, as the difficulty becomes greater, its induced errors transfer from being random to being systematic.

Crowd-workers. A crowd-worker answers the questions contained in a crowdsourcing task according to the requirements of the requester. As mentioned earlier, this worker has certain *motivation* for answering the questions the way they do, which is *unknown* to the requester and supposedly to the other crowd-workers[7]. The crowd-worker also has a certain level of (*domain*)

[7]The motivation of a crowd-worker can be known to other workers when they form collusion.

expertise required by the subject of the task (e.g. knowledge about tumour texture is required for identifying tumours in images). The motivation governs the levels of *effort* exerted by the worker to answer each question, and also the *truthfulness* of the worker's response for each question. The greater the effort is exerted, the higher the quality of each answer from the worker regardless of her expertise. In other words, when two crowd-workers possess the same level of expertise (in the same domain), the worker motivated to exert more effort is more likely to yield responses of higher quality. The truthfulness of the worker's response determines how faithfully she sticks to the response that she believes correct when responding to the requester. If the worker is malicious, she is likely to give a different response from the one she believes correct (e.g. by deliberately flipping the response to binary questions). Finally, the levels of effort and expertise, and the truthfulness of response determine the accuracy/ability of a crowd-worker which determines the correctness of her response to each objective question.

When encountering a subjective question (e.g. about choosing one's favourite clothing), a crowd-worker would exhibit a *preference* towards a particular type of response (e.g. clothing with dark colours). Such a preference is independent from the accuracy and ability of the worker, thereby having no effect on the correctness of the response. Moreover, the preferences vary according to individual crowd-workers (e.g. some like clothing with dark colours and others prefer that with light colours). From an even finer point of view, preferences of individual crowd-workers can be expressed toward different *features* characterizing the responses. For example, some crowd-workers prefer the dark colours (e.g. grey, black, brown) because of their formality and solemnity, while others prefer bright colours (e.g. red, yellow, green) because of their freshness and liveliness.

Response options. In a crowdsourcing task, a crowd-worker typically answers individual questions by choosing one answer from the same set of response options for each question. Although the set of options can be infinite/continuous (e.g. real values measuring the heights of people), in most cases the set is finite with all the elements being either integer valued (e.g. counts) up to some maximum value, ordinal values (e.g. relevance levels of a document to a query) or categorical values (e.g. yes/no, true/false).

It is common that the set of response options for each question in a crowdsourcing task is small. Even when the set is large, past research has applied the "divide-and-conquer" strategy by breaking the set into many much smaller subsets. A typical example of this is the crowdsourcing for the ImageNet database [8] [27] which stores millions of images according to tens of thousands of object

[8] http://www.image-net.org/

types that are connected in a *semantic relational* graph. To facilitate the image annotation in this case, the originally large set of object types was divided in [27] into individual object types called synsets. Each of these synsets then constitutes the binary response options, that is whether an image contains an object type (e.g. Labrador) or not, for a crowdsourcing task that contains images of either that object type or some other object types (e.g. golden retriever).

The "divide-and-conquer" strategy is less practical outside of research. This is because real-world requesters are often unwilling (and also not obliged) to divide a raw problem with a large set of response options into sub-problems with smaller sets of options, due to the sheer complexity of the dividing process. In practice, requesters are customers who would like to pass such raw problems directly to the crowdsourcing platforms. In this case, these platforms should still be able to provide proper quality control for the responses produced for such problems.

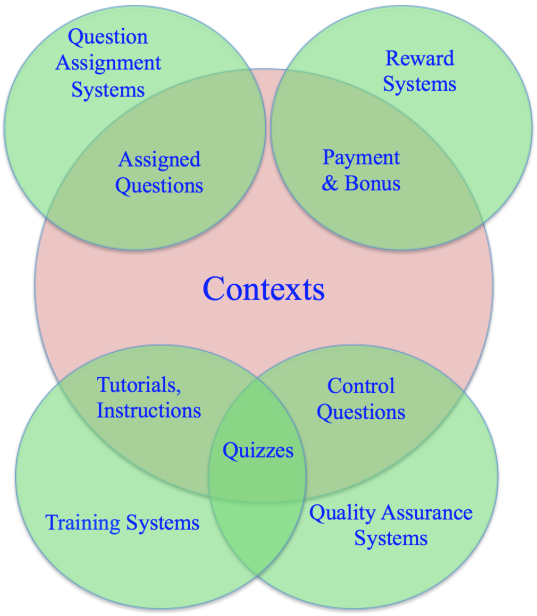


Figure 1.3: The green circles represent the four supporting systems of a crowdsourcing platform: the question allocation system, the reward system, the training system and the quality assurance system. They provide components of worker contexts, illustrated as the intersections with the red circle denoting the contexts. Different settings of these components (e.g. by mechanisms deployed in the systems) change the contexts which further influences workers' behaviour. Note that both the training system and the quality assurance system embed quizzes into the contexts.

Contexts. A context in crowdsourcing is an *environment* in which crowd-workers are situated. Apart from the workers, this environment contains other components which interact with the workers to achieve certain goals of crowdsourcing such as having the workers answer the target questions accurately, rewarding the workers and so on. These components are primarily provided by four important supporting systems of a crowdsourcing platform. They are the *question allocation*

system, the *reward* system, the *training* system and the *quality assurance* system. Figure 1.3 shows the four systems in green circles. The components they provide that constitute a context are the intersected areas in the figure. Considering these components from the perspective of workers situated in such a context, they are:

- provided with *instructions* and *tutorials* that prepare them for the corresponding tasks;
- required to pass some *quizzes* to start the tasks and to answer some *control questions* correctly in order to stay in the tasks;
- assigned with *target questions* they need to answer during the tasks;
- rewarded at certain *pay rates* or *bonus rates* once they finish the assigned questions.

Normally, crowdsourcing platforms allow data requesters to change the settings of the above components before or during the task. There has also been a large body of research in quality control for crowdsourcing which designs *mechanisms* that are deployed in the above four systems to manipulate the status of the components. Such manipulation changes the context/environment which further influences workers situated in it in various ways.

Changing or intervening in the context can affect the motivation of a worker which then affects her question-answering behaviour. For instance, instruction is an indispensable part of the crowdsourcing context that informs crowd-workers of things they need to know in order to perform a task. Suppose that a task requires crowd-workers to classify biopsy images [28] and shows two sets of instructions to two different groups of crowd-workers. The two instruction sets are the same in describing the task except that one tells the workers their contributions are going to be valuable for the study of breast cancer, while the other does not. The first instruction set obviously provides extra information that will affect the motivations of its group of crowd-workers (e.g. inspiring their altruism) differently from the second group.

Moreover, changes to a context can also directly affect the worker's question-answering ability regardless of her motivation. For instance, having the training mechanism tailored for the worker ensures she possess higher levels of expertise before or during crowdsourcing. On the other hand, even though the worker becomes an expert, her motivation can still remain the same as the one prior to the training (e.g. constantly being money-driven or malicious).

Past QCC research has proposed not only to control different components (e.g. payments, training, etc.) of the worker context but also to control them at different *levels* of the context. More

specifically, they studied how changes to (certain components of) the context at different levels of granularity can affect the response behaviour of the crowd-workers. Some investigations looked at crowdsourcing tasks on the same crowdsourcing platform. In these investigations, mechanisms were designed to adapt their quality control to individual tasks but not to any Web-page or response within a task. Statistical models applied to multiple crowdsourcing tasks learn the correlations between the quality of responses and the contextual information about crowd-workers' responses to each task. Such contextual information describes characteristics of a task and its received responses. The former includes a task's duration, location[9], problem domain (e.g. movie rating, image annotation), type of responses (e.g. integers, ordinals, categories), instructions given to crowd workers, and rules that regulate their behaviour and the payment made to them. The latter includes different statistics that summarize crowd-workers' responses to individual tasks, such as micro-averaged and macro-averaged worker response times for each task.

In these investigations, the types and the effects of the quality control mechanisms tend to be similar when the involved tasks and their received responses are similar with respect to the contextual information. For example, tasks whose problem domains concern innovation and creativity (e.g. logo designs) tend to be offered similar methods and degrees of quality control on the responses. They are different from the methods offered to tasks asking crowd-workers to follow certain routines (e.g. counting or identifying certain objects in images).

Other investigations focus on a single task and were conducted at finer granularities for contexts within the task. More specifically, two common types of within-task contexts involve details about workers' responses to individual task pages and about their individual responses.

For the first type of contexts, the contextual information is confined to individual task pages each crowd-worker has finished. It describes a process in which a crowd-worker responds to all the questions on a task page. The process starts from the first time the page is presented to the worker and ends when it is finished and submitted by the worker. Therefore, the contextual information is contained by features relevant to the task page and the page-response process. The features for a task page include the pay rate for its questions, their topics, their total word counts, the total number of clicks on the radio buttons for the questions' response options and so on. The features regarding the page-response process include the average response time a crowd-worker spent answering each question on the page, the total response time of the worker, the device (e.g. PC, mobile phone) and

[9]Knowledge about task locations is common in mobile crowdsourcing [29].

the browser (e.g. Chrome) she used, the time period of the process (e.g. hour of the day, day of the week).

For the second type of contexts, the contextual information is specific to each response made by a crowd-worker. It commonly describes the time duration of a response and its specific place in the sequence of all the worker's responses. Correspondingly in those investigations, the quality control mechanisms were designed to be aware of the different levels of the contextual information to refine their control effects. This was done by learning the correlations between the quality of answers and the within-task contextual information at the specific levels of granularity.

We summarize the aforementioned three levels of contexts as follows:

- *Task* level: a task-level context consists of the features which distinguish multiple tasks on the same crowdsourcing platform. These features contain information about individual tasks such as the task durations, locations, domains, settings including the pay rates, instructions, minimum accuracy for quizzes and so on. They also contain information about the responses from workers who participated in the multiple tasks with statistics such as micro-averaged and macro-averaged worker response time for a task.

- *Session* level: a session-level context within a task corresponds to a Web-page of the task that a crowd-worker has completed and submitted. We call the process of the worker answering all the questions on that page a *working session* of that worker. A session-level context thus consists of features regarding a task page (e.g. the pay rate for answering each question on the page, their topics and total word counts, number of embedded control questions etc.). They also concern the worker's responses within the corresponding working session (e.g. the average response time on each question, the total response time, the device used for the response and etc.).

- *Response* level: a response-level context represents an even finer level of granularity for the within-task contexts. It corresponds to a single response given by a crowd-worker to a question, and consists of features regarding that response (e.g. its payment, duration, location and position in the sequence of all the responses given by the same worker).

Collecting the task-level contextual information is often harder than collecting within-task information since we usually do not see the same crowd-workers participating in multiple tasks in practice. Additionally, past QCC approaches on utilizing within-task contextual information have

considered either the session-level [30] or the response-level [31] granularity in the information, while studies on combining both types of contextual information are still missing.

1.3 Notation

Before concluding this chapter, we introduce the notation and symbols to be used throughout the rest of this thesis in Table 1.1. The layout of this table is based on the crowdsourcing aspects and their attributes shown in Figure 1.2.

Symbols	Description
Symbols Used across Chapters	
$\lvert \cdot \rvert$	the size of a set
$\hat{}$	the estimate of a parameter/variable
f	a function
$\mathcal{I}$	a set of workers
$\mathcal{J}$	a set of questions
$\mathcal{K}$	a set of response options
$\mathcal{R}$	a set of responses given by workers $\mathcal{I}$ to questions $\mathcal{J}$
$\mathcal{L}$	a set of true answers for questions $\mathcal{J}$
i	the i-th worker or worker i where $i \in \mathcal{I}$
j	the j-th question or question j where $j \in \mathcal{J}$
$\mathcal{I}_j$	a set of workers who answered question j
$\mathcal{J}_i$	a set of questions answered by worker i
k	the k-th response option or option k where $k \in \mathcal{K}$
r_{ij}	the response given by worker i to question j, $r_{ij} \in \mathcal{K}$
l_j	the latent true answer of question j where $l_j \in \mathcal{L}$
q_{ij}	the quality of response r_{ij}, $q_{ij} = P(r_{ij} = l_j)$
f_i	a function specific to worker i
f_j	a function specific to question j
f_{ij}	a function specific to response r_{ij}
$\boldsymbol{\theta}$	the probability/proportion vector over response options $\mathcal{K}$
$\boldsymbol{\gamma}$	the Dirichlet prior parameters over response options $\mathcal{K}$
N_i	the number of responses from worker i

e_i	the expertise of worker i
d_j	difficulty of question j
μ_e, σ_e	the mean and standard deviation of the Normal prior for worker expertise
μ_d, σ_d	the mean and standard deviation of the Normal prior for question difficulty

Additional Symbols Used in Chapter 2

$\mathbf{\Pi}_i$	a $\|\mathcal{K}\| \times \|\mathcal{K}\|$ confusion matrix for worker i that represent *worker biases* towards the set of *response options* $\mathcal{K}$
$\pi_{ikk'}$	worker i's conditional probability of responding option k' given the true answer is k

Additional Symbols Used in Chapter 5

$\mathcal{S}$	a set of sessions each corresponding to a task page that contains a number of questions
s	the s-th session or session i where $s \in \mathcal{S}$
z_{ij}	an auxiliary variable corresponding to r_{ij}
$\boldsymbol{x}_i$	a demographic feature vector for worker i
$\boldsymbol{x}_j$	a feature vector for question j
$\boldsymbol{x}_{ij}$	a feature vector for the response r_{ij}
$\boldsymbol{x}_{is}$	a feature vector for the s-th session of worker i
$\boldsymbol{\beta}^{\mathcal{I}}$	a global coefficient vector that multiples the worker features $\boldsymbol{x}_i$
$\boldsymbol{\beta}^{\mathcal{J}}$	a global coefficient vector that multiples the question features $\boldsymbol{x}_j$
$\boldsymbol{\beta}^{\mathcal{S}}$	a global coefficient vector that multiples the session features $\boldsymbol{x}_{is}$
$\mu_\beta^{\mathcal{I}}, \sigma_\beta^{\mathcal{I}}$	the mean and standard deviation of the Normal prior for each component of vector $\boldsymbol{\beta}^{\mathcal{I}}$
$\mu_\beta^{\mathcal{J}}, \sigma_\beta^{\mathcal{J}}$	the mean and standard deviation of the Normal prior for each component of vector $\boldsymbol{\beta}^{\mathcal{J}}$
$\mu_\beta^{\mathcal{S}}, \sigma_\beta^{\mathcal{S}}$	the mean and standard deviation of the Normal prior for each component of vector $\boldsymbol{\beta}^{\mathcal{S}}$
$\boldsymbol{\eta}_i$	a coefficient vector specific to worker i that multiples her response features $\boldsymbol{x}_{ij}$
μ_η, σ_η	the mean and standard deviation of the Normal prior for each component of vector $\boldsymbol{\eta}_i, \forall i \in \mathcal{I}$

Additional Symbols Used in Chapter 6

$\boldsymbol{\rho}_{ij}$	the response probabilities over $\mathcal{K}$ for generating r_{ij} in collaborative filtering
ρ_{ijk}	the response probability $r_{ij} = k$
$\boldsymbol{\psi}_{ijl_j}$	the response probabilities over $\mathcal{K}$ for generating r_{ij} conditioned on true answer l_j in crowdsourcing
ψ_{ijl_jk}	the response probability $r_{ij} = k$ conditioned on true answer l_j

$\mathcal{M}$	a set of preferences from which workers can choose to exhibit over response options when answering questions
m	the m-th preference or preference m
$\boldsymbol{\phi}_i$	the preference selection probabilities of worker i over $\mathcal{M}$
$\boldsymbol{\lambda}$	the Dirichlet prior parameters for the preference selection probabilities
z_{ij}	the preference selected by worker i according to $\boldsymbol{\phi}_i$ to determine the extent to which worker i will prefer each response option as the correct response to question j, $z_{ij} \in \mathcal{M}$
u_{mk}	the extent of the m-th preference over the k-th response option
v_{jk}	the extent to which the k-th response option is preferred to be the correct answer to question j
$\mathcal{C}$	a set of worker clusters obtained from the preference selection probabilities $\boldsymbol{\phi}_i$ of each worker
c	the c-th worker cluster or cluster c

Additional Symbols Used in Chapter 7

$\text{vec}(\cdot)$	the vectorization function for matrices
$\boldsymbol{S}$	a latent symmetric matrix capturing semantic relatedness/similarity between response options in $\mathcal{K}$
$s_{kk'}$	an off-diagonal entry in $\boldsymbol{S}$ capturing the relatedness of option k' to true answer k
μ_s, σ_s	the mean and standard deviation of the Normal prior for each off-diagonal entry $s_{kk'}$ where $k < k'$
$\boldsymbol{X}$	an observed real-valued relatedness matrix counterpart to $\boldsymbol{S}$
$x_{kk'}$	an off-diagonal entry in $\boldsymbol{X}$ capturing the relatedness of option k' to true answer k
β	a global coefficient to be multiplied by $x_{kk'}$

Table 1.1: List of Notation used in this thesis.

1.4 Conclusion

In this chapter, we started by introducing crowdsourcing and especially the paid crowdsourcing with its online platforms such as Amazon Mechanical Turk and CrowdFlower. We then articulated the importance of controlling the quality of responses from crowdsourcing and described two ways of quality control: worker-filtering and wisdom-of-the-crowd aggregation.

We concentrated our studies on the latter technique and showed the strength and weakness of the most basic wisdom-of-the-crowd approach: the majority vote. The weakness that the majority vote ignores the difference in the quality of responses led us to modelling the response quality and its correlations with four major aspects of a crowdsourcing application. They are the workers, the questions, the contexts and the response options. We described each of these aspects by specifying their possible sub-categories and attributes considered by the past research in quality control for crowdsourcing.

More specifically, we discussed three categories of questions: objective, subjective, and partially subjective questions, and whether they possess difficulty. We analysed attributes that determine the question-answering behaviour of workers, including their motivation (which governs efforts and honesty), expertise and preferences. We also emphasized the significance of semantic relationships between response options in indicating correlations within responses.

We defined what a worker context is in crowdsourcing and what are its components, followed by discussion on how controlling these components can affect workers' behaviour. We also found that past QCC research controlled the components at three levels of granularity: the task-level, the session-level and the response-level. We proposed a formal definition for each of the levels with examples for contextual features that describe them.

Finally, we specified a table of notations to be used throughout the thesis based on our discussions about the four major crowdsourcing aspects and their attributes. We provided descriptions for all the symbols in the table.

In the next chapter, we will conduct a comprehensive literature review for quality control in crowdsourcing. Doing so will help us identify the research gaps and corresponding research directions in this area.

Chapter 2

Literature Review

This chapter provides readers with a comprehensive literature review on how previous research in crowdsourcing models the quality of responses (QoR) from individual crowd-workers. The review is structured according to the four crowdsourcing aspects introduced in Section 1.2.3. A summary of the review is given at the end of the chapter with identification of the current research gaps in quality control for crowdsourcing (QCC) and proposal of future research directions that aim to solve these issues.

A variety of literature reviews in the area of crowdsourcing have been written in the past. The subjects of these reviews include general overviews of crowdsourcing [39, 40], management of certain components of crowdsourcing platforms, such as the routing and recommendation of tasks [41, 42], and different applications of crowdsourcing, such as information retrieval [43], software engineering [44], data mining [45], health and medicine [46], music [47] and neogeography [48].

With the development of sophisticated quality control mechanisms in recent years, surveys specifically regarding QCC research have started to appear. Existing surveys on quality control for crowdsourcing mainly fall into two areas: the *design* of crowd-sourcing applications and *statistical methods* for estimating answer quality, which have been summarized in Table 2.1.

Design for quality control purposes in crowdsourcing involves changing existing modules of crowdsourcing tasks, or adding new modules into the tasks, which intervene reactively or proactively in the crowdsourcing process to augment quality of responses (as shown by the work flow inside the red box in Figure 2.1). A task module is any system/program/script which runs to affect behaviour of the workers. For example, a payment (and bonus) system is a common module in crowdsourcing tasks which controls when and how much crowd-workers should be rewarded. Meanwhile, a badge system is a module that gets frequently added to tasks under *gamification*.

Surveys	Areas	Details	Pros	Cons
[32, 33, 34]	Designs for Quality Control	Review of designs for tasks/measures /interfaces that facilitate: - Assessment of response quality (e.g. using quiz, expert/peer review, personality test [35] etc.); - Assurance of response quality (e.g. by worker filtering, selection, training, team work, etc.).	Diverse design problems and strategies related to measuring and controlling response quality are covered with brief descriptions.	Lack of technical details about how statistical methods are used/integrated in the designs for various crowdsourcing applications. Lack of taxonomies on papers in the area.
[36, 37, 38]	Statistical Methods	Review of statistical models for QCC which estimate: - Worker Ability/Expertise - Question Difficulty - Question True Answer - Response Correctness Probability	Technical details about a variety of statistical models are specified. They include model assumptions,variables for crowdsourcing aspects and attributes considered, parameter estimation, etc.	Lack of design information on crowdsourcing applications involving statistical methods to estimate response quality, worker ability, question difficulty and the true answer; Ignoring aspects other than worker and question which also affect response quality: - Context (in which workers are situated, e.g. time, location, session) - Response Options (their semantic relationships) Lack of taxonomies on papers in the area.

Table 2.1: A summary of surveys regarding quality control for crowdsourcing. There are two types of surveys focusing on either the design of crowdsourcing applications or statistical methods. Their strengths and weaknesses have been specified in the table.

Current crowd-sourcing design surveys [32, 33, 34] mainly review *general-purpose* design strategies for building the user interfaces, interaction mechanisms and response quality control measures employed in various task modules, as opposed to *ad-hoc* designs for specific crowd-sourcing applications. For example, all the three surveys described two lines of incentive designs: *extrinsic* incentive designs (e.g. reward-driven by customizing the payment system) and *intrinsic* incentive designs (e.g. enjoyment-driven by adding and customizing a badge system), without diving into application details. Most of the quality control designs require estimates of certain properties of workers or questions as their inputs for triggering actions or responses as prescribed by the particular crowd-sourcing design mechanism. For example, an estimate of response correctness could serve as the input to a reward system to trigger a payment action according to a prescribed threshold set by the incentive design. Worker ability and question difficulty can serve as the input to the worker selection and question assignment system to trigger the corresponding actions according to the designed mechanism. In both cases, statistical methods are needed to provide the attribute estimates. The main weakness of current surveys for quality control designs is that they barely provide any technical details about the statistical methods and how the methods support the designed mechanisms.

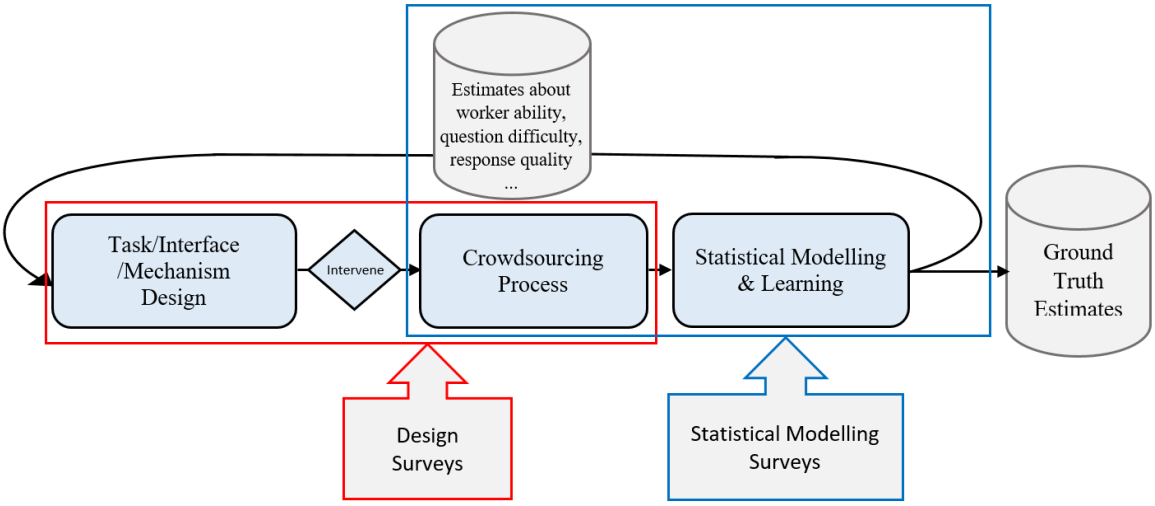


Figure 2.1: A diagram that shows the respective parts played by designs and statistical models in quality control for crowdsourcing. The red box shows the part of QCC designs which aim to intervene in crowdsourcing to improve response quality. The blue box shows the part of statistical models which yield estimates of true answers and other attributes. Each part has its dedicated surveys but there lacks a current survey that links the statistical models with the designs.

Statistical methods for quality control purposes involve estimating response quality and other attributes of important crowdsourcing aspects that are assumed to be correlated with the response quality. The estimates are computed based on the responses collected during or after crowdsourcing. The work flow inside the blue box in Figure 2.1 shows the relationship between statistical methods and the crowdsourcing process. One of the outputs of the statistical methods is the estimate of the ground truth for questions contained in the tasks. Other outputs are those estimates, mentioned earlier, that are required by the quality control mechanisms.

Current surveys of statistical methods in crowd-sourcing [36, 37, 38] focus on the algorithms and inference procedures that model and learn attributes of two crowdsourcing aspects: the crowd-worker and the question. Table 2.1 shows the worker and question attributes that have been reviewed by these surveys, namely: worker ability/expertise, question difficulty and true answer probability. There are other worker and question attributes that might affect or indicate response quality and thus have been modelled by past QCC research. For example, worker *effort* and *honesty* are attributes that have been frequently modelled by game-theoretic methods used for incentive designs [49, 50]. Questions that might contain more than one correct answers (thereby suggesting *subjectivity*) have also been studied and modelled by Nguyen et al. [26]. Current surveys however ignore quality control methods that have modelled and learned these attributes.

Apart from workers and questions, there are other aspects in crowdsourcing that also affect or indicate response quality. The *context* in which each worker is situated is such an aspect. It

consists of features such as the time, location and the session/page being read, etc. Modelling and learning user context for predicting user ratings on subjective items has been thoroughly studied in recommender systems [51]. In a similar set-up, worker context in crowdsourcing has also been modelled and learned for quality control purposes [31, 52, 53]. Another aspect is the *response options* from which workers choose to answer questions. When the set of options is finite and large, their *semantic relationships* might become very useful for indicating the responses given to each question according to recent QCC research [54, 55]. Our survey endeavours to review all the QCC research that deals with these aspects which have not been adequately covered by the past surveys.

Finally, neither the quality control design surveys nor the statistical modelling surveys include taxonomies of QCC papers with respect to methods for quality control (e.g. modelling assumptions, parameter estimation techniques, design features, etc.). The taxonomies are effective means of organizing and describing QCC papers. They clarify what has and has not been done in the area of quality control designs and statistical modelling. Thus, providing them facilitates future QCC research in terms of identifying potential research gaps and new directions.

Contributions. Overcoming the weaknesses of the current QCC surveys, our survey makes the following contributions. We provide:

- A unified taxonomy of the key aspects of crowdsourcing considered by quality control designs and statistical methods to determine response quality, along with their attributes that have been modelled by the two types of methods. Most of these aspects and attributes are considered by a QCC survey for the first time.

- A graph visualization of all the QCC research in which the nodes represent the crowdsourcing aspects, attributes and their combinations that have been considered by the surveyed research. The graph contains directed edges with acyclic paths indicating different lines of research.

- A systematic review of all the QCC research based on the proposed graph. The investigation starts from the most basic work, which considered only the crowd-worker aspect and its attributes, and finishes with the most sophisticated work, which considered multiple aspects and different combinations of attributes.

- Heirarchical categorisation of QCC papers in the areas of quality control designs and statistical methods. These hierarchies are constructed according to the method in which quality control is conducted (e.g. aspects considered, modelling assumptions, parameter estimation techniques, design features, etc.).

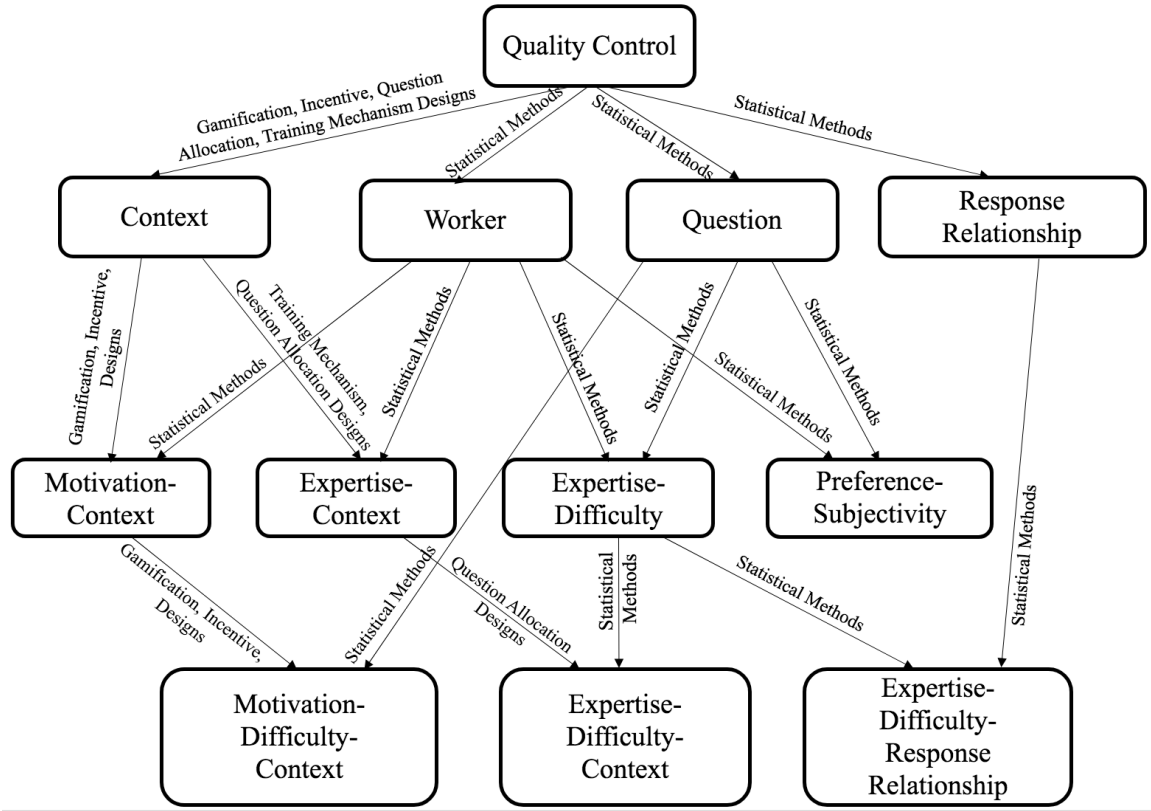


Figure 2.2: A graph visualization of the past QCC research. Vertices (except the root) denote either crowdsourcing aspects, key attributes or their combinations. Paths indicate different lines of QCC research that considered different (combinations of) aspects and attributes. Tags on the edges correspond to the types of QCC methods studied by the particular lines of research.

- An in-depth analysis of current QCC research and identification of research gaps along with the proposal of future research directions.

2.1 Survey Outline

QCC research considers the quality of responses to be dependent on the (attributes of the) crowdsourcing aspects introduced in Section 1.2.3. The QCC methods proposed by the surveyed research encodes the considered aspects and the associated assumptions about their dependency into either mechanism designs or statistical models. To communicate these intricate relationships more effectively, we use a graph visualization over the aspects, their key attributes and the QCC methods under mechanism designs and statistical modelling.

2.1.1 A Graph Visualization on Past QCC Research

This survey is organized according to the graph shown in Figure 2.2. This graph captures the crowdsourcing *aspects*, their key *attributes* and the past QCC research that has considered them

jointly for developing control mechanisms and statistical models. This is a *directed acyclic* graph in which the root node "Quality Control" has outgoing edges to the four major crowdsourcing aspects introduced in Section 1.2.3. Nodes at the second level correspond to *pairs* of attributes which have been jointly considered by some of the past QCC research.

Likewise, each node at the third level involves a *triplet* of attributes. Research considering the triplets of attributes is usually the most sophisticated in terms of the assumptions made. Labels on the edges describe the methods that have been developed by the past QCC research for making use of the crowdsourcing aspects. For example, mechanisms such as *gamification, payment, question allocation*, and *training* design, all make use of the "Context". Edges labeled "Statistical Methods", indicate that the corresponding research focuses on statistical modelling and inference of the attributes.

Edges lower in the graph extend the methods and mechanisms appearing higher in the graph. For example, the edge from node "Context" to node "Expertise-Difficulty-Context" denote the line of research that designed question allocation mechanisms by considering more aspects (e.g. question and its difficulty). Also note that the label "Training Mechanism" does not pass down to the bottom level. This means the research in this regard has not yet considered question difficulty.

Note that edges merging at a particular node indicate the combination of the corresponding QCC methods. For instance, the edges merging at at node "Motivation-Context" indicate the research considering worker motivation and contexts that combines mechanism designs with statistical methods.

2.1.2 Sections of this Survey

We carry out the rest of the survey according to the proposed graph. In Section 2.2, we review the QCC methods which only consider *worker expertise* or *ability*[1] to account for the quality of responses. According to the node "Worker" in the graph, these methods are all *statistical models* which encode the worker's ability and the question's true answer as latent variables. The key assumption that workers yield responses of different quality due to differences in their abilities has been adopted by most QCC methods including those that consider other crowdsourcing aspects.

In Section 2.3, we review QCC methods which consider both *worker expertise* and *question difficulty* to account for response quality. These methods are *statistical models* which additionally

[1]The research in this case ignored the motivation of workers.

encode question difficulty as they assume more difficult questions tend to lower the quality of workers' responses.

In Section 2.4, we review the few QCC methods which suggest that when questions are *partially subjective*, question *subjectivity* can cause workers with different *preferences* to give different (but equally correct) responses to the same question. Such methods involve *statistical models* that encode question subjectivity and worker preferences.

In Section 2.5, we review QCC methods which motivate workers to work harder by controlling their *contexts*. These methods can be further categorized into two types: *payment* mechanisms and *gamification* mechanisms. From Section 2.5.1 to Section 2.5.3, we review the payment mechanisms which incentivize workers by remunerating them more effectively. From Section 2.5.4 to Section 2.5.7, we review the gamification mechanisms which incentivize workers by making their tasks more enjoyable or meaningful. Both of these mechanisms base their designs on statistical modelling and in some cases the estimation of worker attributes, such as worker effort and expertise.

In Section 2.6, we review QCC methods which consider both worker *expertise* and *context*. The various methods make different assumptions about how these aspects interact, and can be divided into two categories: *training* mechanisms and *question allocation* mechanisms. In Section 2.6.1, we review the training mechanisms which assume tailored training can improve workers' expertise. From Section 2.6.2 to Section 2.6.4, we review the question allocation mechanisms which leverage workers' expertise to perform more cost-effective question selection. Both categories require statistical modelling and possibly the estimation of worker expertise to support the mechanism designs.

In Section 2.7, we review QCC methods which incorporate *question difficulty* into the interaction between *worker motivation* and *context*. These methods extend the payment and the gamification mechanisms in Section 2.5, with the aim to incentivize workers more effectively by considering the *heterogeneous* nature of their answered questions.

In Section 2.8, we review QCC methods which consider *worker expertise*, *question difficulty* and *context* together. Such methods can be categorized into two types: *statistical models* and *question allocation mechanisms*. The statistical models encode the three aspects as latent variables, and assume that the expertise varies across different contexts. These models are reviewed from Section 2.8.1 to Section 2.8.3. The question allocation mechanisms incorporate question difficulty into their criteria of selecting the right questions for workers. These mechanisms are reviewed in Section 2.8.4.

In Section 2.9, we review QCC methods which consider *worker expertise*, *question difficulty* and *semantic relationships between response options*. These methods are statistical models which assume that options more related to the true answers are more likely to be selected by workers.

In Section 2.10, we conclude the survey by summarizing the major limitations we found in current QCC methods, and proposing corresponding future directions for the QCC research.

2.2 Modelling Worker Ability

Modelling the effect that individual crowd-workers have on the quality of responses (QoR) has been most widely adopted by quality control methods in crowdsourcing. It also serves as the foundation for more complicated modelling of the correlations between QoR and other crowdsourcing aspects. In this case, the quality of responses is modelled to be fixed for each crowd-worker and varies across different workers. In other words, the quality of each answer from a worker is simply represented by the ability/expertise of that worker.

2.2.1 The Dawid & Skene Model

The work that first introduced a model for estimating worker ability was that of Dawid & Skene [56], which we refer to as the **DS** model for the rest of the thesis. The DS model deals with the scenario where a crowd-worker i reads a question j which has underlying true answer l_j, and then gives her answer/response r_{ij} to the question. Both the true answer l_j and the response r_{ij} are members of a finite set of options $\mathcal{K}$ which is the same for each question (i.e. all questions take the same yes/no responses or the same set of categorical values). In this work, the crowd-worker i is modelled by a $|\mathcal{K}| \times |\mathcal{K}|$ confusion matrix $\mathbf{\Pi}_i$ where $|\mathcal{K}|$ is the size of the option set $\mathcal{K}$. Each diagonal entry of the matrix π_{ikk} records the probability of a response from worker i being correct (equal to option k given the correct answer for the question to which the worker responded is indeed k): $P(r_{ij} = k|l_j = k) = \pi_{ikk}$. Each off-diagonal entry $\pi_{ikk'}$ records the probability of the response being incorrect and equal to option k' given the correct answer is option k: $P(r_{ij} = k'|l_j = k) = \pi_{ikk'}$. Since the k-th row of the confusion matrix delivers conditional probabilities, the entries must sum to one $\sum_{k' \in \mathcal{K}} \pi_{ikk'} = 1$. Based on this confusion matrix for worker i, the quality q_{ij} of her answer r_{ij} to the question j whose true answer is l_j can be represented by the probability $\pi_{il_jr_{ij}}$. This probability has the following interpretations:

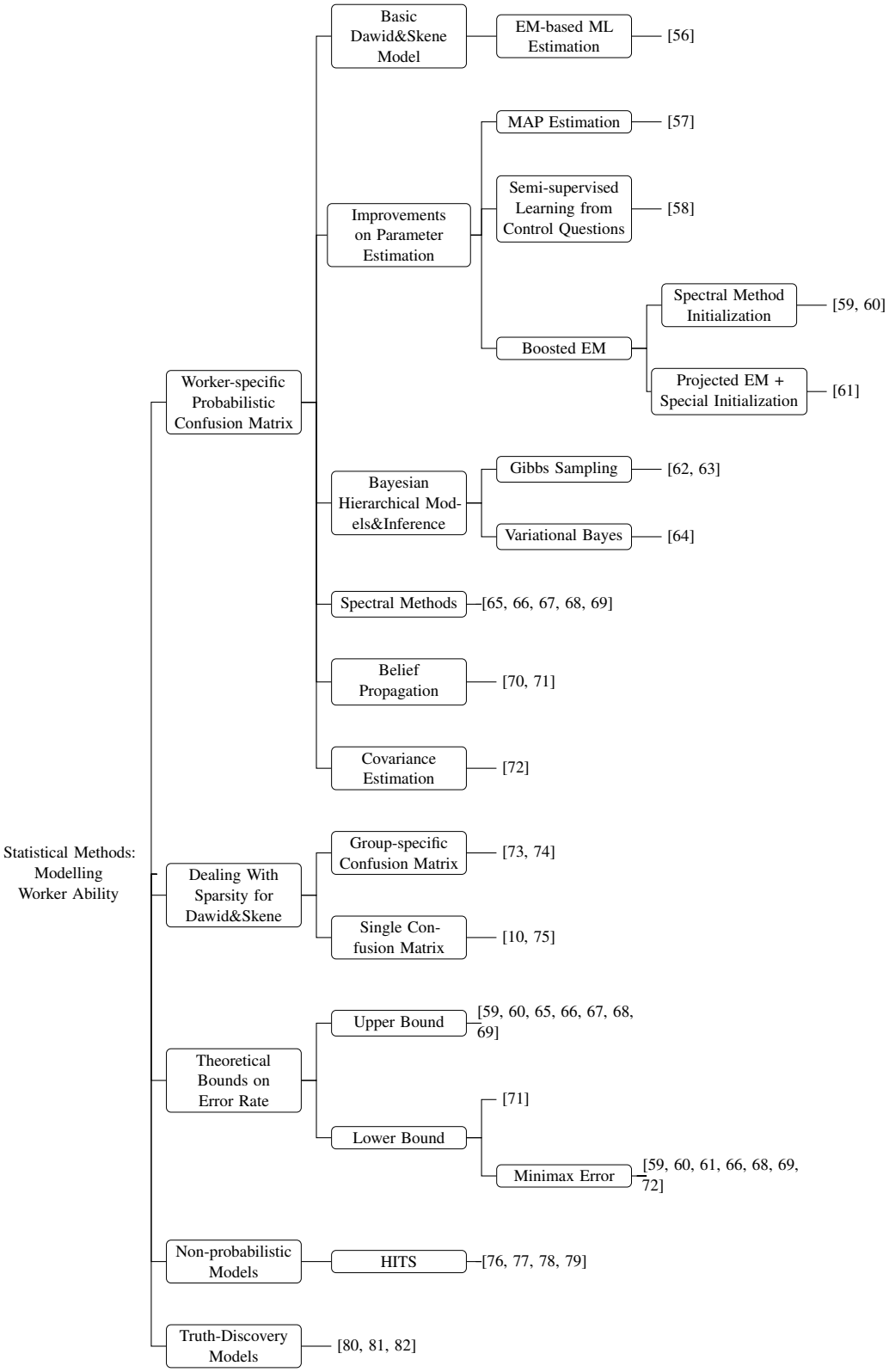


Figure 2.3: A taxonomy of QCC papers that only considered worker ability/accuracy/expertise to account for response quality. These papers focused on statistical modelling and inference.

- if $\pi_{il_jr_{ij}}$ is on the diagonal of the confusion matrix with $l_j = r_{ij}$, the higher the probability $\pi_{il_jr_{ij}}$ is, the higher the quality of each response.

- if $\pi_{il_jr_{ij}}$ is off the diagonal of the confusion matrix with $l_j \neq r_{ij}$, the probability $\pi_{il_jr_{ij}}$ stands for a bias (or equivalently, a systematic error) of the worker i to respond with k' when the true answer is actually k. The higher the probability $\pi_{il_jr_{ij}}$, the larger the bias becomes.

- if all the entries in the k-th row of the confusion matrix are *equal* that is: $\pi_{ikk} = \pi_{ikk'} = \frac{1}{|\mathcal{K}|}, \forall k, k' \in \mathcal{K}$ and $k' \neq k$, the quality q_{ij} of the response r_{ij} is the same as that of an arbitrary guess or a random error.

The DS model adopts the expectation-maximization (EM) algorithm to perform maximum likelihood (ML) estimation over all the worker responses, in order to find the locally optimal estimates for the model parameters: both the probabilistic entries in the worker-specific confusion matrices and the true answers for the questions. In this case, the wisdom-of-the-crowd aggregation for inferring the true answers of the questions is essentially integrated into the EM estimation process.

The EM algorithm comprises two alternate steps which are iterated until the convergence of the likelihood. In the E-step, for each question j, the DS model estimates the probability of the true answer l_j being equal to each possible category $k \in \mathcal{K}$ given the current estimates $\hat{\boldsymbol{\pi}}_{ik}$ for the entries in the k-th row of the confusion matrix $\hat{\mathbf{\Pi}}_i$ as:

$$P(\hat{l}_j = k|\mathcal{R}_j, \{\hat{\boldsymbol{\pi}}_{ik}\}_{i\in\mathcal{I}_j,k\in\mathcal{K}}) = \hat{\rho}_{jk} = \frac{\left(\prod_{i\in\mathcal{I}_j}\prod_{k'\in\mathcal{K}}(\hat{\pi}_{ikk'})^{\mathbb{1}\{r_{ij}=k'\}}\right)P(\hat{l}_j = k)}{\sum_{k''\in\mathcal{K}}\left(\prod_{i\in\mathcal{I}_j}\prod_{k'\in\mathcal{K}}(\hat{\pi}_{ik''k'})^{\mathbb{1}\{r_{ij}=k'\}}\right)P(\hat{l}_j = k'')} \tag{2.1}$$

In equation 2.1, $\hat{l}_j$ is the estimate of the correct answer l_j to the question j; $\hat{\rho}_{jk}$ is the estimate of the probability ρ_{jk} of the correct answer $l_j = k$; $P(\hat{l}_j = k)$ is the estimate of the prior probability of $l_j = k$; $\mathcal{I}_j = \{i|(i \in \mathcal{I}) \wedge (r_{ij} \neq ?)\}$ is the set of workers who have answered the question j, with "?" denoting a missing value; $\mathcal{R}_j = \{r_{ij}|i \in \mathcal{I}_j\}$ are their responses; $\mathbb{1}\{...\}$ is the indicator function. In the M-step, the algorithm estimates the rest of its parameters given the current estimates $\hat{\boldsymbol{\rho}}_j = \{\hat{\rho}_{jk}\}_{k\in\mathcal{K}}$ for the probabilities of the true answer l_j:

$$\hat{\pi}_{ikk'} = \frac{\sum_{j\in\mathcal{J}_i}\hat{\rho}_{jk}\mathbb{1}\{r_{ij} = k\}}{\sum_{k'\in\mathcal{K}}\sum_{j\in\mathcal{J}_i}\hat{\rho}_{jk'}\mathbb{1}\{r_{ij} = k'\}} \tag{2.2}$$

$$P(\hat{l}_j = k) = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\hat{\rho}_{jk} \tag{2.3}$$

In equation 2.2, $\mathcal{J}_i = \{j|(j \in \mathcal{J}) \wedge (r_{ij} \neq ?)\}$ is the set of questions that have been answered by worker i. In equation 2.3, $|\mathcal{J}|$ is the total number of questions.

There has been simplification in some of the subsequent worker in quality control for crowdsourcing, (sometimes without loss of generality), to only consider binary response options for all the questions. This means the confusion matrix specific to each worker can be reduced to only the two free parameters (i.e. the diagonal entries) since the off-diagonal entries are simply one minus the diagonal entries. Such a simplified DS model is called a *two-coin* DS model [83]. If the two diagonal entries (specific to each worker) in this model are assumed equal (i.e. the error probability is independent of the true answer), the model is further simplified into a *one-coin* model [59].

2.2.2 Improvements on Parameter Estimation

As an improvement over the maximum likelihood EM estimation performed for the DS model, Snow et al. [57] employed MAP estimation for the parameters. Later, Tang and Lease [58] leveraged gold-standard control questions for improving the maximum likelihood estimation of the DS model. This was achieved via semi-supervised learning from the true answers of the control questions to boost the EM estimation from the responses to the target questions. Zhang et al. [59, 60] proposed to use spectral methods to initialize the EM algorithm to escape local optimum in the search for the optimal probabilities of the question true answers and for the ability parameters of the crowd-workers. Gao and Zhou [61] modified the M-step of the EM algorithm using a projection strategy which acts as an alternative to prior distributions over worker abilities to prevent EM estimation from over-fitting. The authors also devised a customised initialization procedure for the projected EM to avoid falling into a local optimum.

Instead of point estimation based on EM, Kim and Ghahramani [63] and Carpenter [62] applied their respective Bayesian treatments for building hierarchical DS models and used Gibbs sampling to infer posterior distributions for the models' parameters. Preserving the same Bayesian hierarchical frameworks, Simpson et al. [64] applied variational Bayesian inference to efficiently estimate the joint probabilities of the worker-specific confusion matrices and the true answers for the questions. The authors further extended the frameworks to take into account the time evolution of the worker confusion matrices and adapted the variational Bayesian approach accordingly.

Ghosh et al. [65] proposed a spectral algorithm that decomposes a matrix capturing correlations between questions with respect to responses to learn workers' abilities and questions' true answers. The algorithm works with the one-coin DS formulation and requires the existence of one expert

worker and that every worker answers every question. Removing the last two constraints, Dalvi et al. [67] proposed spectral methods that focus on matrices capturing correlations between workers. Karger et al. [66] proposed to apply spectral decomposition explicitly to the worker-question (response) matrix.

Karger et al. [68, 69] combined the previous spectral analysis with belief propagation algorithms which achieved nearer optimality for true answer estimation. Liu et al. [70] applied full belief propagation to both one-coin and two-coin DS models with proper prior distributions assumed over worker ability variables. In comparison, Ok et al. [71] proposed a practical belief propagation algorithm which works on the one-coin DS model and does not rely on choice of prior distributions over worker abilities. Recently, Bonald and Combes [72] proposed a clever algorithm that utilizes covariances between responses of source crowd-workers to estimate the abilities of some target workers.

2.2.3 Dealing with Sparsity in DS

The main drawback of the worker-specific confusion matrix modelling in the DS model is its vulnerability to sparsity in the responses from workers. When the number of answers per worker is small, the confusion matrix for each worker cannot be estimated reliably. To make matters worse, if the number of response options is also large, each confusion matrix will be massive (quadratic in the number of response options), and estimating each matrix will most certainly result in overfitting to the sparse worker responses. To solve this problem, the sparse response information from individual workers needs to be combined and smoothed so that the overall response information is sufficient for reliable estimation of each confusion matrix. Venanzi et al. [73] applied Bayesian hierarchical modelling to infer clusters of workers, called “communities”. The model allows for combining noisy response information across workers, such that the confusion matrix for each worker is smoothed based on the cluster to which the worker belongs. The model is parametric in the sense that it requires users to set up the number of communities in advance.

In Moreno et al. [74], the authors proposed both the Bayesian non-parametric modelling alternative and its hierarchical extension to enable more flexible partitioning of the workers into communities. The number of clusters is learned jointly with the confusion matrices for the communities and the individual workers.

Instead of determining a number of worker communities, Liu and Wang [75] and Kamar et al. [10] have developed statistical models for an extreme case where the response information

from the individual worker matrices is merged to form a single confusion matrix specific to the entire worker population. This confusion matrix is then balanced against the confusion matrix for each worker to smooth out its noisy information.

2.2.4 Theoretical Bounds on Error-Rate of DS Estimation Techniques

A different line of work has investigated bounds on the error (convergence) rates of various parameter estimation algorithms employed for learning the DS model as the redundancy of crowd-sourced responses increases. Among them, Ghosh et al. [65] first derived the upper bound for the error rate of a spectral inference method for true answer prediction. The technique considered crowdsourced binary responses under the one-coin DS model and assumed that each crowd-worker has answered a large number of questions. Based on the same setting, Gao and Zhou [61] showed that the global maximum likelihood estimator follows a minimax lower bound with respect to the error rate, and proposed a projected EM algorithm that was shown theoretically to achieve nearly that rate.

Changing the setting by allowing each worker to answer just a few (rather than many) questions, Zhang et al. [59, 60] proved their proposed EM with spectral method initialization yielded a tighter upper bound than that of Ghosh et al. [65] and was faster to achieve the minimax error rate than Gao and Zhou [61]. Later, Bonald and Combes [72] showed that their covariance estimation algorithm can match an even stricter lower bound on the minimax error rate employing the EM algorithm. Karger et al. [66] proved that when each worker provides only a few responses, their proposed framework based on low-rank spectral decomposition yielded a strict upper bound on the error rate. Meanwhile, they proved the framework matched a lower bound on the minimax error rate that could only be achieved by the best possible question assignment with an optimal true answer inference algorithm. Later, Karger et al. [68, 69] showed their framework based on belief propagation methods yields a tighter upper bound than Karger et al. [66] and the same lower bound on the minimax error rate. The same strict upper bound was also achieved by the framework based on spectral methods proposed by Dalvi et al. [67]. In Ok et al. [71], the authors proved that their framework based on belief propagation is able to achieve the tightest possible error-rate lower bound under the same setting with an additional requirement that each worker is assigned at most two questions. Recently, Gao et al. [84] established both the lower and the upper bounds of the error rates that match exactly the exponential rates under the setting in Karger et al. [68, 69].

Despite their theoretical soundness and empirical feasibility, state-of-the-art work in QCC error rate analysis has seldom relaxed the binary-response assumption and the one-coin DS modelling assumption. For works relaxing the binary-response assumption, we have only found that of Karger et al. [85] whose inference framework adopted the same setting as Karger et al. [68, 69] but extended the binary response options to multiple response options and proved that a tight upper bound on the error rate can still be reached using proposed spectral methods. For works relaxing the one-coin DS assumption, we have only found that Liu et al. [70] imposed a two-coin DS model and has done empirical error rate analysis based on belief propagation, EM and a mean field method with the conclusion that all of them can achieve nearly optimal rates with proper prior settings on worker confusion matrices.

Despite these limitations, the current results of the QCC error rate analysis provide insights into setting up both the early stopping criteria and the response redundancy requirement for crowdsourcing responses provided that certain parameter estimation methods are used for true answer estimation.

2.2.5 Non-Probabilistic Worker-Ability Models

A number of quality control methods that are not based on probabilistic inference have also been developed. In these methods, the abilities of crowd-workers and the quality of their answers are modelled to mutually support one another. The more workers who give the same answer to a particular question, the higher the quality of that answer will be. Likewise, the more responses of high quality provided by the same worker, the higher the ability of this worker. The above mutually supportive relationship is analogous to the authority-hub relationship modelled by the HITS framework [86]. In this case, the inference of the worker abilities and the quality of worker responses has been conducted in similar ways to HITS by some of the current quality control methods [76, 77, 78, 79]. In terms of inferring the true answer for each question, this can be done by aggregating all the responses received by that question weighed by the responses' respective quality estimates. Alternatively, the weights can be the difference between each response and the true answer estimate for the question [76].

2.2.6 Truth-Discovery Worker-Ability Models

Research on *truth discovery* from different (possibly unreliable) information sources [80] shares similar modelling characteristics to the wisdom-of-the-crowd quality control methods. In this case,

each source of information (equivalently a crowd-worker in crowdsourcing) is associated with a reliability variable, called a weight, which measures the quality of the claim (i.e. a response in crowdsourcing) made by the source about an object (i.e. a data item in crowdsourcing). The general goal in truth discovery modelling is to minimize the sum of the weighted distance between each claim and the latent ground-truth of the corresponding object. This distance function can be any loss function depending on the data type of the claims and the ground truths. For example, the claims and the ground-truths can be observed as real-valued feature vectors, which do not often occur in QCC modelling, and their distances can be measured as the squared or absolute difference between these vectors. In comparison, QCC models mainly focus on minimizing the log-loss during the learning process. A comprehensive review on truth discovery models was provided by Li et al. [80]. Thus, the details of these models will not be covered in our review. Nevertheless, the main idea of jointly modelling the quality of workers/information sources and the ground-truths of items/objects is the same in both truth discovery and quality control for crowdsourcing. Some QCC models have started to adopt the idea of distances in truth discovery when handling crowdsourcing tasks over ordinal or continuous responses such as object counting [82], and percentage annotation [81].

2.3 Modelling Worker Expertise and Question Difficulty

More sophisticated quality control methods have taken into account not only the abilities of workers but also the difficulty of questions. Recent advances in the QCC research on partially subjective questions [26] have also hinted at the necessity of distinguishing the above ability-difficulty interaction from the interactions between worker preferences and question subjectivity. Subsections 2.3.1 and 2.4 will cover the state-of-the-art QCC approaches that have made progress in the modelling of either of the above two types of interactions.

2.3.1 The GLAD Model

In a crowdsourcing task, it is common to find that some questions are inherently more difficult for crowd-workers to answer correctly than others. Only experts are able to reliably answer the difficult questions while even novice workers can correctly answer the easy ones. Thus, the average quality of workers' responses to difficult questions will be intrinsically lower than the quality of responses (from the same workers) to easy questions. Based on this idea, some QCC models have taken the question difficulty into account alongside the worker ability for estimating the quality

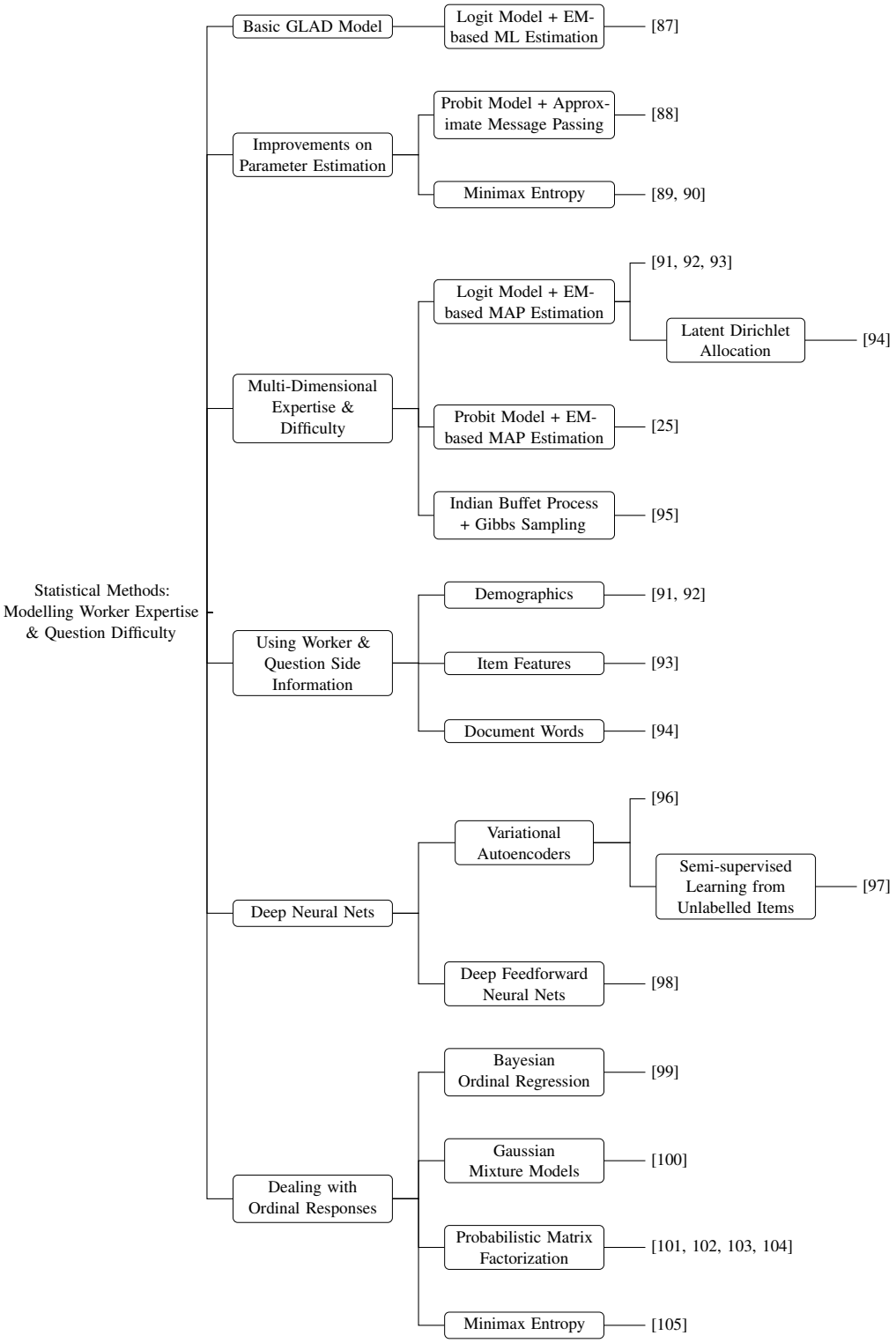


Figure 2.4: A taxonomy of QCC papers that considered both worker expertise and question difficulty to account for the response quality. These papers also focused on statistical modelling and inference.

of each worker response [87, 88, 91, 92]. More specifically, they have modelled the quality of a worker response as a function of each worker's ability and each question's difficulty. The output of the function represents the quality of the worker response as the probability of it being correct. The

most fundamental work in this area is the model of Whitehill et al. [87], referred to as the **GLAD** model, in which the following logistic function δ is used:

$$P(r_{ij} = l_j | e_i, d_j) = \delta(e_i, d_j) = \frac{1}{1 + e^{-(e_i / \exp(d_j))}} \tag{2.4}$$

where e_i is a real-valued parameters that models the ability/expertise of the worker i, and (also real-valued) d_j models the difficulty of the question j. The exponent transformation $\exp(d_j)$ serves to prevent negative difficulty. Compared to the DS model which considers the bias of a crowd-worker towards certain (possibly incorrect) responses[2], GLAD only accounts for the probability of the correct response and ignores any biases by assuming their corresponding probabilities to be uniform as $\frac{\delta(e_i,d_j)}{1-|\mathcal{K}|}$.

In GLAD, the probability of a correct response is considered as the quality of the response and is parametrized by the sigmoid function $\delta(e_i, d_j)$, which means that:

- Increasing the expertise e_i or decreasing the difficulty d_j results in higher quality q_{ij} of the response r_{ij}.
- Decreasing the expertise e_i or increasing the difficulty d_j results in lower quality q_{ij} and it being more likely the response r_{ij} will be a systematic error/bias made by the worker.
- If the value $\delta(e_i, d_j)$ is equal to or approaches $\frac{1}{|\mathcal{K}|}$, then the quality q_{ij} of the response r_{ij} is close to being random. When questions have binary options (i.e. $|\mathcal{K}|$=2), the response r_{ij} being a random error (i.e. q_{ij}=0.5) means either the expertise e_i is zero or the difficulty d_j is $+\infty$.

Like the DS model, the GLAD model also adopts the EM algorithm for parameter estimation. More specifically, in the E-step, for each question j, the GLAD model estimates the probability of the true answer $l_j = k$ as:

$$P(\hat{l}_j = k | \mathcal{R}_j, \{\hat{\boldsymbol{e}}_i\}_{i \in \mathcal{I}_j}, \hat{d}_j) = \hat{\rho}_{jk} = \frac{\left(\prod_{i \in \mathcal{I}_j} (\delta_{ij})^{\mathbb{1}\{r_{ij}=k\}} (\frac{1-\delta_{ij}}{|\mathcal{K}|-1})^{\mathbb{1}\{r_{ij} \neq k\}} \right) P(\hat{l}_j = k)}{\sum_{k' \in \mathcal{K}} \left(\prod_{i \in \mathcal{I}_j} (\delta_{ij})^{\mathbb{1}\{r_{ij}=k'\}} (\frac{1-\delta_{ij}}{|\mathcal{K}|-1})^{\mathbb{1}\{r_{ij} \neq k'\}} \right) P(\hat{l}_j = k')} \tag{2.5}$$

In equation 2.5, δ_{ij} is the shorthand for $\delta(\hat{e}_i, \hat{d}_j)$ = $P(r_{ij} = \hat{l}_j | \hat{e}_i, \hat{d}_j)$, the probability of response r_{ij} being correct, where $\hat{e}_i$ and $\hat{d}_j$ are respectively the estimates of the expertise of worker i and the

[2]For example, in relevance judgement, a worker has a tendency to respond "highly relevant" to any document given any query regardless of whether the true relevance level is "not relevant", "relevant" or "highly relevant".

difficulty of question j. In the same equation, $(1-\delta_{ij})/(|\mathcal{K}|-1)$ is the probability of $r_{ij} \neq \hat{l}_j$ (i.e. response r_{ij} is incorrect).

In the M-step, the expected joint likelihood over the observed responses and unobserved true answers with respect to the true answer probabilities estimated by the E-step is maximized over the rest of the parameters. More specifically, the expected joint likelihood Q is formulated as follows in GLAD:

$$\begin{aligned} Q(\{e_i\}_{i\in\mathcal{I}}, \{d_j\}_{j\in\mathcal{J}} \,;\, \mathcal{R}, \{\hat{l}_j\}_{j\in\mathcal{J}}) &= \boldsymbol{E}\Big[\log\big(p(\{\hat{l}_j\}_{j\in\mathcal{J}}, \mathcal{R} | \{e_i\}_{i\in\mathcal{I}}, \{d_j\}_{j\in\mathcal{J}})\big)\Big] \\ = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \hat{\rho}_{jk} \log\big(P(\hat{l}_j = k)\big) &+ \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}_j}\sum_{k\in\mathcal{K}} \hat{\rho}_{jk} \log\Big((\delta_{ij})^{\mathbb{1}\{r_{ij}=k\}}\big(\frac{1-\delta_{ij}}{|\mathcal{K}|-1}\big)^{\mathbb{1}\{r_{ij}\neq k\}}\Big) \end{aligned} \tag{2.6}$$

where $\mathcal{R}$ is the set of all the responses, and $\boldsymbol{E}$ denotes the expectation. The set of worker expertise parameters $\{e_i\}_{i\in\mathcal{I}}$, and question difficulty parameters $\{d_j\}_{j\in\mathcal{J}}$, are estimated using gradient descent by taking partial derivatives of Q with respect to each element.

2.3.2 Improvements on Parameter Estimation

Many later quality control methods have followed the main idea in GLAD [87]: the quality of a response being a probabilistic function over variables representing the ability of the worker and the difficulty of the question. In Bachrach et al. [88], the probit model is used where the logistic function is replaced by the Gaussian cumulative distribution function with the mean being the worker expertise e_i minus a question-specific bias term, and the variance being the question difficulty d_j. Additionally, instead of using the EM algorithm for the maximum likelihood estimation, this work employs approximate message passing inference on the model parameters.

Zhou et al. [89] used a minimax entropy model to estimate the accuracy of each response. The entropy function is evaluated over all the responses with respect to their probabilities. The ability of worker i and the difficulty of question j are introduced as Lagrangian multipliers for the constraints derived from the i-th row and the j-th column of the response matrix. The authors maximized the constrained entropy function with respect to the response probabilities $P(r_{ij} = k)$, and the ability and difficulty multipliers. Then, the constrained maximization was minimized with respect to the latent true answer of each question, which was shown by the authors to be equivalent to minimizing the KL-divergence between the probability estimates of the true answers and their underlying distribution. Later, Zhou et al. [90] extended their original work by regularizing the

minimax optimization with relaxed constraints to prevent its response probability estimates from overfitting sparse responses.

2.3.3 Multi-dimensional Worker Expertise and Side Information

More recent work has extended the worker-question interaction to be *multi-dimensional* arguing that crowd-workers can have their own areas of expertise and questions can be associated with the different domains of expertise. The authors of GLAD exploited this idea by simply converting the worker expertise and the question difficulty into vectors rather than scalars. Correspondingly, they converted the original scalar product into a dot product between these vectors in Ruvolo et al. [91, 92].

Ruvolo et al. [91] was also the first work to leverage side information of both crowd-workers (i.e. their demographics) and questions (i.e. the characteristics of the data items described by the questions) for further improving the estimation of response quality and the prediction of the true answer. It incorporates the side information into the multi-dimensional GLAD model as the design matrices for the linear regressions that respectively determine the prior means of the expertise and difficulty vectors in the logistic function. Meanwhile, similar work was done by Welinder et al. [25] who used the Gaussian cumulative distribution function to represent the quality of worker responses with the mean being the dot product introduced in Ruvolo et al. [91] minus a worker-specific bias term.

In Wauthier and Jordan [95], rather than set up the dimension for the worker expertise and the question difficulty vectors in advance as the previous work did, the proposed method modelled the dimension and the selection of the underlying latent expertise and difficulty components of the vectors as a Bayesian non-parametric Indian Buffet process [106]. Based on such modelling, this work applied Gibbs sampling to infer the model parameters including the true answers of the questions.

In Kajino et al. [93], convex optimization techniques were proposed for training worker-specific binary classifiers which took side information features about questions (indicating question difficulty) into account when estimating logistic functions. To allow for the multi-dimensionality of the question features, these classifiers are endowed with weight (expertise) vectors in the following logistic function:

$$P(r_{ij} = l_j | \boldsymbol{w}_i, \boldsymbol{x}_j) = \delta(\boldsymbol{w}_i, \boldsymbol{x}_j) = \frac{1}{1 + \exp(\boldsymbol{w}_i^T \boldsymbol{x}_j)} \tag{2.7}$$

In equation 2.7, $\boldsymbol{w}_i$ is the real-valued weight vector specific to worker i and $\boldsymbol{x}_j$ is the real-valued feature vector for question j. The weight vectors follow a Multivariate Gaussian distribution and the maximum a posteriori (MAP) estimate of the mean vector serves as the weight vector of a base classifier to estimate question true answers using equation 2.5.

In Ma et al. [94], the multi-dimensionality of the expertise was estimated in a topic-wise fashion for questions each associated with a text document. The difficulty of each question was modelled to be independent from their topics as a single variable. The proposed model used Latent Dirichlet Allocation (LDA) [107] with a universal distribution of the topics across all the documents (assuming each of them to be short). It draws the topic of a question from that distribution and selects the corresponding topical expertise of each worker (from their topical expertise vectors) to calculate the correctness probabilities of their responses.

2.3.4 Neural Network Approaches

Deep Learning approaches [108] have become very popular over the last few years in Machine Learning applications. Recently, Yin et al. [96] used variational autoencoders [109] to map responses to each question into latent true answer distributions for the questions. The inputs and outputs of the autoencoders are vectors corresponding to individual questions. Each vector is a concatenation of the *one-hot* encoding of the response given by each worker to a particular question. Both the encoder and the decoder are implemented as single-layer networks and the global weight vector for each layer accounts for the response biases across the workers towards different response options. In addition to the weight vectors, a question-specific scalar term is incorporated at each layer of the autoencoders to account for question difficulty. It does this by scaling the input to each layer prior to a softmax transformation.

Atarashi et al. [97] treated the problem as semi-supervised learning and also used variational autoencoders to facilitate latent true answer inference. They leveraged features of unlabelled items to help distinguish the true answers from some (item-specific) latent factors, both of which are assumed to have generated the various feature values of the labelled and the unlabelled items. Instead of using responses and true answers as input-output pairs for the encoder part (as done by Yin et al. [96]), the authors used the labelled and unlabelled item features as the inputs, and both true answers and latent factors as the outputs for the encoder part. The relationships between the responses and the true answers are captured using multi-class logistic regression.

In Gaunt et al. [98], deep feedforward neural networks were trained at two consecutive stages. To train the network at the first stage, accuracy of workers and difficulty of questions were estimated based on degrees of response agreement. The authors then segmented the accuracy and difficulty estimates into different levels (e.g. "low" and "high" levels of accuracy/difficulty) over which they refined the estimates of accuracy and difficulty (to obtain the accuracy of workers on answering easy/difficult questions, and the difficulty of questions for novice/expert workers). The inputs to the network correspond to individual responses. Each input vector contains both the overall estimates of worker accuracy and question difficulty and their estimates across different levels. The outputs of the network are the correctness probability estimates of individual responses. They are used to construct inputs to the neural networks at the second stage. Each of these networks corresponds to a response option for each question. The input to such a network consists of the probabilities of each worker selecting the particular response option. The probabilities equal the outputs from the first stage, i.e. the correctness probabilities $P(r_{ij} = l_j)$, if workers' responses are the same as the response option. Otherwise, the probabilities equal $\frac{1-P(r_{ij}=l_j)}{1-|\mathcal{K}|}$. The outputs of the second-stage networks are finally normalized to obtain the true answer probabilities for each question.

2.3.5 Dealing with Ordinal Response Data

In crowdsourcing, the response options are sometimes not categorical but rather ordinal (e.g. the relevance level of a document to a query) or continuous (e.g. the count of an object in an image). If this is the case, then it is natural to measure the distance between each worker's response and the corresponding correct answer, since they can be calculated on the same real-valued scale. This distance can then be used directly to represent the quality of the response. The function that naturally encodes such a distance is the Gaussian probability density function. In this case, the mean of the Gaussian distribution is set to be the latent true answer, and the precision (i.e. the inverse of the Gaussian variance) typically is set to be the ratio of worker expertise to question difficulty. The function is used to determine the unknown true answer of a question by assuming all of the received responses to be Normally distributed around the true answer. A response with a higher Gaussian density value means that it is closer to the true answer, and thus has higher quality.

In Lakshminarayanan and Teh [99], this framework was followed with one extra treatment that was to use global incremental intercepts of an ordinal regression to model the natural ordering existing in the responses (e.g. no/medium/high relevance of documents to queries). In Metrikov et al. [110], the framework was further extended to be a worker-specific Gaussian mixture with

each Gaussian distribution component being equivalent to an intercept term of a worker-specific ordinal regression. This setting deals with the situation where workers show their individual biases towards different ordinal response options.

In Jung and Lease [101] and their subsequent work [102, 103, 104], a slightly different framework was adopted where each response is set to follow a Gaussian distribution with the mean being the dot product between the latent variable vectors of each worker and the target question. The corresponding variance is always set to be one. Sharing the same idea with Kajino et al. [93], these methods estimate the true answer of each question using the dot product between the question's latent vector and the MAP estimate of the Gaussian mean vector over all the workers' latent vectors. Zhou et al. [105] adapted the minimax entropy principle from their previous work [89] to make it compatible with ordinal responses by modifying the constraints to account for the natural ordering in the response options.

2.4 Modelling Worker Preferences and Question Subjectivity

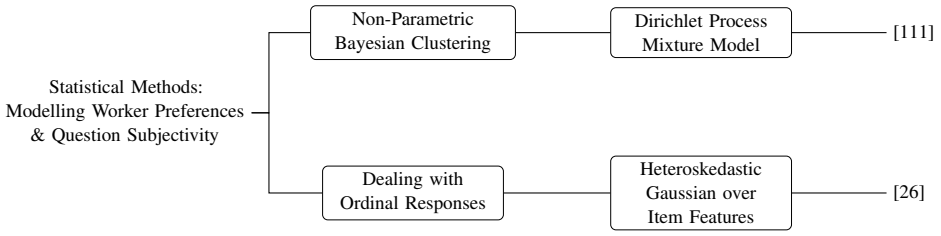


Figure 2.5: A taxonomy of QCC papers that considered worker preferences and question subjectivity. These papers focused on statistical modelling and inference.

As mentioned in Section 1.2.3, some crowdsourcing tasks might contain questions that are at least partially subjective. These questions possess more than one correct answer and in the case of partially subjective questions, also at least one incorrect answer. To answer a partially subjective question correctly, a crowd-worker needs to avoid any incorrect answers, which depends on the interaction between her ability/expertise and the difficulty of the question. In other words, the quality of her response depends on both her level of expertise and the difficulty of the question. At the same time, her subjective opinions on different (subjective) features of the question will cause her to prefer one of the correct answer options over the others.

In recommender systems, users rate purely subjective items according to their preferences for various subjective features of the items. This type of interactions is typically modelled by the dot product between the latent factor vectors of the users and the items. This model is known as *matrix*

factorization in *collaborative filtering* based recommendation [112]. This subject is beyond the scope of our survey and we refer the interested reader to a dedicated survey by Koren and Bell [113] for more details.

To the best of our knowledge, very few current quality control methods have endeavoured to distinguish between question difficulty and question subjectivity in one unified model. One work that has made some progress in this regard is Tian and Zhu [111]. In it, the authors assumed that a higher joint degree of difficulty and subjectivity for an entire crowdsourcing task can increase the number of underlying groupings of responses given to each question in the task, with the expected number of responses in each grouping becoming smaller. The authors proposed to infer the response groupings using a Dirichlet Process Mixture Model [114]. Despite attributing the variance of worker responses to both difficulty and subjectivity, the paper makes no attempt to separately model the two even though they might induce different types of interactions with crowd-workers.

Another work done by Nguyen et al. [26] has achieved similar progress based on a rating dataset containing partially subjective items with observed item features. The rating r_{ij} to item j is assumed to follow a Gaussian distribution with the mean and the variance linearly regressed on the observed features $\boldsymbol{x}_j$ of that item as follows:

$$r_{ij} \sim \mathcal{N}(\boldsymbol{w}^T \boldsymbol{x}_j, \exp(\boldsymbol{v}^T \boldsymbol{x}_j)) \tag{2.8}$$

where $\boldsymbol{w}$ and $\boldsymbol{v}$ are global coefficients.

The mean and the variance can be regarded as accounting for the correct answer and the mixing effects of the subjectivity and difficulty of the item. Therefore, this method is also not intended to separate the modelling of the two properties. Moreover, it is only applicable to the scenario where the answers are ordinal and the item features are observed. A second weakness of this method is that it fails to handle any possible multi-modality in (the distribution of) the responses to each question. In the extreme case, if the response distribution is skewed to both ends, meaning that workers mostly give ratings of either 1 or 5 to a question, the method will perform poorly as it tries to fit a unimodal Gaussian distribution to these ratings. Essentially, one can assume that each mode of the response distribution indicates a (subjective) correct answer to the question. If this model can be extended with worker-specific coefficients (i.e. $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$) to form worker-specific Gaussians, then not only will the model's fit becomes multi-modal, but also the question's subjectivity can

be captured by the modes of the worker Gaussians separately from its difficulty (modeled by their variances).

2.5 Modelling Worker Motivation and Worker Contexts

In crowdsourcing, the motivations of individual workers have significant impact on the quality of

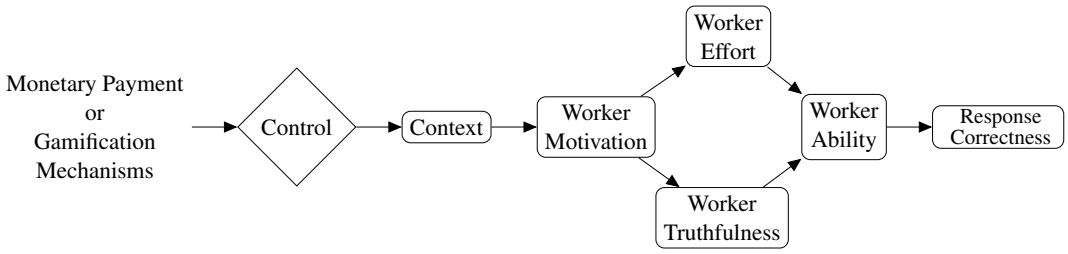


Figure 2.6: A diagram shows payment and gamification mechanisms control worker contexts to affect worker motivation, which further influence worker effort and truthfulness, and eventually the correctness of responses.

their responses. For example, when participating in the same task, workers who are bored by the task tend to make more careless errors than those who are interested in it, even though they possess the same level of expertise to fulfil the task. Furthermore, worker motivation has been observed to vary with the context in which the workers find themselves [115, 116]. Thus, properly manipulating the motivation of workers by controlling their associated contexts may improve the quality of their answers. Past research [117, 118, 119, 120] has proposed to control the context through *incentive mechanisms* which are triggered during crowdsourcing tasks to intervene to positively affect the motivation of the workers. Moreover, based on our anatomy of crowdsourcing contexts in section 1.2.3, the incentive mechanisms can intervene in and control the context at various levels, from the task-level and the session-level to the response-level in order to achieve different impacts on the workers' motivations.

Different incentive mechanisms are designed to favour certain characteristics of worker motivation. According to Ryan and Deci [121], motivations can be broadly characterized as being either *extrinsic*, which is the desire to gain monetary payoffs and avoid costs, or *intrinsic*, which is the desire to achieve fulfilment and enjoyment. Correspondingly, we categorize the past literature on incentive mechanisms into *monetary payment* and *gamification* approaches which are respectively responsible of motivating the workers extrinsically and intrinsically.

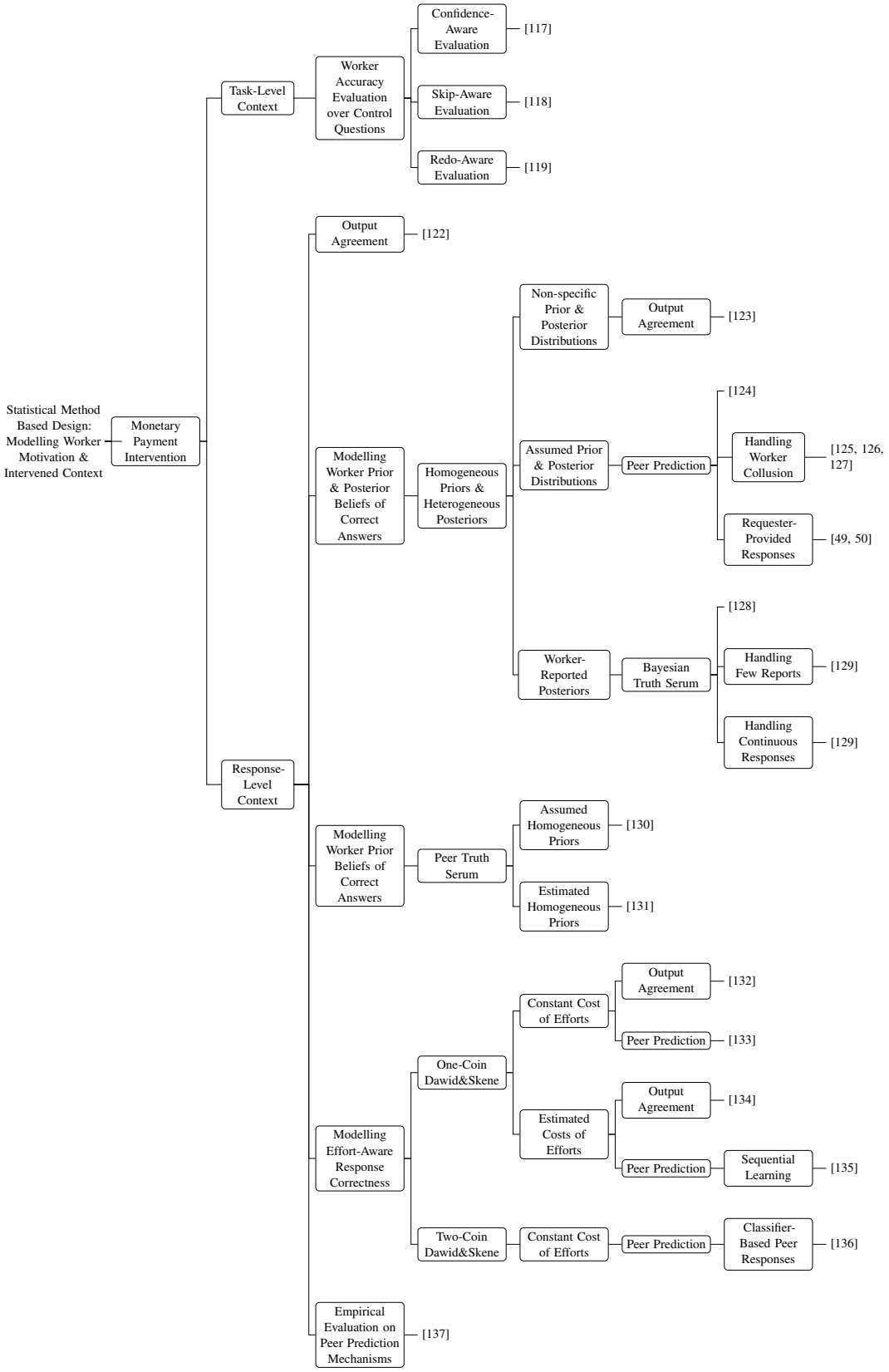


Figure 2.7: A taxonomy of QCC papers that considered the interaction between worker context and (extrinsic) motivation. These papers focused on designing monetary payment mechanisms which rely on statistical modelling (and possibly estimation) of worker attributes.

2.5.1 Incentive Mechanisms Based on Monetary Payment

Monetary payments in crowdsourcing involve two types: *base* payments and *bonus* payments. The extrinsic motivation of rational workers is to choose answer options that maximize their monetary payoffs [138, 139, 140]. The payoffs are usually formulated as the difference between the monetary rewards given to the workers for the responses they provide and the costs incurred (the effort exerted) to generate these responses [132, 140]. Maximizing the payoffs means minimizing the (costs of the) effort exerted. Consequently, one expects crowd-workers to minimize their effort, which generally leads to a deterioration in the quality of their responses. Moreover, some workers could even deliberately choose to not *truthfully* report the answers that they elicited with effort (e.g. by flipping their answers to true/false questions) if they believe that doing so could result in higher payoffs [131].

Thus, the extrinsic motivations of workers in default contexts of paid crowdsourcing tend to cause the workers to invest as little effort as possible to respond to the questions and even respond untruthfully if that benefits them. To address this issue, specialized incentive mechanisms have been devised that alter either the fixed base payment or the fixed bonus payment, causing the payment to become *adaptive* at the task level (or some lower levels of contexts). The aim of the adaptiveness is to ensure that the expected payoffs of the crowd-workers are maximized only when they exert sufficient effort to produce high-quality responses and report these responses truthfully.

2.5.2 Monetary Payment in Task-Level Contexts

In a task-level context, control questions are sometimes available and can be inserted randomly into each question page of the task (such that crowd-workers do not know their whereabouts). Once the workers finish the task, the monetary payments to them can be adapted to be different (i.e. either the base or bonus payments) based on their accuracy on the control questions. To be more specific, the accuracy of the responses to those control questions serve as the inputs to some carefully designed *payment function* which outputs a final reward for the worker and guarantees that this reward be maximized only when she has exerted sufficient effort and reported all her answers truthfully.

In Shah et al. [117], the crowdsourcing task allows workers to express their confidence in different response options as the correct answer for each question. The confidence scores of a worker for all the control questions are then taken into account when estimating the worker's accuracy that is then used to calculate the payment to the worker. In Shah and Zhou [118], the task additionally provides a "skip" option for each question and the final payment is decreased

for each skipped control question and increased (multiplicatively) for each correctly answered one. In later work, Shah and Zhou [119] used a two-stage design was proposed where a worker first answers all questions, and then is provided with the opportunity to change her answers after viewing a reference response from another worker to each of the same questions. The payment function evaluated over the worker's responses to the control questions ensures truthful reporting from the worker in both stages.

2.5.3 Monetary Payment in Response-Level Contexts

Most state-of-the-art incentive mechanisms that are based on monetary payments were developed under the assumption that the control questions are unavailable in crowdsourcing tasks or too scarce to be used reliably. Certain mechanisms were also developed for making the payments adaptive to the response-level context. This means that a worker is not paid the same amount for every question and that different workers may be paid differently for their answers to the same question. The aim is to make the payments dependent on the quality of the response, where a response of higher quality deserves a higher base or bonus payment.

Since the quality of a response is unknown without control questions, the incentive mechanisms in this case resort to a strategy called *peer consistency* to assess the quality of a worker's response. In this case, a response from another random worker, called a *peer worker*, is selected for the comparison with the target response of the current worker. If these two responses are the same, then the current worker will be rewarded according to a payment function that is carefully designed to induce a *game-theoretic equilibrium* among all the participating workers [123, 130, 131, 132]. In such an equilibrium, no worker can improve their expected payoff by acting differently from what is required by the mechanism, namely that they truthfully report their answers to the requester. Such an equilibrium is thus also referred to as a *truthful* one. Moreover, the game-theoretic incentive mechanisms usually model the belief systems of crowd-workers, which consist of their *prior* and *posterior* beliefs about the correct answers to questions. The belief systems are assumed by these mechanisms to be either *homogeneous*, which means they are identical across all workers, or *heterogeneous*, which means that different workers possess different prior and posterior beliefs as well as different ways of updating the beliefs about the correct answers.

Output agreement [122] is the most basic peer consistency mechanism which does not assume any form of belief system (i.e. neither specific distributions over correct answers nor whether the distributions are shared) among workers. It only involves paying a worker for her response

to a question when the response is the same as the one given by a randomly selected peer to the same question. Based on output agreement, Waggoner and Chen [123] assumed a homogeneous prior belief across workers (i.e. a shared non-specific prior distribution over correct answers) and heterogeneous posterior beliefs (i.e. private non-specific distributions) according to workers' individual understanding after reading the question. They defined a broader payment function by replacing the 0/1 error function in output agreement with the Euclidean distance between the response and the peer's response. They showed that output agreement based on this general payment scheme at best results in a strict equilibrium where workers report the correct answer according to the common part of their understanding.

Peer prediction [124] is another early work based on peer consistency assessment which assumes homogeneity for prior beliefs and heterogeneity for posterior beliefs with specific distributions over correct answers. It proposes to use the assumed posterior updated from observing a worker's response to a question together with a random peer's response to the same question to calculate the reward for the worker's response. In the original paper, the authors consider the case in which true answers and responses are continuous and for which they assume the belief systems follow Normal distributions. In general, conjugate distributions are usually selected for the belief systems to facilitate belief updates. A drawback of the peer prediction approach is the existence of multiple undesired equilibria caused by the collusion among crowd-workers that allows them to exert no effort (e.g. by copying each others' answers to every question) and yet gives them higher expected payoffs than the truthful equilibrium does. Thus, subsequent work has focused on removing such equilibria [125, 126] or designing payment functions that penalize the "collusion equilibria" to make them have smaller payoffs than the truthful one [127]. However, the techniques still that true answers and responses are binary.

Peer truth serum (PTS) [130] is an alternative when requesters cannot find appropriate distributional assumptions for the posterior beliefs of workers. This is because PTS does not consider the posterior beliefs in the payment design. Instead, it assumes homogeneity for the prior belief of workers about correct answers for which the requesters need to provide specific distributions. PTS makes use of the assumed prior distributions and the 0/1 distance between the target worker's response and a random peer's response to the same question to calculate the target's reward. Such a payment function was shown by the authors to induce at least one "non-truthful" equilibrium where all workers collude with one another to always give the least likely responses to each question. Instead of using a predefined prior distribution, subsequent work [131] focused on dynamically

estimating the prior using frequencies of responses from other workers to the same question to which the response of the current target worker was given.

Bayesian truth serum (BTS) [128] assumes homogeneous prior beliefs (i.e. a shared non-specific prior distribution) for crowd-workers. On the other hand, it implies that the posterior beliefs are heterogeneous by requiring additional assessments from workers about the probabilities over correct answers to each question along with their responses. BTS obtains the geometric mean of these probability estimates excluding the one from the target worker and combines it with the frequencies of collected responses to calculate the payment for the target. A weakness of BTS is that it needs a large number of workers to answer the same question in order to produce a reliable geometric mean to achieve a truthful equilibrium. Robust BTS [129] was proposed which modified the payment function of BTS to handle the situation where only a few workers answer each question. Furthermore, a divergence-based BTS [141] has been proposed to handle continuous responses (e.g. numbers). The payment function of this mechanism leverages the KL-divergence of the probability estimates over intervals that might contain the correct answer between the target worker and a random peer.

The work reviewed thus far focuses on modelling crowd-workers' beliefs or distributions on the correct answers of questions. This is different from non-incentive quality control models such as DS and GLAD which focuses on modelling the response correctness or biases given a global distribution over the correct answers. The former type of modelling emphasizes the elicitation of honest responses (i.e. workers exert efforts and truthfully report what they think to be the correct responses) which might turn out to be incorrect. The aggregation of these responses to obtain better final answers can come afterwards using the DS or GLAD models.

There have been other incentive mechanisms which directly model response biases of workers (as the DS model does) for deriving payment functions that are able to achieve a truth-telling equilibrium among the workers. Unlike the DS model, they do not explicitly infer question true answers but rather aim at eliciting effort-exerted and honest responses. The first work in this regard was proposed by Dasgupta and Ghosh [132]. They dealt with binary response options based on the one-coin DS model. More specifically, they model the efforts exerted by workers as binary variables that control the switch between arbitrary guessing (i.e. zero-effort) and the workers' response correctness probabilities (i.e. effort-exerted) which were assumed to be always greater than 0.5. The payment function in this case was designed to both recognize the response agreement between the target worker and a random peer for the same question, and penalize zero-effort

(coincidence) agreement given both workers' response statistics calculated from the other questions. The authors proved that this payment function avoided a zero-effort equilibrium by making it always less appealing than a truthful equilibrium in terms of expected rewards over efforts and response correctness.

Based on the same one-coin model, Witkowski et al. [133] additionally considers the scenario where a worker would make the decision on whether to participate in the crowdsourcing task. The probability of participation equals the worker's response correctness probability, which models the worker's self-assessment about their qualification. Correspondingly, the payment function is designed in such a way that unmotivated or unqualified workers will prefer to not participate rather than guess an answer (with zero effort), which also means that those who participate will be qualified and invest efforts in the equilibrium.

Both Dasgupta and Ghosh [132] and Witkowski et al. [133] have assumed the cost induced by non-zero efforts is a constant. Within the same modelling framework, Liu and Chen [134] proposed an extension which considers varying unknown costs randomly drawn from a distribution under non-zero efforts. Learning the cost distribution requires the workers to additionally report their costs of answering each question. The learning process is integrated with an incentivizing process, which aims to reach a truth-telling equilibrium, under a multi-armed bandit framework. The framework optimizes the trade-off between the two processes. Another hybrid mechanism that combines the DG mechanism with a multi-armed bandit framework to realize a similar goal was proposed by Liu and Chen [135]. It learns the optimal choices of bonus levels at each time step for workers categorized into two peer groups that cross validate the truthfulness of each other's answer reporting behaviour.

Based on a two-coin DS model that captures workers' biases in binary labelling, Liu and Chen [136] leveraged binary labels generated by classification algorithms as the benchmark labels against which worker responses were compared for peer consistency assessment. The assessment results are then input to a payment function which guarantees that if the error rates of the classification algorithms on predicting the true labels converge towards zero, the function is able to achieve a truthful equilibrium.

The above mechanisms have more-or-less coped with undesirable equilibria, which yield higher expected payoffs than the truthful ones do, in their theoretical formulation of the payment functions. However, empirical evidence remains insufficient in the following two aspects [130, 137]. First, it is still unclear that whether the existence of these undesirable equilibria actually pose a problem to the

quality of crowdsourced data in practice. Second, it remains to be seen whether these theoretically elegant truth elicitation mechanisms based on peer consistency assessment can work effectively in practice.

After the intervention of any of the above incentive mechanisms into the monetary payments for workers' responses, quality control methods can be further applied to the crowdsourced responses to produce more reliable estimates of the question true answers. There are also unified frameworks proposed by Frongillo et al. [49] and Ho et al. [50] that integrate the above process. More specifically, Frongillo et al. [49] combined a payment function that ensures a truth-telling equilibrium with the Bayesian statistics to allow for both truthful answer elicitation and Bayesian aggregation for inferring the correct answers to questions. In Ho et al. [50], a further step was taken to optimize the multiple-choice interface with confidence shown to each worker (the interface being similar to the one employed by Shah et al. [117]) along with the optimization of the Bayesian aggregation.

2.5.4 Incentive Mechanisms Based on Gamification

Gamification refers to incorporating (video) game elements into various (levels of) contexts of crowdsourcing tasks in an attempt to lift the intrinsic motivations of crowd-workers to make them more engaged in answering questions, which eventually leads to improved quality of their answers. The intrinsic motivations in this case are crowd-workers' feelings of enjoyment, playfulness, and accomplishment (e.g. through improvement of their own skills), and the welfare of the communities. Different from the payment-based incentive mechanisms discussed above which derive theoretical guarantees of expected performance of crowd-workers, the gamified incentive mechanisms have proven themselves empirically to have improved the quality of workers' answers.

2.5.5 Gamification in Response-Level Contexts

Score feedback (aka "scoring") is the most fundamental means of gamification that allocates a certain number of points to a worker once she has answered a question depending on the quality of her response. Higher-quality responses should be rewarded with more points resulting in higher scores. This makes the scoring very much similar to dynamic monetary payments in terms of how they are allocated. They are however different with regard to the types of motivation they deal with. The former provides virtual rewards for increasing intrinsic motivations as opposed to material rewards provided by the latter for increasing extrinsic motivations. In addition, scoring often acts

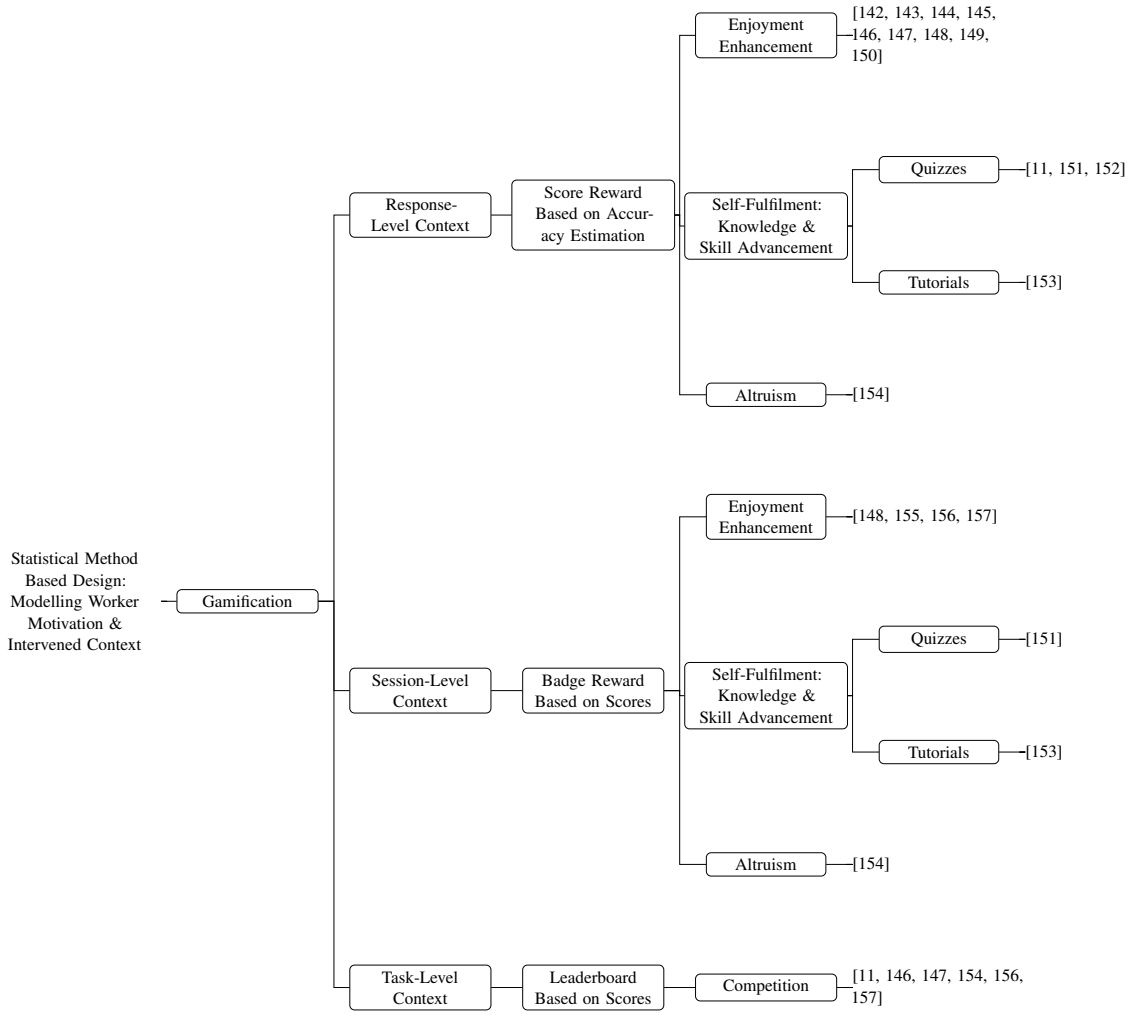


Figure 2.8: A taxonomy of QCC papers that considered the interaction between worker context and (intrinsic) motivation. These papers focused on designing gamification mechanisms which, in most cases, rely on statistical estimation of worker accuracy/expertise.

as the building block of other more complicated game elements such as leaderboards and level systems.

In Brenner et al. [142], a simple and fun online game was developed to help train a face recognition system to refine its classification. In the game, a crowd-worker was rewarded with certain points every time she provided correct feedback regarding an uncertain recognition result from the system. The experiments showed that the game, without paying any bonus, iteratively improved the recognition performance of the system solely based on the enjoyment/fulfilment it brought the workers. The incentivizing aspects of the above game design, including the graphics and the scoring mechanism, made the game enjoyable for workers to engage in, causing them to provide more accurate answers in general. Thus, these design aspects have appeared repeatedly in gamified crowdsourcing over different areas including medical facts elicitation [143], relevance

judgement in Information Retrieval [144, 145, 146], image classification [146], video captioning [147], language translation [148], and communication to the general public about culture [149] and science [150].

Apart from the enjoyment and increased productivity it has brought to crowdsourcing, the scoring mechanism can also help to build more sophisticated incentive mechanisms that advance the skills and knowledge of crowd-workers. Such advancement allows the workers to produce responses of higher quality. A typical skill-development mechanism that has been boosted by the scoring mechanism from gamified crowdsourcing is the online quiz. In Ipeirotis and Gabrilovich [11], online quizzes were advertised as skill/knowledge tests for individuals in order to attract both unpaid volunteers and crowd-workers. These quizzes contain not only control questions but also questions with unknown answers for which the requesters were seeking correct answers from the participants. The scoring mechanism in this case supported both the performance feedback mechanisms, which display each individual's score and the others' average score, and the all-time leader-boards, which rank the participants by their scores. The experiment results show that such quizzes can attract a large number of participants with diverse skills over a relatively short period, and the total payment is much lower than what would have been required on online human intelligence marketplaces such as Mechanical Turk.

The main idea of the above work to use gamified quizzes to attract participation of (and contributions from) workers or volunteers who seek enjoyable learning experiences, has also been adopted in Stanculescu et al. [151] and Boyd-Graber et al. [152]. The former work leveraged the idea for engaging employees in learning about enterprise history, products and services while crowdsourcing some subjective data from them (e.g. their opinions). The latter leveraged the intrinsic fun of quiz bowl [158] for engaging online players to provide answers to questions as the labels to be used in training classifiers that can perform better question-answering.

Apart from quiz testing, tutorial training/learning is another means of stimulating workers' intrinsic needs for knowledge and skills, and has been seamlessly combined with the scoring mechanism. In Dontcheva et al. [153], a gamified crowdsourcing platform for image editing was developed which attracted large numbers of workers as they could learn skills for producing high-quality and creative editing of images. The basic game element employed by the platform was scoring, which again also supported an all-time leaderboard. Worker satisfaction surveys were collected and showed that most of the workers appreciated the sense of achievement created by the scores when learning the image editing skills. Moreover, feedback from the requesters showed that

the number of images with better quality was almost double with respect to those produced by the originally novice workers.

Scoring mechanism can also help incentivize workers to make more altruistic contributions to their communities. In Lee et al. [154], the focus is on motivating workers to contribute to the construction of a new community. The scoring mechanism in this case quantifies the amount of contribution a worker has made, and supports more advanced game elements including a badge system and an all-time leaderboard, each of which has been able to motivate the workers to perform better in terms of the amount contributed and the quality of the contributions.

2.5.6 Gamification in Session-Level Contexts

Badges are typical game elements that are awarded to people for recognizing their achievements and contributions at different levels (usually with bronze, silver and gold badges corresponding to the increasing levels). Many crowdsourcing marketplaces (e.g. CrowdFlower) have implemented their own badge systems that award workers within task-level contexts according to the numbers of crowdsourcing tasks they have successfully completed. In a gamified crowdsourcing task, badge awarding usually happens in session-level contexts. More specifically, when a crowd-worker completes a session/page of questions, she gains some points and whenever her total points exceed a certain threshold, a badge system is triggered to award her with the corresponding badge. Such a badge system has been integrated into the session-level contexts of various paid crowdsourcing tasks for various motivational purposes. The purposes range from making laborious and tedious work (such as image annotation [155], proofreading [156], language translation [148] and mobile application testing [157]) more enjoyable, confirming one's learning progress, (e.g. on image editing [153] and enterprise knowledge [151]), to encouraging workers' commitment to building online communities [154].

2.5.7 Gamification in Task-Level Contexts

The most notable game element that has been utilized for gamifying task-level contexts in paid crowdsourcing is the *leaderboard*. Typically, a leaderboard exists throughout the entire duration of the task and is accessible by all the crowd-workers at any time during their participation in the task. The aim is to ignite *competition* amongst the workers, which motivates them to work harder to either overtake those above them in the ranking or to maintain their current rank positions. However, the past research on using all-time leaderboards to incentivize workers has yielded conflicting empirical

results. In the study of Eickhoff et al. [146], steady improvements were observed in the quality of workers' relevance judgements for documents to search queries. Quality was measured in terms of the level of agreement between workers on the same document-query pairs. In Itoko et al. [156] and Saito et al. [147] where workers were required to do proofreading, senior workers were *demotivated* by the competition brought about by the presence of a leaderboard while younger workers found it the other way round. In Lee et al. [154], workers constantly returned to the communities as they would like to follow their status on the leaderboard, and were encouraged by doing so to make more contributions, although their quality varied significantly. In Ipeirotis and Gabrilovich [11], an all-time leaderboard was set up which provided two types of ranking: the percentage of correct answers and the total number of correct answers submitted. The leaderboard in both cases showed positive effects on the quality of workers' answers only in the early stages of the crowdsourcing tasks as it *discouraged* the workers who came late to the tasks when other workers had already amassed a large number of points and had well-established positions on the leaderboard. A similar phenomenon has also been observed by Massung et al. [157] where many new workers collected pro-environmental behaviour data for a mobile application. Performance was initially high before dropping for later arriving workers due to the large difference in the contribution points between the leading workers and themselves. Thus, Ipeirotis and Gabrilovich [11] suggested the leaderboard be embedded in session-level contexts which means that there is a leaderboard dedicated to each page of a task. In this case, workers need to answer correctly much fewer questions to reach the top of a page leaderboard than they need to do with respect to a global leaderboard. As a result, workers who arrive late at a task page are less likely to be intimidated by the (page) leaderboard rankings of those who completed the page earlier. In Lee et al. [154], the experiment results suggested that the leaderboard should only be "switched on" after a certain "warming-up" period for each worker, by which time she will have completed enough questions to make herself feel less disadvantaged by the late starting point.

2.6 Modelling Worker Expertise and Contexts

Not only can the motivation of workers be affected by intervened contexts but also their expertise can interact with the contexts in different ways. According to the literature of quality control for crowdsourcing, we have found the following three mechanisms in which the interaction takes place:

- Worker expertise is improved by *training mechanisms* deployed at different levels of contexts.

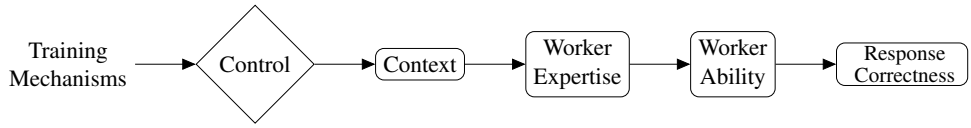


Figure 2.9: A diagram shows training mechanisms control worker contexts to improve worker expertise, which further improves the correctness their responses.

- Worker expertise is leveraged by *question assignment mechanisms* to control the allocation of questions into different levels of worker contexts.
- Expertise of a worker varies across different contexts in which this worker has been situated. For example, the worker's expertise is dynamic across different tasks she has participated, or across different pages of the same task. It is often assumed that workers exhibit similar (levels of) expertise towards similar tasks or sessions.

The first two methods of interaction concern mechanism designs based on modelling worker expertise. The last method of interaction is frequently encoded by statistical models which, according to the literature, also consider the difficulty of questions.

2.6.1 Improving Worker Expertise Using Training Mechanisms at Different Levels of Contexts

According to the literature on quality control for crowdsourcing, the current approaches that intervene in the context to directly affect the *expertise* of workers are *training mechanisms*. In the literature, this effect is assumed to be *orthogonal* to the effect on workers' intrinsic motivation of self-fulfilment brought by gamification. In other words, all the work to be reviewed in this section aims at improving workers' expertise regardless of their motivation. Thus, they are different from the aforementioned gamification work done by Dontcheva et al. [153] which focuses on inspiring people's need of self-fulfilment and the pursuit of new knowledge and skills.

By default, the training of crowd-workers is performed prior to their participation in a task and aims to teach the workers basic skills and expertise required to answer the questions in the task. It has been shown in Gadiraju et al. [159] that the default training can significantly improve the quality of worker responses on a variety of crowdsourcing tasks. The mechanism is restricted however to a task-level context in which each worker only receives the training once throughout the entire task. As a result, even when the later performance of the worker is undesirable, or they do not demonstrate the expected levels of expertise, they are not given a second chance to learn the requested skills (be retrained) to perform better. To solve this issue, Bragg et al. [30] proposed a training mechanism that

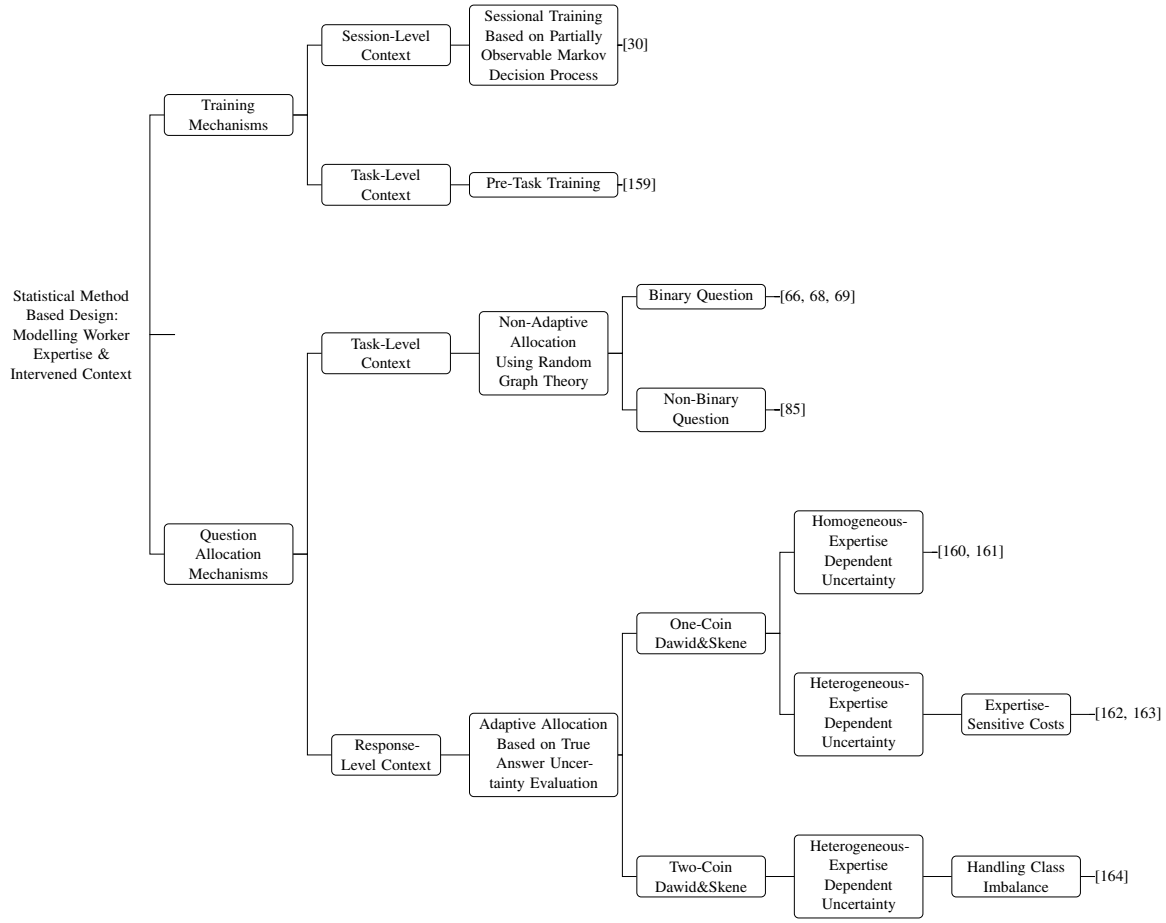


Figure 2.10: A taxonomy of QCC papers that considered the interaction between worker expertise and context. These papers focus on designing either training mechanisms that alter contexts to improve worker expertise or question allocation mechanisms that use worker expertise to determine questions to be answered in the contexts.

functions on the session-level contexts of the crowdsourcing. The mechanism models the decision-making process of whether to train a worker or collect answers from her in each of her working sessions as a partially observable Markov decision process (POMDP) [165]. This mechanism then employs reinforcement learning to estimate the parameters of the POMDP including the expertise vectors of individual workers and the correct answers to the questions.

2.6.2 Question Allocation in Crowdsourcing

Reducing the number of responses collected in crowdsourcing to lower costs while maintaining the prediction accuracy on question true answers given the collected responses has been an important QCC research topic over the years. A common crowdsourcing process involves assigning each participant worker a set of questions which is selected uniformly at random according to a value set up prior to the task by the data requester for the number of questions answered per worker. Each

worker answers the set of assigned questions only once. Unfortunately, such a process often leads to a higher total (monetary) cost than necessary. This is because the uniformly random assignment of the questions is independent of all the informative characteristics of the workers (e.g. their domain expertise, interests, etc.) and the questions (e.g. their domain difficulty, genres, etc.), and thus fails to leverage these characteristics for more efficient performance. On the other hand, if the question assignment can be designed to be biased towards these characteristics, then the question's true answer prediction can potentially be improved with lower costs.

2.6.3 Non-Adaptive Question Allocation Based on Worker Expertise in Task-Level Contexts

In this case, the allocation of questions happens before any worker enters the task. The total number of allocated questions equals the batch size multiplied by the number of workers (if each worker is assigned questions only once and never reused once they finish their batches). Once the task begins, workers arrive in sequence to pick up the corresponding allocated batches of questions. Such pre-task simultaneous allocation of questions relies on designing a bipartite graph which contains two types of nodes: questions and workers, where edges between them correspond to the assignment of a question to a particular worker. In the work of Karger et al. [66, 68, 69], the authors proposed to draw a regular random bipartite graph based on the *configuration model* from the random graph theory [166]. In the graph, the degrees of the question and the worker nodes represent how many workers to assign to each question and how many questions to assign to each worker respectively. The goal of their work is to realize a particular error rate on true answer prediction with minimum costs (i.e. minimum degree for the question nodes or equivalently, minimum number of responses[3]). The authors proved that using a regular random graph to achieve a target error rate was sufficient. This graph's actual error rate was within a constant factor of the target rate using the underlying graph (which is possibly neither regular nor random) with the best possible inference algorithm. The authors also showed that the cost incurred by each binary question (i.e. number of responses per question) to achieve a target error is scaled by the inverse of the expectation of the individual workers' expertise. In their following work [85], the same authors investigated the same subjects but with non-binary questions. They derived similar results in terms of the near-optimality

[3]The minimum number of responses equals the minimum degree for the question nodes multiplied by the number of questions.

of the regular random graph in achieving any target error rate and the scaling effect of expertise on the cost per question.

2.6.4 Adaptive Question Allocation Based on Worker Expertise in Response-Level Contexts

For the adaptive schemes, the question allocation happens within an ongoing task and is dependent on the current estimate of each worker's expertise based on their responses so far. In their pioneering work, Sheng et al. [160] and Ipeirotis et al. [161] proposed to model one key aspect in the adaptive allocation, that is the *uncertainty* of each question's true answer. In that work based on the one-coin DS model for binary questions, the uncertainty of a question depends on the expertise/ability of the workers who answered it. The higher the expertise, the lower the uncertainty will become. The authors simplified the scenario by assuming that all the workers shared the same level of expertise and were non-adversarial (i.e. their response correctness probability always greater than 0.5). They proposed a design in which at each timepoint, the question with the largest amount of uncertainty in its correct answer will be assigned to an arbitrary worker. As a result, the same question might be assigned to multiple workers. In this case, the expertise of individual workers governs the quality of their responses from which an integrated response can be derived for the question. The lower the expertise of each worker, the more responses are needed to generate an integrated response for the question that has a low true answer uncertainty. In their following work [162, 163], Wang and Ipeirotis addressed heterogeneous worker expertise. In this case, the entropy of the probability estimates from equation 2.1 embodies the uncertainty about the true answer of each question. The question with the highest entropy was selected for assignment at each time. The payment for each worker is proportional to their current expertise estimates. The lower the expertise, the less payment a worker will receive for a response.

In Zhang et al. [164], the question assignment strategies were further extended to be based on two-coin models which encode binary biases of individual crowd-workers corresponding to the positive and negative question types (as determined by the binary response options). They also devised an adaptive decision boundary for determining the true answer of each question and further, degrees of the uncertainty when class imbalance exists in the true answers. In Donmez et al. [167], a hidden Markov model was proposed to capture the correlation in the time-varying ability of each crowd-worker. At every time step, the workers were ranked based on estimates of their current

abilities and only the top worker for each question was assigned the question to answer. Compared to Wang et al. [162, 163], this work fails to utilize all the crowdsourcing power available.

2.7 Modelling Worker Motivation, Question Difficulty and Contexts

In section 2.5, the modelled correlations between response quality and worker/context ignore the *heterogeneity* in the questions answered. Different questions possess different levels of inherent difficulty however. This heterogeneity was ignored in order to simplify the problems under investigation including the derivation of theoretical equilibrium guarantees, and the practical design of the gamification techniques. When heterogeneity in question difficulty is considered, the above problems become more complicated since the response quality becomes harder to measure and estimate.

According to the literature, both the payment-based and the gamification-based incentive mechanisms have successfully controlled the response quality by intervening with the context (to properly motivate the workers) while considering the difficulty of the questions answered. In this case, the payment-based mechanisms are typically applied to the response-level context while the gamification-based mechanisms are applied to the session-level context.

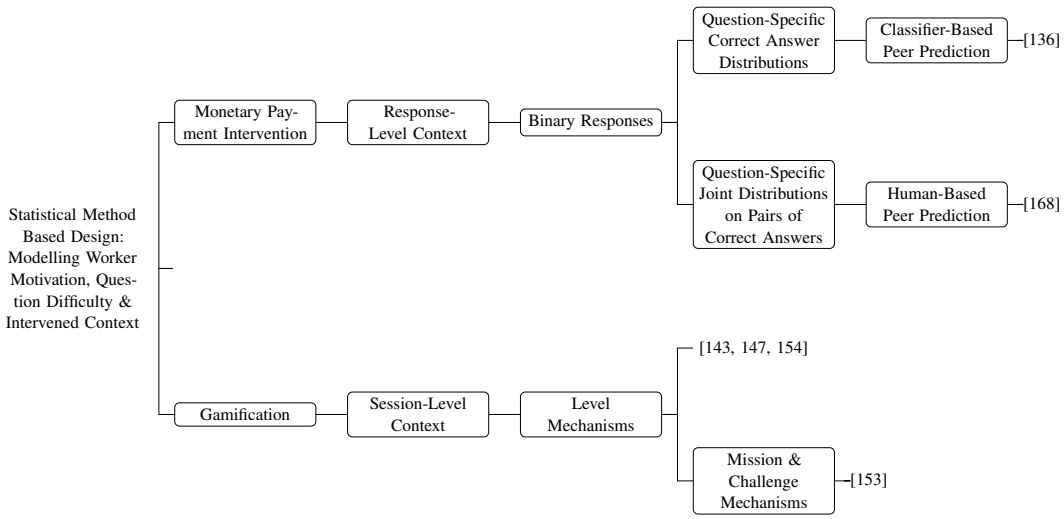


Figure 2.11: A taxonomy of QCC papers that considered worker motivation, context and question difficulty. These papers focused on designing either monetary payment mechanisms which additionally modelled question difficulty or gamification mechanisms which increase question difficulty to challenge workers.

2.7.1 Monetary Payment in Response-Level Contexts

Only recently have payment-based incentive mechanisms started to take into account the heterogeneity in question difficulty in order to refine their design of the response-level payment functions. They do this by increasing the payment for responses whose quality is low due to the fact that the difficulty of the corresponding questions is high rather than a lack of effort on the part of the corresponding workers.

In the peer prediction mechanism proposed by Liu and Chen [136], the difficulty of a binary-response question is encoded by a dedicated probability distribution over its correct answers. If this distribution is (nearly) uniform a priori or a posteriori, it means the question is considered to be so difficult under the corresponding situations that its correct answer remains uncertain. In this case, the payment function was designed to achieve a truth-telling equilibrium for each question with respect to their correct answer distributions.

In Mandal et al. [168], the proposed peer prediction mechanism is also applied to binary-response questions except that it relies solely on *human* peer consistency assessment rather than the hybrid assessment with machine learning done by Liu and Chen [136]. The difficulty of a question is captured by a symmetric matrix of joint probabilities of each pair of response options (including each option with itself). Each entry represents the chance that any random pair of workers agree (the diagonal entries) or disagree (the off-diagonal entries) with one another on the correct answer for the question. The larger the summation of the diagonal entries is, the easier the question and vice-versa for the off-diagonal entries. (The entries denote joint probabilities on each pair of response options and thus their summation equals 1.) The payment function takes in a matrix of posterior joint probabilities given responses thus far to a question and rewards a random pair of workers according to the joint probabilities indexed by their responses to the question. The paper provided synthetic experiment results suggesting that by considering heterogeneity in question difficulty, the proposed mechanism achieved improved incentives for workers to be truthful and was less sensitive to their collusions compared to the previous mechanisms.

2.7.2 Gamification in Session-Level Contexts

The level mechanism is the most common game element that leverages differences in the difficulty of the questions for motivating the crowd-workers. In gamified crowdsourcing, the mechanism sets up different difficulty levels for the questions to be answered in a task so that the workers progress from the easiest level to the hardest level to finish the task. In this case, proceeding to a higher level

that contains more difficult questions requires the workers to exert more effort and show higher levels of expertise. In Dumitrache et al. [143], in addition to the scoring mechanism, the level mechanism controls the timing of when to change the difficulty levels of the medical documents used for fact elicitation for each worker according to the estimate of their current expertise (based on their current scores). The mechanism was appreciated by the workers with 50% of them praising the level progression.

The level mechanism has played a similar role in Lee et al. [154] and Saito et al. [147] where higher worker scores trigger higher difficulty levels in the game for the workers to play. In Dontcheva et al. [153], variants of the level mechanism were proposed, namely the *mission* mechanism and the *challenge* mechanism, to inspire crowd-workers to learn and develop new image editing skills and meanwhile complete the editing tasks posted by the requesters, which requires utilizing the skills they have learned. The mission mechanism issues increasingly difficult sets of questions packaged in the form of increasingly advanced sessions of training for workers to answer in order to improve their skills. Once a worker has successfully completed a session, she can proceed to a more sophisticated one. The challenge mechanism lists all of the image editing tasks from the requesters with difficulty levels matching the current skill level of the worker, so that she can select any of them that interests her.

2.8 Modelling Worker Expertise, Question Difficulty and Contexts

The quality control methods that model the interactions between the three crowdsourcing aspects of worker ability, question difficulty and context primarily extend the GLAD model described in section 2.3.1. The modules added to the GLAD model account for the correlations between the quality of responses and the different levels of contexts which had previously not been factored into the interaction between expertise and difficulty that is modelled by GLAD.

2.8.1 Modelling Interactions in Task-Level Contexts

In this case, the considered context in which the interactions between the worker ability and the question difficulty take place is the whole crowdsourcing marketplace. Thus, the corresponding quality control methods utilize the response information from the same workers across various source crowdsourcing tasks in which they have participated to improve the estimation of the quality of their answers in a target task where these answers are generally sparse. In the Machine Learning

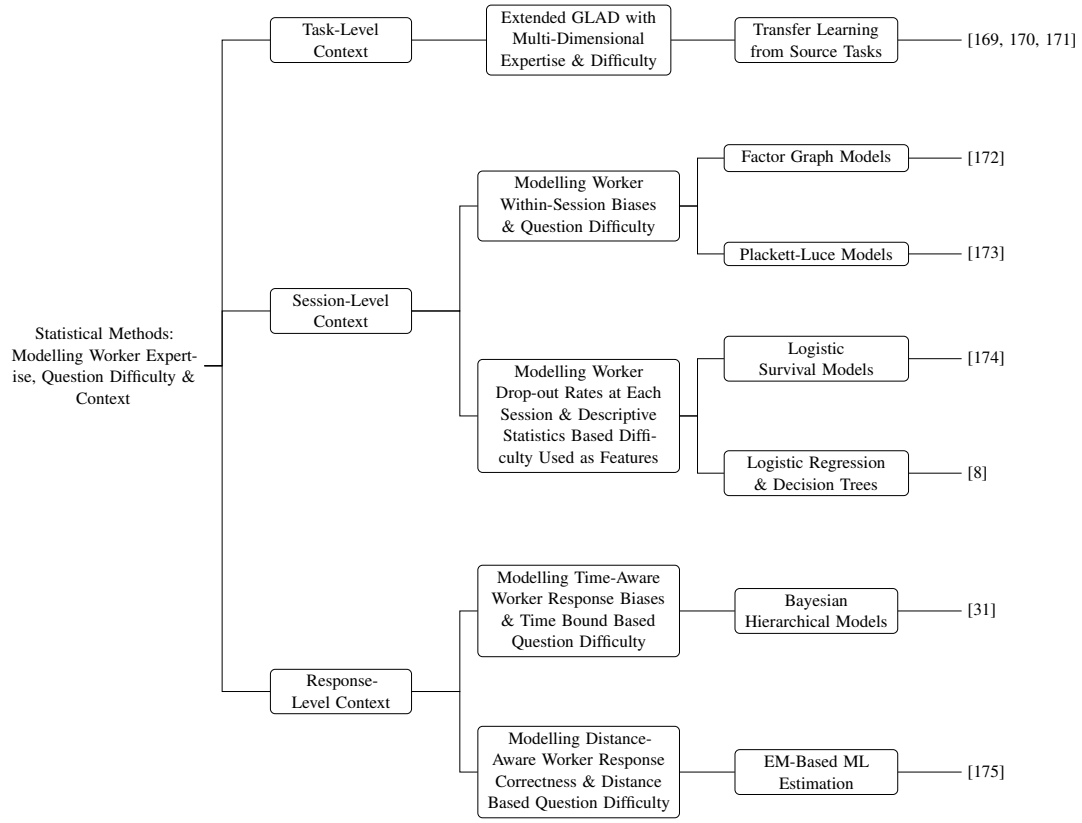


Figure 2.12: A taxonomy of QCC papers that considered worker expertise, context and question difficulty. These papers focused on statistical modelling of how worker expertise varies with contexts at specific levels, and the interaction between the context-aware expertise and the difficulty.

literature, the transfer of response information from a source task to a target task is referred to as transfer learning [176].

In Mo et al. [169], the GLAD model was extended to be multi-dimensional in both worker ability and question difficulty (i.e. workers could exhibit expertise in certain domains). Since such a multifaceted model has a larger number of parameters, it is intrinsically more vulnerable to few responses on both the worker and the question sides. The authors thus decided to transfer crowd-worker information from source tasks to target tasks based on estimated similarity of the tasks. This cross-task transfer of estimated worker expertise was able to smooth out the unreliable response information especially in cold-start situations with few responses per workers. However, the source-task response information was unexploited in this work for calibrating the estimation of question difficulty in the target task. In contrast, in the work of Fang et al. [170, 171], the transfer learning helps to learn a better latent feature representation for each question or data item in the target task from the observed features of the questions in the source tasks. The learned latent feature representation for each target question can be interpreted as the multi-dimensional difficulty of the

question. Active learning was also employed to determine which question to query next given its augmented feature representation estimated through cross-task transfer learning.

2.8.2 Modelling Interactions in Session-Level Contexts

The literature has also considered the effect that each session/page has on the interaction between worker ability and question difficulty. The corresponding quality control methods learn patterns of responses as latent (bias) variables specific to particular (types of) sessions. Zhuang and Young [172] does this by learning a factor graph model which encodes biases in workers' responses to questions within each session. The corresponding factor function is defined as the exponent of a linear regression over counts of different responses (within a particular session). The regression coefficients are global, meaning that they encode a latent bias structure shared across the sessions. This latent structure is inferred to map the response count distribution over each session to a bias value which "offsets" the response accuracy (determined by the interaction between worker expertise and question difficulty) within the particular session. In their another work [173], Zhuang et al. assume that a worker annotates a data item within a session either *independently* from the other items or *relatively* according to a ranking of the items' response correctness probabilities (determined by their difficulty). The ranking is inferred using the Plackett-Luce model [177]. The top-N items in the inferred ranking are considered to be the ones that are responded correctly. The parameter N, which is smaller than or equal to the number of questions within a session, is estimated using maximum likelihood estimation from a categorical distribution over all the sessions.

In Kobren et al. [174], a worker drop-out modelling framework was proposed which consists of a sequence of logistic survival models each corresponding to a particular session/task page. Each model determines probabilities of workers surviving a particular task page and moving to the next. The logistic coefficients for a model map features about workers (e.g. average response time, response accuracy over control questions, etc.), questions (e.g. difficulty, skip rate, average response time, etc.) and consecutive pairs of questions (e.g. same topic or not, their average skip rate, etc.) on the corresponding page. The framework is also integrated with active learning in which each survival model determines both the next page assignment and the next goal (in terms of badges) assignment for each worker's session sequence. The authors showed in the experiments that both of these assignments improved the prediction of question true answers. Similarly, Mao et al. [8] endeavoured to predict the survival rates of the workers in their respective sequences of sessions using session-specific classification models such as logistic regression and decision trees.

The side information features used by these models included the worker's dwell time on a page, the entropy of her responses from past sessions, the number of past sessions, the average response time for past sessions, etc.

2.8.3 Modelling Interactions in Response-Level Contexts

We now consider how features that describe the context under which each response is made, such as response duration, order and location, can be used to improve the estimation of response quality. In Venanzi et al. [31], the response time was aggregated in a Bayesian manner across the responses to each question to obtain the lower and upper bounds of an acceptable response duration for each question. The inferred bounds not only indicate the difficulty of the questions (i.e. higher upper bounds suggesting more difficult questions), but also help detect spam responses (i.e. abnormally long or short response time compared to the bounds). In Hu et al. [175], the quality of a worker's answer to a spatial question (e.g. labelling point of interests) is jointly determined by three factors: the worker's intrinsic quality (i.e. the probability of her being reliable[4]), her (response-specific) location-aware quality assuming that she is reliable, and the difficulty influence of the question on the worker. The location-aware quality decays as the worker's response location moves farther away from the question. The question's difficulty increases as its distance from the worker becomes larger. Shared by both the workers and the questions, the decay factor is treated as one of the model parameters over a finite set of ordinal values and estimated during the model inference.

2.8.4 Leveraging Question Difficulty in Adaptive Question Allocation

Research on incorporating question difficulty into question allocation mechanisms in crowdsourcing has only focused on the adaptive allocation. In this case, the mechanisms are informed of the characteristics of both the workers and the questions to be able to better balance the total cost and accuracy of the true answer prediction.

In Yan et al. [23], which focuses on binary-response questions, the uncertainty of the correct response l_j for question j is modelled as the squared Euclidean distance between 0.5, denoting a random response, and the correct response prediction from a binary classifier. The classifier is based on a logistic function which has global coefficients including an intercept term, and receives inputs which are the observed features of the question (i.e. $\boldsymbol{x}_j$ in equation 2.7). At each time

[4] An unreliable worker randomly responds to questions and thus has a correctness probability on binary responses of 0.5.

point, a question with the greatest uncertainty (i.e. the minimum squared Euclidean distance) is selected for assignment. Each worker is also represented by a logistic function with worker-specific coefficients (i.e. $\boldsymbol{w}_i$ in equation 2.7). The selected question is then assigned to worker i who is able to maximize the probability of seeing the response r_{ij}:

$$P(r_{ij}|\boldsymbol{w}_i, \boldsymbol{x}_j) = \delta(\boldsymbol{w}_i, \boldsymbol{x}_j)^{\mathbb{1}\{r_{ij}=l_j\}}\big(1 - \delta(\boldsymbol{w}_i, \boldsymbol{x}_j)\big)^{\mathbb{1}\{r_{ij}\neq l_j\}} \tag{2.9}$$

where $\delta(\boldsymbol{w}_i, \boldsymbol{x}_j)$ is defined by equation 2.7.

In Ho et al. [178], workers were first asked to answer a set of test questions (with known responses) in different areas to estimate their expertise levels in those areas. The quality of a worker's response to a new question is then estimated by the difficulty of the corresponding area, and the estimated expertise levels of the worker in that area. The optimization of the question assignment is then modelled as an integer programming problem with the objective being to minimize the total number of responses collected, and thereby the overall cost. The total quality of the responses given to the corresponding assigned questions serves as a constraint in the optimization: $\sum_i \sum_j o_{ij} q_{ij} > 2\ln(1/\epsilon)$. The term ϵ is set to be a value smaller than one, the binary variable o_{ij} denotes whether question j is assigned to worker i, and the term $q_{ij} = (2P(r_{ij} = l_j) - 1)^2$, indicating the uncertainty of the response r_{ij}.

In Khetan and Oh [179], the authors drew inspiration from their previous work [66, 68, 69] which modelled question allocation as a regular random bipartite graph. Rather than construct such a graph before any worker arrives in the task, the authors in this work constructed the graphs over multiple rounds during the task. In each round, the allocation mechanism optimizes a graph for achieving a target accuracy (on true answer prediction) with respect to the degree of the question nodes. Any question with the prediction accuracy higher than the target one has its predicted true answer accepted while the rest of the questions proceed to the next round. The total number of edges in the graph represents the fraction of the remaining budget spent at the particular round. Rounds stop when all the questions have their predicted true answers accepted or the budget is exhausted. The average estimated difficulty of all the remaining questions before a particular round controls the target accuracy to be achieved at that round. Finally, the authors proved that using regular random graphs with the proposed adaptive mechanisms guarantees near-optimality in the rate of convergence to the target accuracy.

2.9 Modelling Worker Expertise, Question Difficulty and Response Relationships

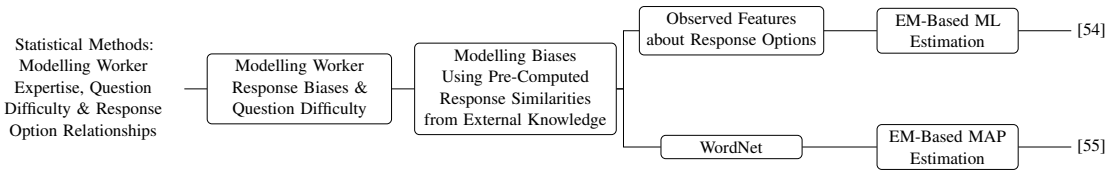


Figure 2.13: A taxonomy of QCC papers that considered worker expertise, question difficulty and semantic relationships between responses. These papers focused on statistical modelling which leveraged pre-computed response similarities from external knowledge to account for response biases.

It is not unusual in crowdsourcing tasks that response options are correlated with one another due to semantic relationships that exist between the options. An example is *object identification* in images where in the ImageNet[5] dataset a large number of label categories are related through the semantic relationships that can also be found in WordNet[6]. The Stanford dog breed identification task [180, 181] uses the "Dog" subset of WordNet as the set of the response options which involves 120 dog breeds. These breeds are related to one another to various degrees (e.g. Labrador being much more closely related to Golden Retriever than it is to Chihuahua).

When semantic relationships between response options are present in crowdsourcing, current quality control methods measure the relationships using some *distance* metric such that options with closer relationships have smaller distances. A common distance metric used by the state-of-the-art methods is the *length* of (or equivalently, the number of edges in) the *shortest path* between two response options in the semantic structure (e.g WordNet). The distances can be computed independently from the crowdsourcing tasks that provide the set of response options to their questions. This also means that the quality control methods will have access to the pre-computed distances when estimating the quality of each response.

Two papers have considered leveraging semantic relationships between response options [54, 55] for improving quality control of crowdsourced responses. The core idea of both papers is to use the precomputed semantic distances to calculate response similarities which indicate possible correlations in workers' responses (both correct and incorrect responses) as shown by Figure 2.14.

In Han et al. [55], a model was proposed in which the probability of each response option a worker could give to an item is conditioned on its true answer and provided by a *soft-max* function.

[5] http://www.image-net.org/
[6] https://wordnet.princeton.edu/

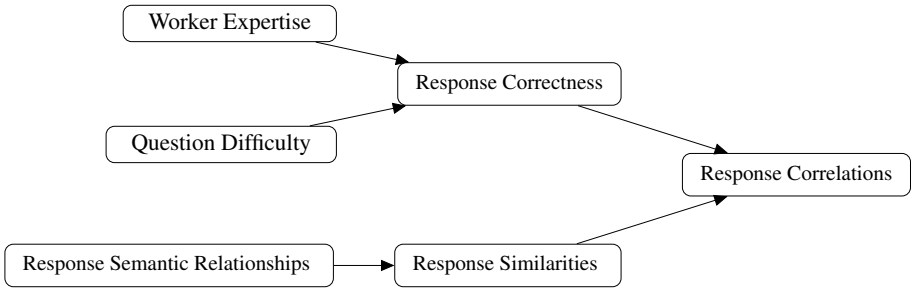


Figure 2.14: A diagram shows how worker expertise, question difficulty and response semantic relationships contribute to the correlations within responses.

This function takes in the normalized distances between each response option and the correct response, along with the question difficulty and the worker expertise. The difference between the difficulty and the expertise is scaled by the normalized distance before computing the corresponding conditional probabilities. Due to the scaling effect, the log-odds of the probabilities are *inversely proportional* to the normalized distances. The larger the distances are, the smaller the probabilities. In other words, a response option that is inherently less related to the correct answer is less likely to be selected irrespective of the worker or question.

In Fang et al. [54], the authors proposed a model which shares the same idea as Han et al. [55] except that *similarity scores* between response options are pre-computed as the inverse of the Euclidean distances between the options in terms of their *observed features*. Both of these models rely on the availability of the external knowledge about the relationships between response options, while models that can automatically infer such relationships from the responses are still missing.

2.10 Conclusion

In this chapter, we reviewed past research which studied quality control for crowdsourcing. The research focused on either *quality control designs* or developing *statistical methods*. Quality control designs are conducted for mechanisms that control the context in which workers are situated to positively affect their question-answering behaviour (e.g. to answer more accurately and quickly). Quality control statistical methods aim at inferring key attributes of important crowdsourcing aspects which are modelled to determine the quality of responses, and further the question true answers. To the best of our knowledge, this survey is the first to clarify the relationship between quality control designs and statistical methods. The latter supports the former by providing descriptive statistics or estimates of the attributes as inputs to the designed mechanisms.

To conduct a systematic and in-depth review of the QCC research, we started by organizing all the aspects of crowdsourcing and the primary attributes considered by the research into a taxonomy. To the best of our knowledge, this survey is the first to provide such an organisation. Based on the taxonomy, we then designed a graph that visualizes the two types of QCC methods (i.e. quality control designs and statistical methods) and the aspects and attributes of crowdsourcing that have been considered in previous work. This graph allows us to easily present to the readers the different lines of QCC research, and also facilitates the arrangement of different sections in our survey.

Based on the QCC research graph, we carried out the survey, starting from the research considering only worker ability to decide the response quality, and ending with those considering multiple attributes from different aspects to jointly determine the quality. We dedicated not only a section to each research but also a taxonomy that summarizes the technical details of the QCC papers from it. The taxonomy includes considered attributes, modelling assumptions, parameter estimation techniques, design features, etc. To our best knowledge, we are the first to provide such intuitive taxonomies to QCC researchers to inform them of the research frontiers and facilitate their identification of new research directions.

As the final step for concluding our survey, we report the major issues of both quality control designs and statistical models found during the literature review, and propose corresponding research directions that aim to resolve these issues. In the next chapter, we convert these directions into a series of research questions to be addressed in this thesis.

2.10.1 Payment, Gamification and Training Designs

The design of incentive mechanisms can be categorized into the *payment* (function) design, the *gamification* design and the *training* design. The three design types describe how the mechanisms should intervene into the contexts to affect worker motivation and expertise to improve the response quality. According to our survey, we claim that both the training design and the payment design have been studied by the past QCC research. For the training design, only one paper by Bragg et al. [30] has been found. This paper has successfully addressed the fundamental issue in the area which is how to replace the traditional pre-task quiz training with more flexible and personalized training sessions for individual workers. They addressed the issue elegantly using reinforcement learning and decision-theoretic approaches. As for the payment design, most techniques rely on game-theoretic approaches whose modelling focus has progressed over the years from interactions between motivation and context, to also considering question difficulty. Modelling motivation,

context and question difficulty all together provides the state-of-the-art and the most sophisticated methods for designing payment functions.

2.10.2 Issues with Designing Gamified Crowdsourcing Tasks

We believe that *gamification* design requires much more research attention. This is because the current gamification methodologies are rather new to crowdsourcing compared with the payment design methodologies, which are theoretically sound since they are predicated on game theory. We have found from our survey that most papers on gamifying crowdsourcing tasks rely on either authors' impressions about various game elements or conventions from video games to draw the designs for their tasks. This suggests the need for empirical insights into both the individual and joint effects of the various game elements on the quality of worker responses in crowdsourcing applications, such as relevance judgement for Information Retrieval, image annotation, etc. These insights will be critical for constructing a common gamification methodology in those areas. The methodology provides empirically justified guidelines on what combinations of game elements should be used (possibly together with other mechanisms) to improve the quality of responses.

- ⋆ **Future direction 1**: Empirical studies on what combinations of game elements should be incorporated into *paid* crowdsourcing tasks to improve the quality of worker responses.

2.10.3 Statistical Modelling and Estimation

According the literature review, we found that a large part of QCC research has focused on *statistical modelling* and *estimation* of attributes of various crowdsourcing aspects assumed to jointly determine the response quality. These aspects include crowd-workers, questions, various levels of contexts and (semantic relationships between) response options. The estimation can be carried out during the crowdsourcing (e.g. active learning or reinforcement learning) or subsequent to the response collection from the crowd (i.e. via post-crowdsourcing wisdom-of-the-crowd aggregation). We also found that the estimation of these attributes has been involved with certain issues in the past QCC research.

2.10.4 A Common Issue of Attribute Estimation

Sensitivity to the *sparsity* in responses has been a common issue for the estimation of attributes like worker expertise and question difficulty which further adversely affects the estimation of response

quality and true answers. If few responses are collected per worker and per question, estimating parameters from them and using their values to predict the response quality becomes statistically unreliable. To resolve this issue, previous QCC research has focused on *calibrating* the parameters in two ways:

- The first way is to build hierarchical models that merge [10] or group [73] the response information from the considered individuals (e.g. workers, questions, contexts, etc.) in a Bayesian manner. The models use such merged or grouped response information to smooth the individual's response information during the estimation of their associated parameters to achieve the calibration.

- The second way is to leverage the side information about the individuals which potentially groups their response information based on the assumption that individuals with similar side information should share similar values for their attributes (e.g. expertise, difficulty, etc.) and response information. The grouping in this case can again be conducted using the same Bayesian hierarchical modelling as the previous work, which ignored the side information. The QCC work that considered side information has reported superior true answer prediction performance compared to systems not considering such information [31, 91, 94, 169, 174].

In the literature, different types of side information have been leveraged by statistical models for quality control, including information about worker demographics [91], question content [94], and different levels of contexts [31, 169, 174].

Unfortunately, all these quality control methods focus on a single type of side information, and none of them has managed to incorporate *all* types of side information into a *single coherent framework*. Consequently, the proposed modelling and estimation frameworks have been *ad-hoc* and specific to certain types of side information and thus cannot be generalised to cover other types of side information.

- ⋆ **Future direction 2**: A more *unified* and *scalable* quality control framework is needed that can incorporate arbitrary types of side information and utilize it to better control the quality of crowdsourced responses. Such a framework would be particularly important for improving true answer prediction when only a few responses are collected per worker and per question.

Apart from the common issue of being sensitive to response sparsity, the learning can face other issues specific to particular types of interactions involving workers and questions. According to

the literature, notable issues arise when *worker-question* interactions and *worker-question-answer* interactions are being inferred.

2.10.5 Issues of Modelling and Estimation with Partially Subjective Questions

When a crowdsourcing task involves *partially subjective* questions, the current statistical modelling of worker-question interactions becomes problematic and the response quality estimation based on it becomes unreliable. More specifically, current modelling techniques are unable (i) to distinguish the *difficulty* of a question from its *subjectivity* and (ii) to distinguish expert workers with distinct *subjective opinions* regarding the true answer from either misled or adversarial workers who give objectively incorrect answers. The reason is that the current modelling techniques assume all questions to be completely objective (with a single correct answer). Consequently, their corresponding inference procedures will falsely increase the difficulty of a partially subjective question every time they receive a correct answer that is different from the single latent correct answer as determined by the estimation algorithm up to that moment. Eventually, the question can be misjudged by the estimation algorithm as being very difficult (or even deceptive) while most of the variation in its received responses is actually due to subjectivity (e.g. where 4 out of 5 response options are indeed correct). Likewise, the inference also penalizes the ability of any worker who gives a different response from the one deemed to be correct for the question by the model at that point. As a result, an expert worker, who has answered all the objective questions correctly, could still be misjudged to be novice or even adversarial by the estimation procedure if she happens to have preferences that differ from the majority of the workers for the partially subjective questions.

- ⋆ **Future direction 3**: More effective modelling of worker-question interactions is required for partially subjective questions which should (i) distinguish question *difficulty* from question *subjectivity* and (ii) distinguish expert workers with different *subjective assessments* of true answers from either misled or adversarial workers who always give different answers from the true answers.

2.10.6 Issues of Modelling and Estimating Response Semantic Relationships

Only two papers by Han et al. [55] and Fang et al. [54] have modelled the correlations of response quality with worker ability, question difficulty and the *semantic relationships that may exist between response options*. The models proposed by these papers incorporate the response relationships by

leveraging the relatedness/similarity scores between the response options which were pre-computed from external knowledge graphs/hierarchies such as WordNet. This means that no previous work has attempted to jointly learn the worker ability, the question difficulty and the (latent) relationships that exist between the response options directly from the crowdsourced responses. The fourth

Moreover, the previous papers have assumed that response options which are more related to the correct answer for the question are more likely to returned irrespective of worker ability and question difficulty. Whether this is the case in practice remains to be investigated.

- ⋆ **Future direction 4**: Effective modelling and learning is needed for the joint interactions between the worker ability, the question difficulty and the semantic relationship between response categories.

Chapter 3

Research Questions

Based on the research directions identified in Section 2.10, in this chapter we propose a corresponding set of research questions to these questions to advance the QCC research in the particular directions. Each proposed question consists of multiple sub-questions that need to be addressed. The following chapters of this thesis are dedicated to answering each of the research questions proposed in this chapter.

3.1 Research Question 1

The first research question is proposed under the research direction 1 identified in Section 2.10.2 that deals with issues of designing gamification mechanisms. This direction focuses on empirical studies of *what* combinations of game elements need to be incorporated into paid crowdsourcing tasks for improving the quality of worker responses. The state-of-the-art work in this regard has focused on voluntary crowdsourcing [11] which bears significant difference from the paid crowdsourcing.

In this thesis, we lay the foundation for future work under research direction 1. We focus on exploiting the most significant part of games: *competition*. A competitive game is usually characterized by two elements: the scoring mechanism (more precisely, the provision of *performance feedback* to competitors), and the leaderboard. We investigate the individual and combined effects of these two elements on the response quality. We also investigate how the effects change when there is a bonus payment or control question based filtering of crowd workers. The first research question is thus the following:

- ***RQ*1**: In order to maximise response quality for a given crowdsourcing budget, which *competition* elements should be leveraged for gamifying crowdsourcing tasks: a *real-time*

performance *score* for each crowd-worker, a performance leaderboard accessible by any worker in the task, or a combination of these competition elements?

To answer this research question, we need to first answer the following sub-questions:

- ***RQ*1.1**: Does providing real-time performance feedback alone affect the average performance of the crowd-workers? And if so, is their performance improved or worsened?
- ***RQ*1.2**: Does *additionally* providing an all-time leaderboard along with the real-time performance feedback to the crowd-workers affect their average performance? And if so, does it make their performance better, worse or both (i.e. more varied) than the performance achieved alone by the feedback?
- ***RQ*1.3**: Is the performance of the workers from *RQ*1.2 affected by the *bonuses/prizes* to the best performers?
- ***RQ*1.4**: Are the effects observed for *RQ*1.3 affected by the use of *worker filtering* using *control questions*?

We will tackle the research question *RQ*1 and all its sub-questions from *RQ*1.1 to *RQ*1.4 in Chapter 4.

3.2 Research Question 2

The second research question is proposed under research direction 2 that tackles the issue of learning from *sparse responses*, i.e. where few responses are collected per worker and per question. In this case, estimation of the parameters corresponding to the attributes of different crowdsourcing aspects, typically *worker expertise* and *question difficulty*, becomes unreliable.

We pointed out in section 2.10 that the state-of-the-art statistical models focus on only a specific type of side information. Each of them provides ad hoc techniques for modelling the particular type of information and using it to improve estimation of the associated attributes. Therefore, to explore this direction, we envision an unified scalable framework. This framework should be able to handle arbitrary types of side information and utilize them to calibrate the parameter estimation of worker expertise and question difficulty, the two most frequently modelled attributes. To achieve this, the second research question is proposed as follows:

- ***RQ*2**: Can we build a *unified scalable statistical modelling and inference framework* that is able to integrate and utilize arbitrary types of side information about different crowdsourcing

aspects for better controlling the quality of crowdsourced responses, thereby improving the true answer prediction under *response sparsity*?

To answer this research question, we need to first answer its four sub-questions:

- ***RQ*2.1**: How should one design or choose a *basic* quality control framework which possesses the capability of plugging in arbitrary types of side information?
- ***RQ*2.2**: How can one best incorporate the different types of side information into such a basic framework?
- ***RQ*2.3**: Once incorporated, which (combinations of) types of side information are the most useful in general for predicting the correct answers of questions under response sparsity?
- ***RQ*2.4**: To what degree of sparsity can such (combinations of) types of side information improve the true answer prediction under the framework?

We will tackle the research question *RQ*2 and all its sub-questions from *RQ*2.1 to *RQ*2.4 in Chapter 5.

3.3 Research Question 3

The third research question is proposed corresponding to research direction 3 that aims at distinguishing question *difficulty* from question *subjectivity* in terms of their effects on the worker-question interactions. Exploring this research direction is useful for better modelling and learning the quality of responses given to the *partially subjective* questions which are characterized by both difficulty and subjectivity, and are not unusual on crowdsourcing marketplaces in practice. For example, as a popular type of crowdsourcing tasks, relevance judgement in information retrieval is widely recognized as being partially subjective in nature [182]. The third research question asks the following:

- ***RQ*3**: Can we better control the quality of worker responses, thereby improving true answer prediction for *partially subjective* questions, by distinguishing question *difficulty* from question *subjectivity* in the modelling and inference of the worker-question interaction which generates the responses?

To answer this research question, we need to first answer the following sub-questions:

- ***RQ*3.1**: How do the difficulty and the subjectivity of questions each affect the interaction between the questions and the workers which generates the responses?
- ***RQ*3.2**: Does any dependency exist between the two parts of the interaction respectively affected by the difficulty and the subjectivity?
- ***RQ*3.3**: Is a statistical model that encodes this dependency effective in improving the true answer prediction of partially subjective questions?
- ***RQ*3.4**: Can the statistical model yield similar judgements to human experts with respect to the degrees of difficulty and subjectivity of the partially subjective questions?

We will tackle the research question *RQ*3 and all its sub-questions from *RQ*3.1 to *RQ*3.4 in Chapter 6.

3.4 Research Question 4

The fourth research question is proposed under the research direction 4 which aims to build statistical models that can effectively (i) capture the interaction between worker ability, question difficulty and semantic relationships between response options, and (ii) infer the semantic relationships directly from the responses. To the best of our knowledge, none of the current work has accomplished these two goals. More specifically, this research question asks the following:

- ***RQ*4**: Can we build a statistical model that can effectively learn the joint interactions between workers, questions and response semantic relationships for better controlling the quality of responses, thereby improving the true answer prediction, when the number of response options is sufficiently large?

To answer this research question, we need to first answer the following sub-questions:

- ***RQ*4.1**: How can one best capture the semantic relationships between response options in a statistical model?
- ***RQ*4.2**: Once the semantic relationships have been encoded, how to further capture its interactions with worker ability/expertise and question difficulty in the model?
- ***RQ*4.3**: How can external domain knowledge about the semantic relationships be incorporated so that it can be useful for learning the latent representation of the relationships in the model?

- ***RQ*4.4**: What is the quality control efficacy of the model that encodes the relationships and the corresponding interactions with and without the external knowledge?

We will tackle research question *RQ*4 and all its sub-questions from *RQ*4.1 to *RQ*4.4 in Chapter 7.

Chapter 4

Quality Control Designs with Leaderboards and Performance Feedback

This chapter is dedicated to answering the first research question (i.e. *RQ*1) by addressing all its sub-questions (i.e. *RQ*1.1 to *RQ*1.4) based on carefully designed empirical studies. The studies will shed light on how to design effective gamification tasks that incentivize workers to be more accurate. The gamification design proposed in this chapter leverages the *competition* among workers with *performance feedback* and a *leaderboard* accessible to every worker at any time. The studies address whether these competition elements are best used individually or in combination with one another. They also address how the previous results change when the competition mechanism is combined with the *bonus payment* to top competitors and the filtering of *low-performing* competitors using *control questions*. The studies consist of a series of *controlled tests* [183], from which we show that displaying a leaderboard to workers can motivate them to improve their performance, providing a bonus is offered to the best performing workers. This effect is observed even when control questions are used to enforce quality control during task completion. We also show that it is straightforward to augment a crowdsourcing task with real-time feedback and leaderboards while preserving privacy and anonymity of crowd-workers.

4.1 Related work

The closest work to this chapter was done by Ipeirotis and Gabrilovich [11]. The authors leveraged the scores as performance feedback to each worker, and a task leaderboard. They found that such a leaderboard tends to discourage latecomers to the games by showing them accomplishments of well-established workers who have already answered many questions. They suggested to use one leaderboard for each Web-page of the task providing that the number of questions are the same and much fewer on each page. However, the game elements in this work were dedicated to *unpaid* crowdsourcing involving only online volunteers, while our work studies paid crowdsourcing with bonus incentives rewarded to crowd-workers. Therefore, whether a task-level leaderboard can negatively affect the performance of workers in paid crowdsourcing remains to be investigated in this chapter.

Based on our literature review from Section 2.5.4 to Section 2.5.7 that concern designing gamification mechanisms, we are unaware of any other work in paid crowdsourcing that has *empirically* studied how to apply *competition games* among *individual* workers with *performance feedback* and *task leaderboards* to crowdsourcing.

4.2 Experiments on Gamification Designs with Performance Feedback and Leaderboards

We perform three experiments to address the research sub-questions from *RQ*1.1 to *RQ*1.4. The particular type of tasks we consider here is *relevance judgement* of retrieved *documents* for matching the Web search *queries* in the area of *Information Retrieval* (IR). The reason for choosing to perform the empirical studies on this type of task is due to the long-time application of crowdsourcing to the construction and evaluation of the IR systems (e.g. learning-to-rank systems) [184]. In this case, quality control of crowdsourced relevance judgements is critical for making the constructed IR systems reliable.

We made use of the "assigned-judged.balanced" data subset[1] from the TREC 2011 crowdsourcing track[2] with 750 expert judgements provided by NIST[3] assessors for 60 queries and the ClueWeb09 collection[4] from which the documents judged for relevance were drawn.

[1]This data subset is balanced across three relevance levels (i.e. highly relevant, relevant and non-relevant)
[2]`https://sites.google.com/site/treccrowd/2011`
[3]National Institute of Standards and Technology at https://www.nist.gov/.
[4]`http://lemurproject.org/clueweb09/`

In each experiment, we collected relevance judgements for a crowdsourcing task undertaken by workers on the CrowdFlower platform. Within each task, we performed A/B/n testing of some control and treatment groups. We followed standard techniques for online experimentation [183], randomly dividing workers as equally as possible across different control and treatment groups in each task. Moreover, we randomly divided the chosen data subset into three disjoint sets of document-query pairs in order to prevent crowd-workers who might partake in multiple experiments from ever judging the same document-query pair twice. Within an experiment, the same set of query-document pairs were shown to all crowd-workers across (and within) the control and treatment groups in order to remove variability due to differences in the judgement difficulty for different query-document pairs.

The CrowdFlower platform is designed for ease of use in creating the crowdsourcing tasks, but not for the A/B/n testing within each task. Thus in order to perform controlled experiments we employed the experiment architecture shown in Figure 4.1. The crowd-workers first join our tasks held on CrowdFlower platform, causing a task page to load on their browser. At this point, client-side Javascript sends a request for additional page content to the A/B/n testing server. The server then randomly allocates each individual to one of the control and treatment groups for the duration of the experiment[5], and returns content to be inserted into the task page.

The inserted content provided different information to crowd-workers in different experimental groups. For example, the task page shown in Figure 4.2 contains the performance statistic (e.g. answering accuracy) for the worker and her own view on the shared leaderboard, while the task page shown in Figure 4.3 (for a different group) contains no inserted content. The performance statistics for each crowd-worker is updated by the server based on the ground-truth relevance judgements from NIST assessors at the completion of each task page (consisting of 5 relevance judgements). The crowd-worker's view on the leaderboard is updated each time they load *or refresh* the task page. The leaderboard shown for the various experiments contained: (i) the rank of the crowd-worker, (ii) anonymous usernames for each of the crowd-workers, (iii) the score for each worker (computed using metrics to be specified in the following sections), and (iv) the change in the rank of each worker since the previous refresh. Crowd-workers were free to quit a task at any point, sometimes even without making any judgement, meaning that the number of workers can differ across different groups due to both random assignment and self-removal. We adopted the setup used in the TREC 2011 crowdsourcing track where a worker judged the relevance of 5 documents

[5] We used random assignment rather than sequential assignment to prevent biases due to the order in which individuals join the tasks.

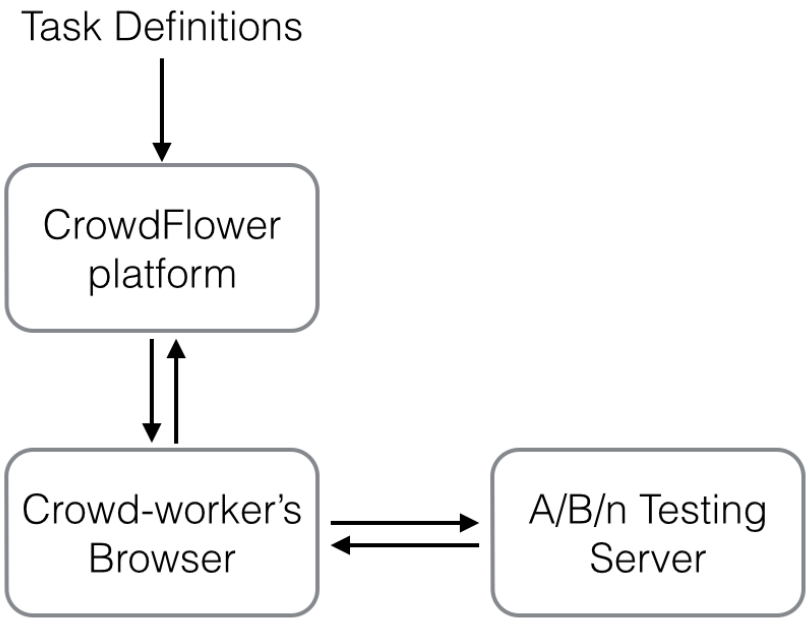


Figure 4.1: Architecture of experiment environment. Crowd-workers are randomly assigned to a particular group after joining the task on CrowdFlower, and see the same version of the interface for each subsequent interaction from their browsers. The server updates their performance statistics (or rank positions) once each page of judgements was completed.

per query. The 5 documents per query were embedded as images in one page in random order for all the tasks except the one which will be specified in Section 4.2.3.

4.2.1 Experiment 1: Real-time Feedback

In the first experiment, we investigated whether providing real-time performance feedback to crowd-workers affects their performance. To do this, we randomly assigned crowd-workers to one of four groups:

- **Control Group:** *No performance feedback* was provided to the workers (see for example Figure 4.3).
- **Treatment 1:** The workers were told their *current estimated accuracy* and the *current average estimated accuracy of all the workers* in the same group. (The latter information was provided for anchoring purposes, so that the workers knew whether their accuracy scores were relatively high or low.)
- **Treatment 2:** The performance feedback from treatment 1 was preserved in this case and in addition, directly underneath the feedback area, an *all-time task leaderboard* was provided to the workers showing their current ranking in the group scored by their estimated accuracy.

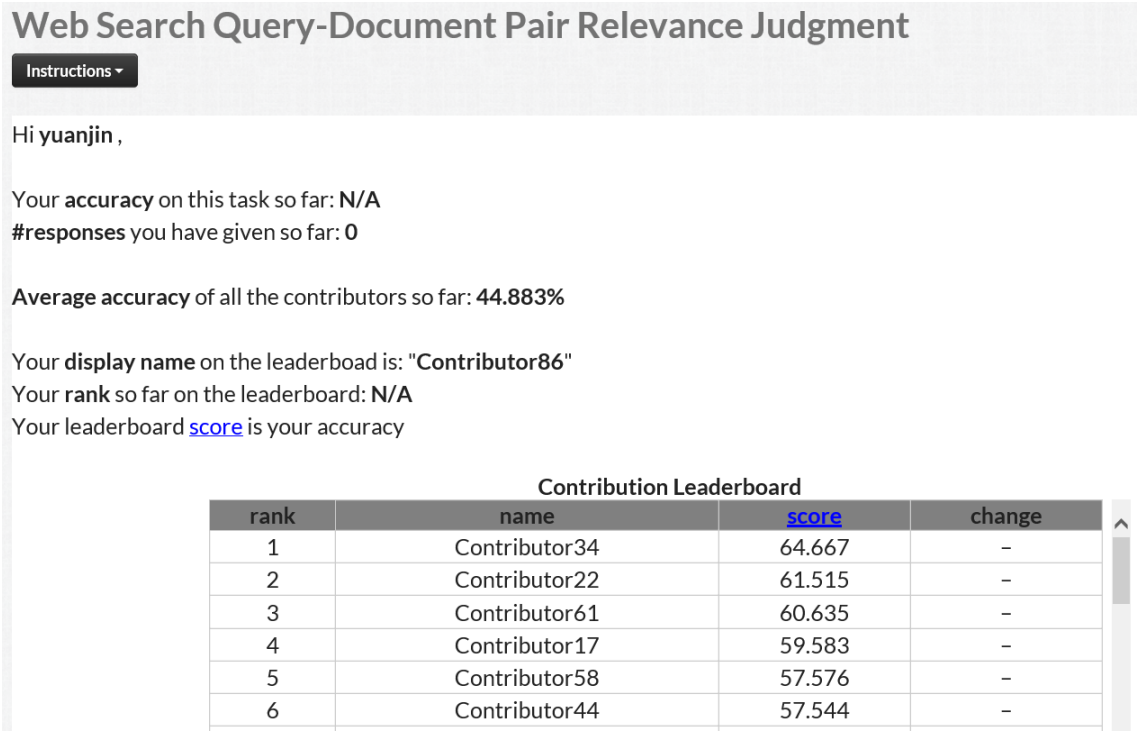


Figure 4.2: A screenshot of the interface shown to crowd-workers assigned to Treatment group 2 for Experiment 1, containing a leaderboard ranking contributors based on their labelling accuracy in percentage.

On the board, workers were referred to as "Contributor$i$" where i denoted the order in which they were assigned to this group (see for example Figure 4.2). Thus no personally identifying information (such as CrowdFlower usernames) was leaked across contributors. Moreover, workers were able to see the changes in their ranking every time they reloaded current pages or loaded new ones.

- **Treatment 3:** The setting here was exactly the same as in treatment 2 except the scores for the workers on the leaderboard was changed to be the product of their accuracy times the number of responses they have made. The idea was to encourage workers to keep annotating in order to move up in the ranking on the leaderboard.

For this experiment, each task consisted of labelling 5 documents for the same query as {*highly relevant*, *relevant*, *non-relevant*}. Crowd-workers could complete a maximum of 20 tasks each (i.e. label 100 documents) and each task was made available to 40 crowd-workers. The documents used to build the tasks were randomly sampled from the balanced data subset of the TREC2011 dataset.

In all experiments, we used the term "estimated accuracy", even though we calculated the accuracy based on the ground-truth relevance judgements, in order to be consistent with the general

Figure 4.3: A screenshot of the interface shown to crowd-workers assigned to the control group for Experiment 1, which contained no additional information about the workers' performance.

case where the ground-truth judgements are unknown and need to be estimated. Moreover, we applied Dirichlet smoothing to the accuracy estimate[6]:

$$\hat{e}_i = \frac{(\sum_{j=1}^{N_i} \mathbb{1}(r_{ij} = l_j)) + \alpha\frac{1}{3}}{N_i + \alpha} \tag{4.1}$$

Where N_i is the number responses from worker i, r_{ij} denotes the individual's response to question j, l_j is the ground-truth judgement for the question, and α is a smoothing parameter set to 5 so that the number of pseudo-counts equals the number of questions per page. The main reason for smoothing was to prevent crowd-workers in Treatment 2 from providing a small number of judgements on which they performed unusually well (e.g. judge all 5 documents correctly on the first page) and then quitting the task to win the competition.

Some contributors might have provided fewer relevance judgements than others because they (i) arrived late to the task when few pages were left or (ii) gave up early on. To prevent problems due to poor estimation of worker accuracy and/or weaker effects due to receiving feedback for shorter periods of time, data from workers who had provided less than 50 judgements were discarded across all groups (and were not included in the following analysis).

[6]More formally, we employed an estimate of the posterior mean assuming a Binomial likelihood (over correct/incorrect answers) and a Beta prior with concentration parameter α and prior mean (probability of a correct answer) of $1/3$ due to the balanced classes used.

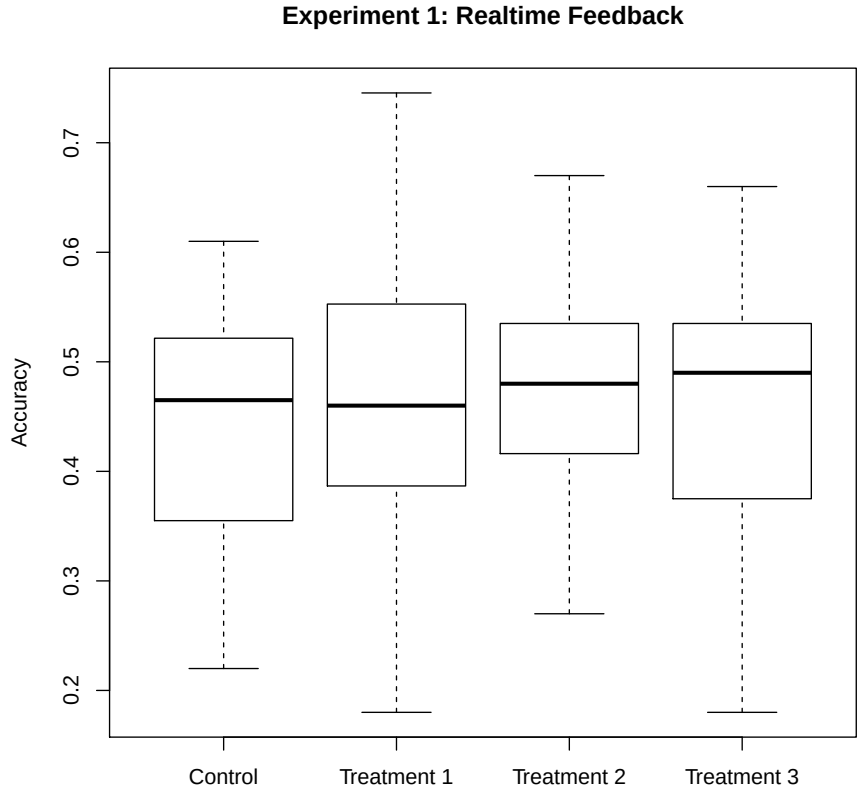


Figure 4.4: A boxplot showing the distribution of Accuracy values across the crowd-workers for different groups in Experiment 1.

Table 4.1 shows experimental results across the four groups measured in terms of both micro and macro-averaged Accuracy[7]. The former is averaged equally across all judgements and the latter equally across all crowd-workers. Due to the differing levels of ability/accuracy of the crowd-workers, the macro-average is the main measure of interest for determining changes in worker accuracy across the groups.

A boxplot in Figure 4.4 shows the distribution (median and inter-quartile ranges) of performance for workers in the various groups. We see that the spread of values is similar across the four groups.

To determine whether the differences in the mean performance across the groups were significant at the worker level, we employed a pairwise t-test (with non-pooled standard deviations and a Bonferroni correction for multiple comparisons). Given the test procedures, we did not find any significant difference between the groups[8], indicating that the effect size for the three treatments is

[7]Accuracy was calculated treating the graded relevance levels {*highly relevant*, *relevant*, *non-relevant*}, as separate classes. Thus the assignment of label *highly relevant* to a document with true label *relevant* is considered incorrect. We repeated the analysis using Accuracy computed over binary relevance judgements, i.e. where the labels *highly relevant* and *relevant* were collapsed into the same class, finding similar results.

[8]The smallest P-value observed (before the Bonferroni correction) was 0.19 between treatment 2 and the control group, so not significant at the 0.05 confidence level even before correction for multiple comparisons.

Table 4.1: Results across the four groups for Experiment 1. Accuracy is given as both Micro and Macro averages, with the latter being aggregated at the *worker level* (i.e. by treating the average performance of each worker as a single observation). Crowd-workers who judged less than 50 documents were excluded from the analysis.

Group	# Workers	# Judgements	Accuracy (Micro)	Accuracy (Macro)
Control	44	3874	45.04%	45.39%
Treatment 1	40	3449	45.26%	43.20%
Treatment 2	48	3947	47.76%	47.86%
Treatment 3	39	3613	46.39%	47.06%

small at best, i.e. the effect on performance that results from additionally providing a leaderboard to crowd-workers is small relative to the variation of performance across workers, and experiments with larger numbers of workers would be required to determine the size of the effect.

We also note from Table 4.1 that the score (the estimated number of correct judgements) used to rank crowd-workers in treatment 3 didn't appear to work quite as well as the score (the estimated accuracy) used in treatment 2 in terms of both micro and macro-averaged Accuracy. Assuming this was indeed the case, we conjecture that the reason for this might have been that new workers started very low on the leaderboard in the former case, and it took them a long time to rise up amongst the leaders, which caused apathy towards higher rankings. A similar conjecture was mentioned in Ipeirotis and Gabrilovich [11].

4.2.2 Experiment 2: Adding a Bonus

Given the results of the first experiment, we believed the crowd-workers needed to be further motivated and therefore moved to investigate whether incentivising the workers by offering to pay them a bonus at the end of the task affected their performance and if so, whether it caused them to perform better or worse.

More specifically, a $1 bonus was offered and paid to those who ended up within the top 10 of the leaderboard. We followed the same setup used in experiment 1 except the number of responses collected for each document-query pair for all the groups was now 30. The control and the treatment groups involved in this experiment are listed as follows:

- **Control Group 1:** *Neither performance feedback nor bonus* was given to the workers assigned to this group.

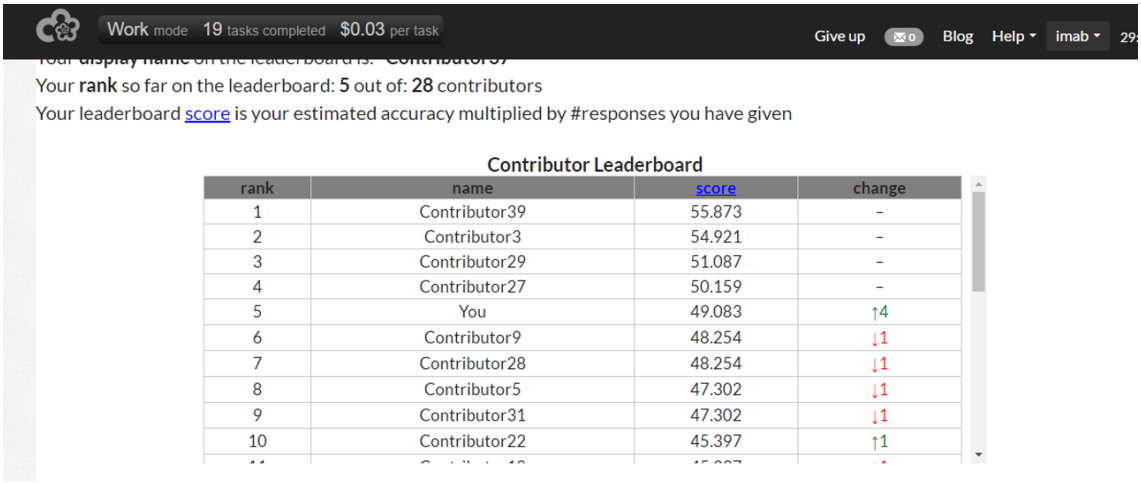


Figure 4.5: A screenshot of the interface shown to crowd-workers assigned to Treatment Group 2 in Experiment 2.

- **Control Group 2:** *No performance feedback* was provided to the workers while they were performing their annotation, but they were informed that the *top 10* workers would be paid a $1 bonus at the completion of the task (once all the workers submitted their responses). Workers were also informed that their performance would be judged according to their *estimated numbers of correct responses*[9].
- **Treatment 1:** the configuration was the same as for control group 2 except that a *leaderboard* was provided to the workers in the group showing their rankings in terms of the *Dirichlet-smoothed accuracy estimate*.
- **Treatment 2:** the configuration was the same as for treatment group 1 except that the scores shown on the *leaderboard* were the *estimated number of correct responses* (i.e. the product of their accuracy times the number of responses), a value that increased as the worker responded to more questions.

The results for the second experiment are shown in Table 4.2 with distributions over worker accuracies shown in Figure 4.6. Of note in the figure is the smaller interquartile range for treatment group 2 with respect to the other groups (or the previous experiment). This may indicate that crowd-workers tend to perform more consistently when competing for bonuses on a leaderboard.

Treatment group 2 exhibits approximately 3% higher accuracy than the other groups for this experiment. However, pairwise T-Test results show no significant difference between the various

[9]Estimated number of correct responses from Worker i is the product of the individual's Dirichlet-smoothed accuracy estimate and the number of responses he/she has made.

Table 4.2: Results for Experiment 2, where a bonus was offered to participants in Control group 2 and Treatment groups 1 and 2.

Group	# Workers	# Judgements	Accuracy (Micro)	Accuracy (Macro)
Control 1	26	2419	41.71%	42.19%
Control 2	28	2624	42.11%	42.59%
Treatment 1	27	2665	43.83%	43.84%
Treatment 2	33	3139	46.54%	46.85%

groups at the 0.05 level after a Bonferroni correction[10]. The smallest P-value is 0.14 between treatment 2 and control 1. Values indicate that significant differences would likely be observed for an experiment with a larger number of crowd-workers.

4.2.3 Experiment 3: Control Questions

Control questions are often used in crowdsourcing systems to check whether crowd-workers are (i) qualified for a certain task and (ii) constantly motivated to perform at a reasonable level throughout the task, i.e. not assigning random answers or the same answers to all the questions[11]. In CrowdFlower, conditions (i) and (ii) are implemented as respectively a quiz that workers must pass in order to embark on a task and the inclusion of one control question at a random position in each task page.

In experiment 3, we investigated what effect introducing such control questions would have on crowd-workers' performance of judging relevance on CrowdFlower and whether the real-time feedback and the competition functionality provided by the leaderboard were still useful for motivating the workers to annotate more accurately given the certain level of quality control provided by the control questions.

More specifically, for each group in experiment 3, we collected 50 responses for each of the 100 task questions, which were a different set from those used in experiments 1 and 2. The task was organised to comprise one quiz page which contained 5 control questions and 20 task pages each of which contained one control question randomly inserted by CrowdFlower and four target questions (in random order). As a result and differently from the previous experiments, the 5 candidate documents to be judged on each task page corresponded to different queries. There were in total 25 control questions (5 for the quiz and 20 for the task pages to filter low-performing workers) and

[10] Were it not for the correction for multiple comparisons, significant differences would have been claimed. Moreover, one-way ANOVA rejects the null that the group means are equal with a P-value of 0.019.

[11] This behaviour was observed in experiment 1 for two workers in the control group.

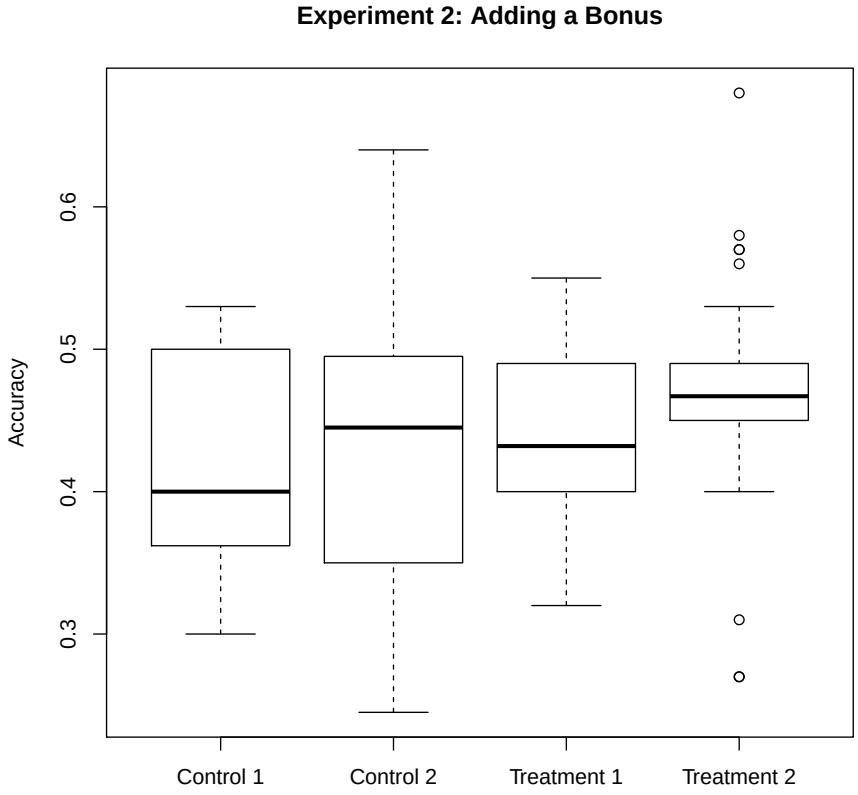


Figure 4.6: A boxplot of Accuracy across the different groups in Experiment 2 where a $1 bonus was paid to the top 10 contributors in Control group 2 and Treatment groups 1 and 2 at the end of the task.

80 target questions in this experiment. Thus a qualified crowd-worker could judge a maximum of 105 documents (5 in the quiz and 100 in the task). Moreover, we set the minimum accuracy on the control questions to be 0.6, higher than the average Macro Accuracy (around 0.45) achieved by the groups in the previous experiments to allow the control questions to affect crowd-workers' annotation processes, while not so high that many workers are removed from the task early on.

Only two groups were investigated in this experiment:

- **Control Group:** *control questions* were used to evaluate crowd-workers based on their corresponding accuracy to see if any of them should be removed from the task. The control accuracy in turn was provided back to the workers by CrowdFlower. No other performance information was provided, but workers were informed that the *top 10* would be paid a $1 bonus at the completion of the task.
- **Treatment 1:** The configuration was the same as for the control group except that a *leaderboard* was provided to the workers showing their rankings in the group in terms of the

Table 4.3: Results for Experiment 3, where control questions were used (in both the control and the treatment groups) to vet and remove crowd-workers based on their corresponding accuracy.

Group	# Workers	# Judgements	Accuracy Micro	Accuracy Macro
Control	40	3495	52.85%	52.53%
Treatment	45	4070	54.84%	54.82%

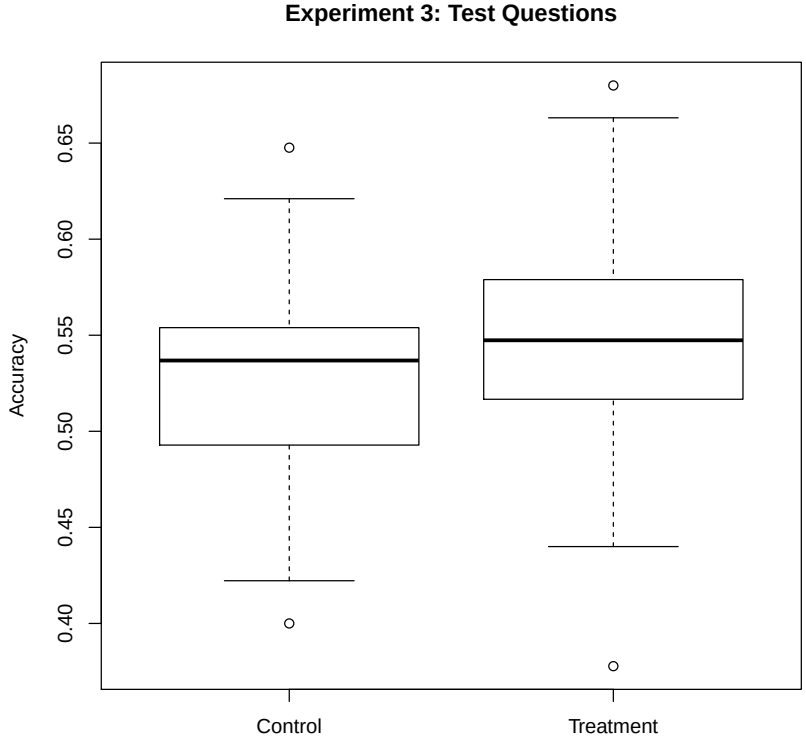


Figure 4.7: A boxplot of Accuracy across the different groups in Experiment 3 where control questions were used to guarantee a certain level of relevance judging accuracy.

estimated number of correct responses across all the documents judged (i.e. both control and target questions).

The results of the third experiment are shown in table 4.3. We note that the overall accuracy across both the control and treatment groups is much higher than it was for the previous experiments, as was to be expected, given that the poor performing crowd-workers (as judged by their accuracy on the control questions) were prevented from further relevance judging and their previous judgements removed from the analysis[12].

[12]Assuming they had not managed to judge more than 50 documents prior to being disqualified

The spread of worker performance shown in Figure 4.7 is similar for the control and treatment groups, with the treatment group showing higher median (and mean) performance across crowd-workers, which is consistent with the hypothesis that providing a leaderboard motivates improved judging performance. However, the difference between the means was not found to be statistically significant at the 0.05 level with P-value of 0.071, indicating that a larger experiment was required. We therefore repeated the experiment three weeks later[13], doubling the sample size of crowd-workers involved. For the repeated experiment, the average accuracy was 52.9% for the 105 crowd-workers in the treatment group versus 51.2% for the 92 workers from the control group[14], and a significant difference in crowd-worker accuracy across the groups was observed with P-value of 0.039.

4.3 Conclusion

In this chapter, we investigated the quality control efficacy of incorporating competitive game elements into crowdsourcing task designs, namely *real-time performance feedback* and a *leaderboard*, both with and without the use of *bonus payments* and *control question filtering*. The main findings of the investigation are able to answer the four sub-questions: $RQ1.1$ to $RQ1.4$, and these answers constitute the overall answer to our research question $RQ1$:

- **Answer to *RQ*1.1**: Solely providing real-time performance feedback to crowd-workers appears to have only a small effect on their response accuracy (compared to providing no feedback to them).

- **Answer to *RQ*1.2**: Additionally providing an all-time task leaderboard to crowd-workers has little effect on their response accuracy (compared to providing no feedback to them).

- **Answer to *RQ*1.3**: Combining the two competition elements with a bonus incentive to the best-performing workers on the leaderboard appears necessary to motivate all the workers to achieve higher accuracy (i.e. average 3% in the relevance judgement task). We note that this finding was not significant at the 0.05 level. Nevertheless, it should be emphasized that the controlled manner in which we performed the tests, and the consistency of the improvement

[13] We waited to repeat the experiment in order to reduce the chance that any crowd-workers from the previous experiment would participate in the repeated experiment.

[14] The fact that the macro-accuracy values (for both control and treatment groups) were lower for the repeated experiment than in the original experiment may be due to the fact that the same number (10) of $1 bonuses were offered twice the number crowd-workers.

across experiments suggests the fact that larger repeated experiments would likely observe statistically significant results for the above case without control questions.

- **Answer to *RQ*1.4**: Including control questions into the above configuration (to eliminate low-performing workers) improves the overall response accuracy of workers, and does not adversely affect the usefulness of the leaderboard. This finding was statistically significant at the 0.05 level for the relevance judgement task.

- **Overall answer to *RQ*1**: Gamifying paid crowdsourcing tasks to make them more competitive with scoring-based real-time performance feedback, a task leaderboard and bonus incentives can improve the quality of worker responses. More specifically, combining the feedback with the task leaderboard by showing them on the same page to workers while offering the best-performing workers fixed bonuses can motivate the workers to respond more accurately than they would have otherwise. This holds true even in the presence of pre-task quizzes and in-task worker filtering using control questions.

There are a number of interesting directions for future work here:

- It would be insightful to repeat experiments for the real-life use-case where the ground truth relevance judgements are not known, and the accuracy of crowd-workers must be estimated based on the relevance judgements collected across workers up to that point. In this case, EM-based algorithms such as Dawid-Skene [56] can be used to estimate quality of each crowd-worker (as well as the posterior over relevance for each document). The estimated accuracy values will likely exhibit greater variability and it would be interesting to see whether that increased variability has an effect on crowd-workers actual accuracy.

- It would be interesting to investigate other performance measures such as the amount of time taken to provide relevance judgements and the number of relevance judgements per worker to see whether real-time feedback and competition causes crowd-workers to annotate more items. One could also deepen the analysis to investigate whether the accuracy of workers improves over time and what effect competition has on workers' performance over time.

- Finally, there is an enticing opportunity to provide real-time rewards to crowd-workers based on their movements on the leaderboard. It is possible to link the amount of bonus to the degree and the direction of movement of the crowd-worker, leading to in some sense "economically optimal" (from a decision theoretic point of view) crowdsourcing approaches.

Chapter 5

Leveraging Side Information for Improved Quality Control on Sparse Responses

In this chapter, we answer the second research question *RQ*2 with all its sub-questions (i.e. *RQ*2.1 to *RQ*2.4). More specifically, we investigate the possibility of leveraging *side information* regarding different aspects of crowdsourcing for improving the quality control of crowdsourced responses when they are sparse (i.e. few responses are collected per worker and per question).

Crowdsourcing facilitates large-scale online data labelling (in the form of question answering) and collection in an inexpensive and timely manner. However, its effectiveness is limited by workers with various motivations and abilities, who end up producing conflicting labels for the same data items. Moreover, an ever-growing number of unlabelled items versus limited budgets in most crowdsourcing projects often results in a small number of labels assigned to an item. Aggregating such small numbers of conflicting labels using majority vote to infer the true labels for the data items is often unreliable.

To overcome the above problem, the quality of responses from crowd-workers must be controlled in a principled manner such that the influence of “high-quality" responses can outweigh that of “low-quality" ones when aggregated for the true labels of the data items. This activity is known from the previous chapters as quality control for crowdsourcing (QCC) [185]. The statistical models for the QCC purpose consider the *abilities/expertise* of workers to govern the quality of the answers they produce with greater abilities indicating higher quality [56, 83]. Some of the models

also consider the *difficulty* of the questions which counteracts the worker's ability to undermine the quality of their responses [87, 88]. These models have overall achieved superior performance over the conventional majority vote, and the basic pre-task or in-task worker filtering using control questions.

There exists, however, one major pitfall of the current QCC models which is their vulnerability to the *response sparsity* problem. This problem happens frequently in real-world crowdsourcing scenarios where only a few labels are collected for each data item or from each worker [102]. Consequently, the procedure for estimating the parameters of the models (for each question and worker) is bound to become unreliable, causing the model's QCC performance to deteriorate.

As an example, due to the paucity of her provided answers, an expert worker could be erroneously considered inaccurate by the QCC models if most of her answers happen to disagree with the majority. Meanwhile, a novice could be erroneously considered accurate if she happens to have made some lucky guesses. In this case, extra side information about the demographics of these workers, the questions they have answered, the time they have taken to respond, or even their situated environments could possibly help to improve the estimation of their associated parameters in the models. Following the previous example, if we know that the demographics of the expert (e.g. her education) are very similar to those of the other experts who have been answering correctly (e.g. by agreeing with the majority), then the belief that she is being a novice due to her poor performance thus far can be weakened by her similarity with the other experts. As a result, her next answer will be more trusted by the models.

In most crowdsourcing tasks, there is extra side information already available (e.g. payment), or which can be collected with some minor efforts (e.g. by designing simple surveys to collect demographic information or programming scripts to collect them). The side information can be elicited from the crowd-workers, the questions and the contexts in which workers are situated while answering questions. Relevant work in crowdsourcing thus far has focused on exploiting only very specific types of side information for improving the QCC performance. To the best of our knowledge, no study has developed a *unified scalable framework* able to *integrate and utilize arbitrary types of side information* for better controlling the quality of worker responses, thereby improving the true answer prediction.

5.1 Related Work

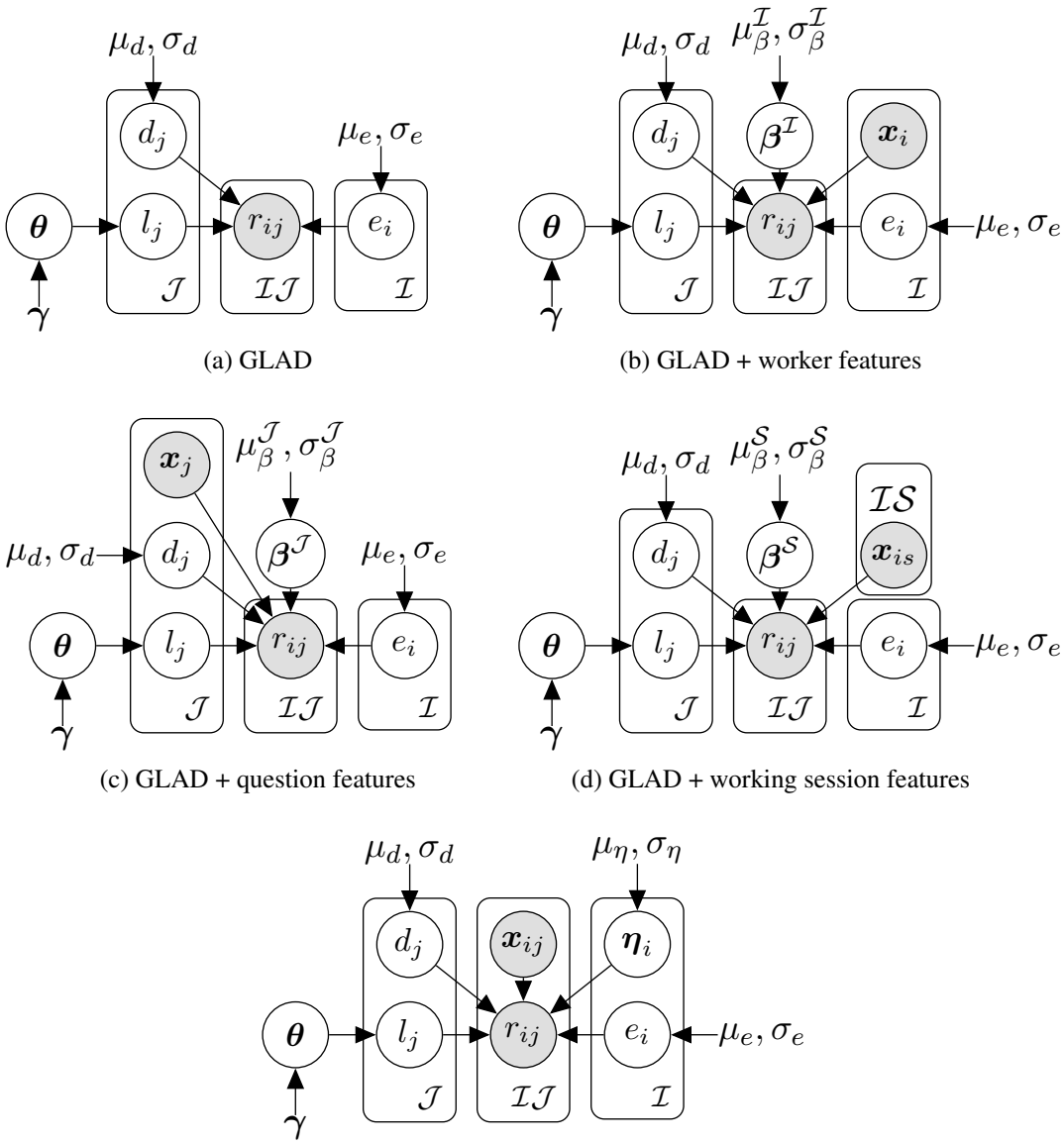
As noted in Chapter 2, research that considers differences in worker abilities/expertise and infers the correct answers for questions dates back to the work of Dawid and Skene [56]. The authors in this work encode each worker's ability in the form of response biases into a single confusion matrix. This matrix accommodates the worker's conditional probabilities of all possible responses given each possible correct answers. Since then, many models have been proposed to prevent the learning of the confusion matrix for each worker from over-fitting the corresponding responses. They have done this by either simplifying the confusion matrix setting by making it symmetric [70, 83, 87, 95], or by grouping workers with similar confusion matrices together to smooth the worker-specific confusion matrices with the ones learned at the group-level [73, 74]. Sometimes modelling worker abilities alone is not enough for accurate estimation of response quality. Accordingly, there has been a large amount of research [87, 88, 89, 94, 99] that takes into account another set of model parameters known as *question difficulty*. Some research further considered the *multi-dimensional* interactions between the worker expertise and the question difficulty [25, 92].

Among all the prior work, research investigating the use of side information to improve the QCC performance is limited. Kamar et al. [10] studied utilizing observed side information, in particular features of the questions, to estimate question-side confusion matrices to account for task-dependent biases. Kajino et al. [93] developed convex optimization techniques using worker-specific classifiers centered on a base classifier which takes in question features for inferring their correct answers. Ruvolo et al. [92] built a multi-dimensional logit model for predicting the correct answer probability based on observed worker features. Ma et al. [94] took into account "bag-of-word" information for learning the topical expertise of individual workers. Venanzi et al. [31] considered response delay information to better distinguish spammers from genuine workers. We can see that each of these relevant works has focused on exploiting only one specific type of side information for improving the performance of the quality control of worker responses.

5.2 Proposed Framework

We are interested in developing a unified framework that is able to predict the correct answers for questions based on both worker responses and various types of side information about the workers, the questions and the contexts. The framework should be scalable for incorporating new types of side information. Table 1.1 has summarized the notation to be used in this chapter including

the observed features that encode the different types of side information, model parameters and hyper-parameters of the proposed framework.



(a) GLAD

(b) GLAD + worker features

(c) GLAD + question features

(d) GLAD + working session features

(e) GLAD + response features

Figure 5.1: Models we have developed for extending the *GLAD model* (a) with *worker features* (b), *question features* (c), *worker session features* (d), and *response features* (e).

5.2.1 Basic Framework

Our proposed framework extends the GLAD model [87] which is shown in Figure 5.1a. GLAD applies a *logistic* function to the product between worker expertise and question difficulty variables to calculate the probability of the answer being correct. More precisely, GLAD defines the

probability of a response r_{ij} being correct, (i.e. equal to the correct answer l_j), as follows:

$$\begin{aligned} p(r_{ij} = l_j) = f_{ij} &= \frac{1}{1 + \exp(-z_{ij})} \\ z_{ij} = f_i / f_j, \quad f_i &= e_i, \quad f_j = \exp(d_j) \end{aligned} \tag{5.1}$$

The functions f_{ij}, f_i, and f_j are specific to response r_{ij}, worker i and question j respectively. According to equation 5.1, they can be interpreted as the response quality function, the expertise function and the difficulty function respectively. According to GLAD, $e_i \in \mathbb{R}$ models the expertise of worker i, and $\exp(d_j)$ with $d_j \in \mathbb{R}$ models the difficulty of question j. Moreover, GLAD assumes Normal priors over e_i and d_j with μ_d, σ_j^2 and μ_e, σ_e^2 being their prior means and variances respectively. The authors further simplified the GLAD model so that instead of needing to infer a $|\mathcal{K}| \times |\mathcal{K}|$ matrix of parameters for each worker as done by Dawid and Skene [56], the model makes use of the following conditional assumption:

$$p(r_{ij}|l_j) \;=\; \begin{cases} f_{ij}, & r_{ij} \;=\; l_j \\ \frac{1-f_{ij}}{|\mathcal{K}|-1}, & \text{Otherwise} \end{cases} \tag{5.2}$$

We choose GLAD as the basis of our framework because of its simple factorization form which easily allows for linear combination of different factors encoding arbitrary information about workers, questions and their dyadic relations.

5.2.2 Incorporating Worker Information

We start by encoding side information about workers into the basic framework as shown in Figure 5.1b. In this case, the worker-expertise function f_i in equation 5.1 is modified to become:

$$f_i = e_i + \boldsymbol{x}_i^T \boldsymbol{\beta}^{\mathcal{I}} \tag{5.3}$$

Here the dot product between the multi-dimensional feature vector $\boldsymbol{x}_i$ of worker i and the weight vector $\boldsymbol{\beta}^{\mathcal{I}}$ forms a global regression across all the workers with $\boldsymbol{\beta}^{\mathcal{I}}$ learned to bring the expertise offsets of similar workers closer together. This helps to smooth the irregular expertise estimates that result from the sparse responses across workers. Moreover, we assume a Normal prior over each component of $\boldsymbol{\beta}^{\mathcal{I}}$ with mean $\mu_\beta^{\mathcal{I}}$ and standard deviation $\sigma_\beta^{\mathcal{I}}$.

Note that when responses per worker are abundant, directly inferring worker expertise from them is sufficient without needing much help from the side information. This trade-off between the response data and the side information has already been achieved in our framework without needing extra weights on e_i and $\boldsymbol{x}_i^T \boldsymbol{\beta}^{\mathcal{I}}$. To see this, consider factorizing out a variable $\alpha \in \mathbb{R}$ from the vector $\boldsymbol{\beta}^{\mathcal{I}}$ to represent the weight for $\boldsymbol{x}_i^T \boldsymbol{\beta}^{\mathcal{I}}$ in equation 5.2.2 while the weight for e_i is 1:

$$f_i = 1 \times e_i + \alpha \times \boldsymbol{x}_i^T \boldsymbol{\beta}^{\mathcal{I}'} \text{ and } \boldsymbol{\beta}^{\mathcal{I}} = \alpha \times \boldsymbol{\beta}^{\mathcal{I}'} \tag{5.4}$$

Thus, learning $\boldsymbol{\beta}^{\mathcal{I}}$ entails learning the weight α denoting the relative importance of the side information. In other words, using equation 5.2.2 avoids over-parametrizing our framework.

5.2.3 Incorporating Question Information

The question information is incorporated into the basic framework as shown in Figure 5.1c. The question-difficulty function f_j in equation 5.1 now has the following form:

$$f_j = \exp(d_j + \boldsymbol{x}_j^T \boldsymbol{\beta}^{\mathcal{J}}) \tag{5.5}$$

where the dot product between the multi-dimensional feature vector $\boldsymbol{x}_j$ of question j and weight vector $\boldsymbol{\beta}^{\mathcal{J}}$ forms a global regression over all the questions. The vector $\boldsymbol{\beta}^{\mathcal{J}}$ serves the same purpose as its worker-side counterpart, that is to smooth the unreliable estimate of d_j by adjusting it with a scaler $\boldsymbol{x}_j^T \boldsymbol{\beta}^{\mathcal{J}}$. We again assume a Normal prior over each component of $\boldsymbol{\beta}^{\mathcal{J}}$ with mean $\mu_\beta^{\mathcal{J}}$ and standard deviation $\sigma_\beta^{\mathcal{J}}$.

5.2.4 Incorporating Session-Level and Response-Level Contextual Information

We consider contextual information at both the response level and the session level. The former type of contextual information is specific to each response given by a worker to a question, encoded by features including response delay and order within the task. The latter is specific to each working session of a worker which we define to start when a task page is loaded and to end once the page is submitted with no time-out in between. In this case, session features can include working devices (e.g. laptop, tablet, mobile, etc.), rendering browsers, the work time periods (e.g. the day of the week, the hour of the day), locations and etc.

The session-level features should be more varied across workers than they are for each worker. This is because most workers usually remain situated in the same environment for the duration

of a task (e.g. using the same device, browser and working in the same location). In contrast, the response-level features should provide much more insight into the variation within each worker's answering behaviour.

To leverage the advantages of both types of contextual information, we incorporate them into the basic framework as shown in Figure 5.1d and Figure 5.1e. The symbol $\mathcal{S}$ denotes the set of working sessions each corresponding to a task page containing a number of questions. The corresponding changes made to the worker-expertise function f_i by the worker session features $\boldsymbol{x}_{is}$ and to the response quality function f_{ij} by the response features $\boldsymbol{x}_{ij}$ are respectively the following:

$$f_i = e_i + \boldsymbol{x}_{is}^T \boldsymbol{\beta}^{\mathcal{S}} \tag{5.6}$$

$$z_{ij} = f_i / f_j + \boldsymbol{x}_{ij}^T \boldsymbol{\eta}_i \tag{5.7}$$

The dot product between the multi-dimensional feature vector $\boldsymbol{x}_{is}$ of worker i within the current session s and the weight vector $\boldsymbol{\beta}^{\mathcal{S}}$ forms a global regression over all the sessions of all the workers. The result affects the correctness probabilities $p(r_{ij} = l_j)$ of any response given by worker i to any question within her current session s. Note that equation 5.7 is for incorporating only the session features, while equation 5.2.2 is for incorporating only the worker features. They are different instantiations of function f_i with only one type of side information.

The vector $\boldsymbol{\eta}_i$ is specific to worker i, serving as the weight vector of a local linear regression over the feature vectors of all the responses made by worker i. Such local regressions aim at addressing worker-specific biases that the GLAD model fails to handle properly [25]. We again assume Normal priors over each component of both $\boldsymbol{\beta}^{\mathcal{S}}$ and $\boldsymbol{\eta}_i$ with means $\mu_\beta^{\mathcal{S}}, \mu_\eta$ and standard deviations $\sigma_\beta^{\mathcal{S}}, \sigma_\eta$, respectively. When responses per worker are scarce, the procedure for inferring $\boldsymbol{\eta}_i$ needs to be more conservative. We can realize this by either having a smaller prior variance σ_η or sharing a coefficient vector $\boldsymbol{\eta}$ across workers.

5.3 Parameter Estimation

In this section, we describe the stochastic parameter estimation procedure used to obtain posterior probabilities for the set of latent true answer variables $\mathcal{L} = \{l_j | j \in \mathcal{J}\}$. More specifically, in each iteration of the estimation procedure, we alternate between a collapsed Gibbs sampling [186] for estimating the posterior over the true answers L given the current estimates of the other model

parameters, and a *one-step (first-order)*[1] gradient descent for updating these model parameters given the current assignment to $\mathcal{L}$.

5.3.1 Collapsed Gibbs Sampling for Estimating True Answer Posterior

At this stage, we employ a collapsed Gibbs sampler to obtain posterior samples for $\mathcal{L}$ given the current estimates of the model parameters $\{e_i\}_{i\in\mathcal{I}}$, $\{d_j\}_{j\in\mathcal{J}}$, $\{\boldsymbol{\eta}_i\}_{i\in\mathcal{I}}$ and $\{\boldsymbol{\beta}^{\mathcal{I}}, \boldsymbol{\beta}^{\mathcal{J}}, \boldsymbol{\beta}^{\mathcal{S}}\}$. In this case, the conditional probabilities of correct answer l_j is obtained by marginalizing out the multinomial probability vector $\boldsymbol{\theta}$, which ends up being:

$$P(l_j = k|\mathcal{L}_{\setminus j}, \mathcal{R}_j, \{f_{ij}\}_{i\in\mathcal{I}_j}, \boldsymbol{\gamma}) \propto \frac{N_{\setminus jk} + \gamma_k}{\sum_{k'\in\mathcal{K}}(N_{\setminus jk'} + \gamma_{k'})} \times \prod_{i\in\mathcal{I}_j}\left((f_{ij})^{\mathbb{1}\{r_{ij}=l_j\}}\left(\frac{1-f_{ij}}{|\mathcal{K}|-1}\right)^{\mathbb{1}\{r_{ij}\neq l_j\}}\right) \quad (5.8)$$

where $\mathcal{I}_j$ is the set of workers who responded question v with a set of responses $\mathcal{R}_j$, $\mathcal{L}_{\setminus j}$ is the set of current correct answer assignments to all the questions excluding question j, and $N_{\setminus jk}$ is the number of questions excluding j whose correct answers are currently inferred to be k.

5.3.2 Gradient Descent for Estimating Other Model Parameters

The respective conditional probability distributions of the model parameters are hard to compute analytically due to the presence of the logistic function. Instead, we run gradient descent for one step with respect to each model parameter on the negative logarithm of their joint conditional probability distribution. More specifically, suppose we use $\mathcal{H}$ to denote the set of model paramters excluding $\mathcal{L}$, that is $\mathcal{H} = \{\{e_i, \boldsymbol{\eta}_i\}_{i\in\mathcal{I}}, \{d_j\}_{j\in\mathcal{J}}, \boldsymbol{\beta}^{\mathcal{I}}, \boldsymbol{\beta}^{\mathcal{J}}, \boldsymbol{\beta}^{\mathcal{S}}\}$, and use $\mathcal{H}_0$ to denote their respective hyper-parameters. We have the negative logarithm of the joint distribution of the model parameters conditioned on the current true answer estimates $\hat{\mathcal{L}}$ as follows:

$$\mathcal{Q} = -\log P(\mathcal{H}|\mathcal{L}) = -\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}_j}\log\left((f_{ij})^{\mathbb{1}\{r_{ij}=l_j\}} \times \left(\frac{1-f_{ij}}{|\mathcal{K}|-1}\right)^{\mathbb{1}\{r_{ij}\neq l_j\}}\right) - \log p(\mathcal{H}|\mathcal{H}_0). \quad (5.9)$$

The first term of equation 5.9 is the negative log-likelihood of the response data, and the second term is the negative log-prior with $\mathcal{H}_0$ being the set of hyper-parameters for $\mathcal{H}$. To minimize $\mathcal{Q}$, we take partial derivative of $\mathcal{Q}$ with respect to each element in $\mathcal{H}$.

[1] We conducted a series of experiments for GLAD on simulated data. The experiments compared Gibbs sampling with full convergence gradient descent against it with one-step gradient descent for MAP estimation. Both approaches converged to almost the same prediction accuracy (using the parameter and hyper-parameter settings specified in the original GLAD paper [87]) with the one-step approach being significantly faster. Thus, we chose to use the one-step gradient descent for all experiments.

5.3.3 Estimating Worker Expertise and Question Difficulty e_i and d_j

When estimating worker expertise e_i and question difficulty d_j, the relevant log-prior terms in $\log p(\mathcal{H}|\mathcal{H}_0)$ for e_i and d_j are $\frac{(d_j-\mu_d)^2}{2\sigma_d^2}$ and $\frac{(e_i-\mu_e)^2}{2\sigma_e^2}$ respectively. The gradients of e_i and d_j are thus:

$$\frac{\partial \mathcal{Q}}{\partial d_j} = -\sum_{i\in\mathcal{I}_j}\left(\delta_{ij} f_i f_j\right) + \frac{d_j-\mu_d}{\sigma_d^2} \tag{5.10}$$

$$\frac{\partial \mathcal{Q}}{\partial e_i} = -\sum_{j\in\mathcal{J}_i}\left(\delta_{ij} f_j\right) + \frac{e_i-\mu_e}{\sigma_e^2}, \tag{5.11}$$

where $\delta_{ij} = [\mathbb{1}\{r_{ij}=l_j\}(1-f_{ij}) - \mathbb{1}\{r_{ij}\neq l_j\}f_{ij}]$, and $\mathcal{J}_i$ is the set of questions responded by worker i.

5.3.4 Estimating Task-, Worker- and Session-level Regression Coefficients

For simplicity of presenting the gradient derivation, we use the same symbol m to denote each component of the task-, worker- and session-level regression coefficient vectors $\boldsymbol{\beta}^{\mathcal{I}}$, $\boldsymbol{\beta}^{\mathcal{J}}$ and $\boldsymbol{\beta}^{\mathcal{S}}$. In this case, the relevant log-prior terms in $\log p(\mathcal{H}|\mathcal{H}_0)$ for the m-th element of the three vectors are $\frac{\beta_m^{\mathcal{I}}-\mu^{\mathcal{I}}}{\sigma^{\mathcal{I}2}}$, $\frac{\beta_m^{\mathcal{J}}-\mu^{\mathcal{J}}}{\sigma^{\mathcal{J}2}}$ and $\frac{\beta_m^{\mathcal{S}}-\mu^{\mathcal{S}}}{\sigma^{\mathcal{S}2}}$ respectively. Taking derivatives with respect to the m-th element of the three vectors yields the following gradients:

$$\frac{\partial \mathcal{Q}}{\partial \beta_m^{\mathcal{I}}} = -\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}_j}\left(\delta_{ij} f_j x_{mi}\right) + \frac{\beta_m^{\mathcal{I}}-\mu^{\mathcal{I}}}{\sigma^{\mathcal{I}2}} \tag{5.12}$$

$$\frac{\partial \mathcal{Q}}{\partial \beta_m^{\mathcal{J}}} = -\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}_j}\left(\delta_{ij} f_i f_j x_{mj}\right) + \frac{\beta_m^{\mathcal{J}}-\mu^{\mathcal{J}}}{\sigma^{\mathcal{J}2}} \tag{5.13}$$

$$\frac{\partial \mathcal{Q}}{\partial \beta_m^{\mathcal{S}}} = -\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}_j}\left(\delta_{ij} f_j x_{mis_j}\right) + \frac{\beta_m^{\mathcal{S}}-\mu^{\mathcal{S}}}{\sigma^{\mathcal{S}2}} \tag{5.14}$$

In equation 5.14, the symbol $s_j \in \mathcal{S}$ corresponds to the session of worker i in which she encountered question j.

5.3.5 Estimating Response-level Regression Coefficients

The log-prior term for the m-th element of the response-level regression coefficient vector $\boldsymbol{\eta}_i$ is $\frac{\eta_{mi}-\mu_\eta}{\sigma_\eta^2}$. The gradient with respect to η_{mi} is thus given by:

$$\frac{\partial \mathcal{Q}}{\partial \eta_{mi}} = -\sum_{j\in\mathcal{J}_i}\left(\delta_{ij} x_{ij}\right) + \frac{\eta_{mi}-\mu_\eta}{\sigma_\eta^2} \tag{5.15}$$

We applied the Gibbs sampling plus one-step gradient descent to our experiments with the above gradients and the step size set according to Section 5.4.4. We ran the algorithm until convergence, which was determined automatically by the stopping threshold $|Q_t - Q_{t-1}|/|Q_{t-1}| < 10^{-5}$.

5.4 Experiments

We present experiments that study the performance of our framework for combining different types of side information to improve the quality control for crowdsourcing on three real-world datasets.

5.4.1 Datasets

The three datasets were collected in separate crowdsourcing tasks on CrowdFlower with three responses collected for each question. As a basic quality control measure, we filtered out workers who did not achieve 88% accuracy on pre-defined control questions. The qualified workers were also asked for additional information including demographics and personal traits[2]. Each qualified worker is allowed to label a certain number of questions for each task and is free to quit labeling at any time. Nineteen questions were randomly selected and shown to the a worker on each task page. Table 5.1 provides a summary of the three datasets.

Dataset	$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{R}\|$
Stack Overflow	505	14,021	42,063
Evergreen Webpage	434	7,336	22,008
TREC 2011	160	1826	5,478

Table 5.1: Dataset Summary. The headers correspond to the notation introduced in Table 1.1.

TREC 2011 crowdsourcing Track: Each CrowdFlower worker was asked to judge the relevance levels (i.e. *highly relevant*, *relevant* and *non-relevant*) of 38 Web-pages to their corresponding queries in the TREC 2011 crowdsourcing Track dataset[3]. Figure 5.2 shows a question for collecting the relevance judgements for a pair of query (i.e. “french lick resort and casino") and document. The ground truth for each question had been judged by multiple experts from the National Institute of Standards and Technology[4] (NIST).

[2]This was done by mixing the demographic survey questions with the control questions in the quiz for the workers to answer prior to starting the crowdsourcing task. As a result, there exist a small number of missing values in the demographic data collected as not all the survey questions were chosen by CrowdFlower to appear on the quiz page.

[3]https://sites.google.com/site/treccrowd/2011

[4]https://www.nist.gov/

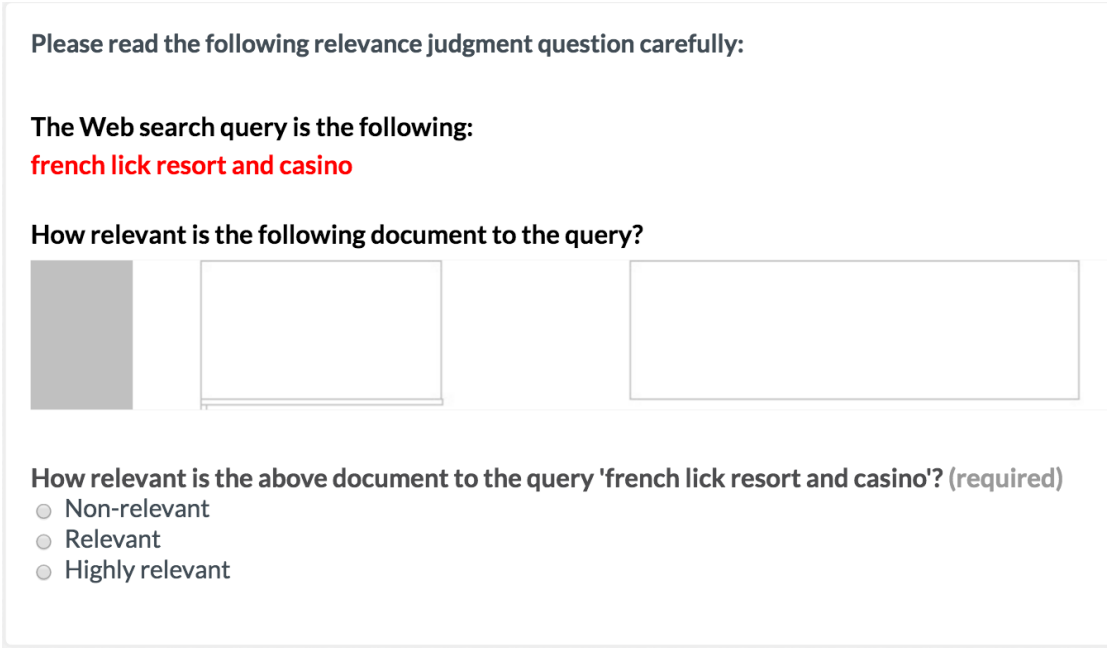


Figure 5.2: a question for relevance judgement

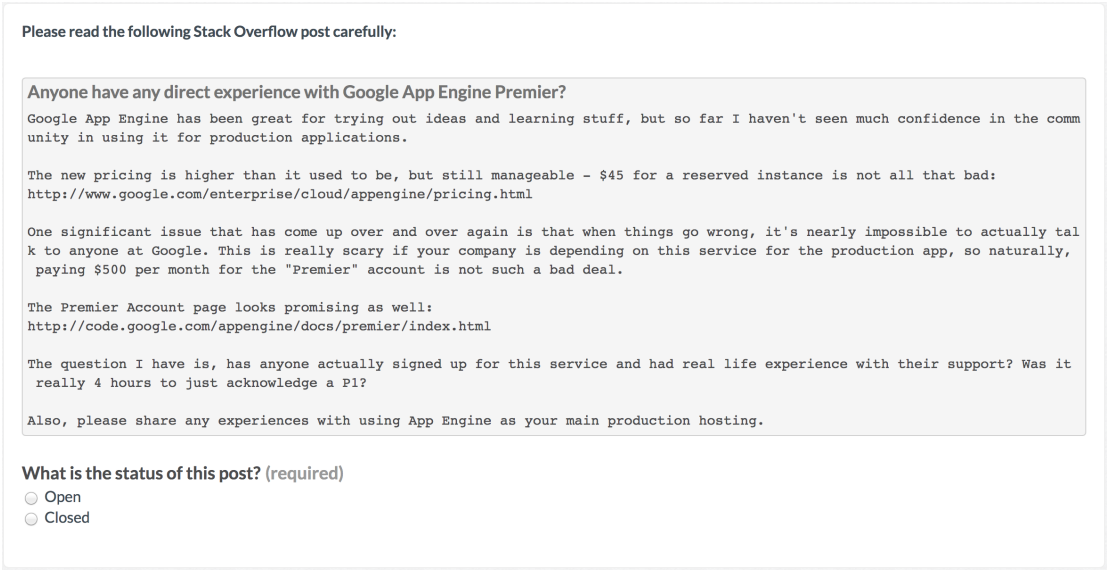


Figure 5.3: a question for Stack Overflow post status judgement

Stack Overflow Post Status Judgement: Each CrowdFlower worker was asked to judge the status of 95 archived questions from Stack Overflow[5]. The status of a question can be either *open*, meaning it is regarded suitable to stay active (i.e. visible, answerable and editable) on Stack Overflow, or *closed*, meaning the opposite for reasons including that it is *not a real question* (i.e. questions that are ambiguous, too broad or "show no efforts" in seeking answers), *not constructive* to the Website (i.e. questions that are subjective and have no correct answers) and *too localized*

[5]The set of questions judged is a random subset of the training dataset used in the Kaggle competition "*Predict Closed Questions on Stack Overflow*" (https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow).

(i.e. questions that are not reproducible, thus useless to other workers in the future). Figure 5.3 shows a question for collecting the status judgements for a Stack Overflow post. The ground truth of each question had been judged by multiple "moderators". Moderators are StackOverflow users with the highest reputations and privileges and are considered as experts in their respective fields of programming.

Evergreen Webpage Judgement: Each CrowdFlower worker was asked to judge 57 Web-pages[6] on whether they think these Web-pages have a timeless quality or, in other words, will still be considered by average workers as valuable or relevant in the future. The ground truth of each question had been judged by hundreds of users of the StumbleUpon Web content recommendation engine[7].

Side Info.	Dataset		
	Stack-Overflow	**Evergreen**	**TREC**
Worker	1. *age*, 2. *gender*, 3. *education*, 4. *personality traits*		
	self-appraisal about: 1. *programming experience* 2. *frequency of search/post/ edit/answer -ing questions* 3. *diversity of questions dealt with*	1. *mother tongue* **self-appraisal about:** 2. *frequency of online search* 3. *frequency of bookmarking Web-pages* 4. *frequency of revisiting Web-pages*	1. *mother tongue* **self-appraisal about:** 2. *frequency of online search* 3. *diversity of online search topics* 4. *average # search result pages checked*
Question	1. *content length*, 2. *question genre*		
	—	1. *Web-page features*	—
Response	1. *response time/delays*, 2. *response order*		
Session	1. *weekends or weekdays*, 2. *time of the day* 3. *labeling devices (e.g. PC, tablets, etc.)*		

Table 5.2: features encoding different types of side information.

5.4.2 Feature Collection

We collected various types of side-information as summarized in Table 5.2.

Worker Features: The worker features include both the demographic and personality trait features which are common to all the three tasks, and the "self-appraisal" features which vary from task to task. The ages of workers (starting from 18) were discretized into 9 groups with the first 8 groups each having a 5-year gap onward and the last being of age 60 or over. The

[6]We used the training set from the Kaggle Competition "*StumbleUpon Evergreen Classification Challenge*": https://www.kaggle.com/c/stumbleupon.

[7]https://www.stumbleupon.com/

education backgrounds of workers were divided into 5 categories from “Less than high school" to “Master degree or above". To collect information about the personality traits of workers, we directly employed the 10 survey questions used by Kazai et al. [35] based on the so called five personality trait dimensions [187]. As for the “self-appraisal” features, these were designed to capture the possible nuances in workers’ expertise levels on different tasks from their own perspectives.

Question Features: The question features include both the common features which are the content length and question genres (already known for each question of each dataset), and those unique to the Evergreen dataset which are the original Web-page features[8] provided in the Kaggle competition.

Response and Session Features: The contextual features were collected at both the response and the session levels, and were the same across all the tasks. Response-level feature “response delay” records the amount of time each worker took to label each question. Its value was calculated by subtracting the click time of the previous question (or the page load time if it was more recent) from the click time of the particular question. We also computed the “response order” of each question by ordering questions by their “last click time”. As for the session feature “time of the day", we set its value, either “daytime", “night" or ‘late night", corresponding to the periods [6am, 7pm), [7pm, 23pm), and [23pm, 6am), respectively.

Feature Normalization: The pre-processing of the features involved binarizing the non-numeric features, and normalizing the numeric features using a Z-score transformation, (except for the numeric feature “response order" for which we used a min-max normalization as its values were always uniformly distributed). For numeric features with highly skewed empirical distributions, a log-transformation[9] was applied prior to normalizing with the Z-score transformation. Finally, we normalized the response-level feature “response delay” on a per-worker basis (i.e. using worker specific mean and standard deviation values), in order to facilitate local linear regression with the worker-specific weight vector η_i as specified in equation 5.7.

5.4.3 Experiment Setup

Baselines: We verify the efficacy of our model by comparing it with the following three baselines:

- **Majority Vote**: the predicted correct answer for a question is the response given by the majority of the workers.

[8] https://www.kaggle.com/c/stumbleupon/data

[9] $\log(c + x)$ where $c = min(0.1, min(x))$

- **GLAD**[87]: the probability of a correct response is a logistic function over the product between the worker expertise and the question difficulty variables.
- **Community-based DS** [73]: to smooth out unreliable estimates of the confusion matrix entries due to label sparsity, the matrix is drawn (row-wise) from one that is shared by a *community* to which the worker is inferred to belong.

Evaluation Metrics: We use the following metrics to evaluate the performance of the baseline methods and our proposed framework in terms of the question correct answer prediction:

- *Predictive accuracy (Accu)*:
  $\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{1}\{l_i = \hat{l}_i\}$
  where $\hat{l}_i = \arg\max_{l \in \mathcal{K}} P(l_i = l | \mathcal{R}, \textit{Model})$
- *Log-Loss (Log)*:
  $-\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{K}} \mathbb{1}\{l_i = l\} \log \big(P(l_i = l | \mathcal{R}, \textit{Model})\big)$

We use the following metric to evaluate the performance of the baselines and our proposed framework with respect to the unseen worker response prediction.

- *Mean Absolute Error (MAE)*:
  $\frac{1}{|\mathcal{R}_h|} \sum_{r_{ij} \in \mathcal{R}_h} \mathbb{1}(r_{ij} \neq \hat{r}_{ij})$
  where $\hat{r}_{ij} = \arg\max_{l \in \mathcal{K}} P(r_{ij} = l | \mathcal{R}_{\setminus h}, \textit{Model})$

The first measure (Accu) tells us the performance in terms of the number of correctly predicted correct answers. The second measure (Log) tells us how confident the model is in its predictions of the correct answers. The second measure is often more sensitive than the first and can therefore be useful for comparing similarly performing models. Both measures do not rely on held-out testing as our framework is *unsupervised* with respect to the question true answers $\mathcal{L}$. On the other hand, the third measure (MAE) concerns *supervised* learning from *training* responses to predict *unseen held-out test* responses. It informs us of the performance regarding the average number of correctly predicted held-out responses $\mathcal{R}_h$ based on the posterior mode $\hat{r}_{ij}$ given the training responses $\mathcal{R}_{\setminus h}$ and the specific model being tested. In the experiment using the entire responses, the proportion of the held-out test data $\mathcal{R}_h$ is set to be 30% and the accuracy of the unseen worker response prediction of a model is obtained by averaging its performance over 10 such held-out tests.

5.4.4 Hyper-parameter Setup

For GLAD, Whitehill et al. [87] imposes normal priors over both e_i and d_j with means and standard deviations both set to be 1. This setting assumes that *a priori* approximately 75% of workers are reliable and 75% of questions are relatively easy for average workers. Although straightforward, we believe that the above hyper-parameter settings for GLAD is not particularly effective for modeling common crowdsourcing scenarios for two main reasons.

First, crowdsourcing tasks conducted on CrowdFlower can only be attempted, by default, by *leveled* [10] (in other words, experienced and high-quality) workers. We adopted this default setting for all our tasks. Moreover, control questions are very often used to vet and remove low-performing workers before and during the tasks. Thus, we believe that *a priori* no workers should be considered unreliable (that is e_i below or close to 0). Instead, all workers should be expected to have abilities close to the prior mean.

Secondly, it is common to collect only a small number of responses for each question during crowdsourcing. Such few responses can hardly provide sufficient information for GLAD to reliably estimate d_j. Moreover, GLAD applies an exponential transformation on d_j to ensure its non-negativity, which could further "inflate" its inaccurate estimation. As an example, when $d_j = 3$ (i.e. two positive standard deviations from its prior mean), $\exp(d_j) = 20.1$, much larger than $e_i = 3$ which is also at two positive standard deviations. As a result, the log-odds of correct labeling, modeled as their product in GLAD, is likely to be dominated by the potentially inaccurate estimates of d_j. Thus, we suggest a stronger regularization penalty for d_j to suppress these problems. Based on the above analysis, we set the prior means for e_i and d_j to be 2 and 0 respectively. Note that assigning the latter to zero imposes an uninformative setting for the prior of d_j. We set the prior variance for both to be 0.1 in accordance with the arguments above (to limit the variance of e_i and to increase the regularization on d_j).

Each component of the Dirichlet prior vector γ is set to be 1. For each regression weight vector (e.g. $\beta^{\mathcal{I}}$), we set the prior means of its components to be 0, and the prior standard deviations to be (0.01 / #features) so that the influence brought by each global/local regression is comparable to that brought by its affecting factor (e.g. e_i). As for the gradient descent step size, we define a default step size of $\eta = 0.001$ and calculate a parameter-specific size based on the number of data instances available for estimating the parameter, that is $(\eta/\#datapoints)$. Discounting η is

[10] http://crowdflowercommunity.tumblr.com/post/80598014542
/introducing-contributor-performance-levels

necessary as parameters are estimated at different (global or local) levels in our framework. Except for the Dirichlet prior for the class proportions to be fixed all at 1, the other hyper-parameters of the community-based DS, including the number of communities, is set through 10-fold cross-validaton repeated and averaged over 5 iterations, evaluated upon the likelihood of the validation responses.

5.4.5 Prediction with Subsampled Responses

While the default evaluation task is to check the efficacy of our framework under question-side label sparsity (where only 3 responses were collected for each question), we would like to further investigate whether the framework can handle even greater degrees of label sparsity which happens not only on the question side but also on the worker side. To do this, we randomly subsampled a fixed number of responses from each worker. By merging all the subsampled responses from each worker, we obtained a data subset with far fewer questions from each of the three datasets. We varied the number of responses subsampled per worker from 1 to 12 (beyond which we observed very marginal differences in the model performence), and ran all the models for the correct answer prediction as well as the unseen (held-out) worker response prediction at each subsampling point. The unseen response prediction is evaluated on the remaining responses. Both prediction tasks are evaluated using the same metrics before as used in the experiments with the full responses. The whole subsampling procedure was repeated 5 times before we obtained the average predictive accuracy of each model. The hyper-parameter setup in this case remained unchanged as we employed the same hyper-parameter setting for our framework and GLAD, and the 10-fold cross-validation for finding the optimal number of communities.

5.5 Results

The results of the experiment using the entire response data are summarized in Tables 5.3 and 5.4. For clarity, the side information of the workers, the question items, the sessions and the responses is abbreviated to “L", “I", “S" and “R" respectively.

5.5.1 True Answer Prediction

From Table 5.3, our framework with all the features outperforms the three baselines (with the comunity-based DS optimized at 4 communities) on all the datasets. The largest improvement in predictive accuracy is seen over Majority Vote (MV) on the TREC dataset (i.e. by 1%). Marginal

	Stackoverflow		Evergreen		TREC	
	Accu	Log	Accu	Log	Accu	Log
MV	0.6083	0.9748	0.7630	0.6635	0.4720	1.124
4-Community DS	0.6088	0.9292	0.7633	0.6248	0.4818	1.112
GLAD	0.6084	0.9323	0.7631	0.6385	0.4747	1.126
GLAD+I	0.6085	0.9306	0.7632	0.6280	0.4765	1.122
GLAD+L	0.6084	0.9311	0.7631	0.6296	0.4751	1.126
GLAD+R	0.6087	0.9294	0.7633	0.6271	0.4766	1.122
GLAD+S	0.6085	0.9306	0.7631	0.6292	0.4760	1.123
GLAD+I+L+R	0.6088	0.9294	0.7633	0.6256	0.4808	1.116
GLAD+I+L+S	0.6087	0.9301	0.7633	0.6252	0.4804	1.116
GLAD+I+R+S	**0.6090**	0.9289	**0.7634**	0.6245	0.4819	1.112
GLAD+L+R+S	0.6088	0.9298	0.7633	0.6258	0.4802	1.118
GLAD+I+L+R+S	**0.6090**	**0.9288**	**0.7634**	**0.6238**	**0.4820**	**1.108**

Table 5.3: True label/answer predictive accuracy of the models across the three datasets. We denote the side information about the workers, the question items, the sessions and the responses respectively with capital letters “L", “I", “S" and “R".

improvements in accuracy have been observed over the three baselines on the Stackoverflow and the Evergreen datasets. We believe the reason for observing only marginal improvements on these datasets is that all workers have exhibited similar levels of ability for these tasks, producing labels of similar quality across the items.

We note from Table 5.3 that incorporating the observed features about question items appears to produce a larger reduction in the log-loss than that is achieved by adding worker or session features. This is in line with our expectation from the first experiment which is that the item features help to reduce the uncertainty in the question difficulty d_v, which likely has suffered from label sparsity across the three crowdsourcing tasks with only three labels collected for each question . In contrast, reduction in the uncertainty of worker expertise e_u, which is attributed to the addition of the worker and the session features into our framework, appears far less beneficial given that there is already abundant response data available for the estimation.

Moreover, it appears that incorporating response information brings systematic improvements in both accuracy and log-loss. The result confirms that our framework is able to consistently utilize such information to mitigate the bias specific to each worker.

5.5.2 Unseen Held-out Response Prediction

From Table 5.4, we can see that when equipped with all the side information features, our framework again defeats the baseline models GLAD and community-based DS (the majority vote intrinsically

	Stackoverflow MAE	Evergreen MAE	TREC MAE
4-Community DS	0.2306	0.1858	0.5176
GLAD	0.2438	0.1901	0.5216
GLAD+I	0.2347	0.1874	0.5176
GLAD+L	0.2396	0.1898	0.5195
GLAD+R	0.2288	0.1854	0.5173
GLAD+S	0.2356	0.1901	0.5211
GLAD+I+L+R	0.2245	0.1801	0.5162
GLAD+I+L+S	0.2325	0.1854	0.5187
GLAD+I+R+S	0.2276	0.1812	0.5169
GLAD+L+R+S	0.2318	0.1860	0.5184
GLAD+I+L+R+S	**0.2226**	**0.1801**	**0.5160**

Table 5.4: Unseen (held-out) response prediction error of the models across 30% held-out response data from the three datasets.

not suitable for predicting worker responses) with the minimum mean absolute error (highlighted in bold font) for the 30% held-out response data across all the datasets.

5.5.3 True Answer Prediction with Subsampled Responses

The results of the experiment using the randomly subsampled worker response data are summarized in Figures 5.4 and 5.5. We observe from Figure 5.4 that when the number of responses subsampled per worker drops below 6, our framework significantly outperformed GLAD and Majority Vote aross all the three datasets by leveraging only one type of side information. When the number of responses is 12, our framework with combined types of side information still distinctly exceeds the performance of the two baselines over the TREC dataset. The community-based DS model, whose optimal number of community is 2 in this case, is clearly beaten by our framework with (1) a single type of side-information over the Evergreen dataset when the number of subsampled responses is below 4, and (2) the combined types over the TREC and the Stackoverflow datasets when the number is below 3.

We also note that there is larger disparity in performance between different methods when the number of responses varies from 1 to 3 compared to the other sampling counts. We conjecture that in this case, the response sparsity is the severest (due to few labels per worker and per question) and extra side information is most prominent in alleviating the data sparsity. This results in our model significantly outperforming Majority vote and GLAD, which lack mechanisms to fight the response sparsity. Our model outperforms, albeit to a lesser extent, the Community BCC model, which relies on just the response information.

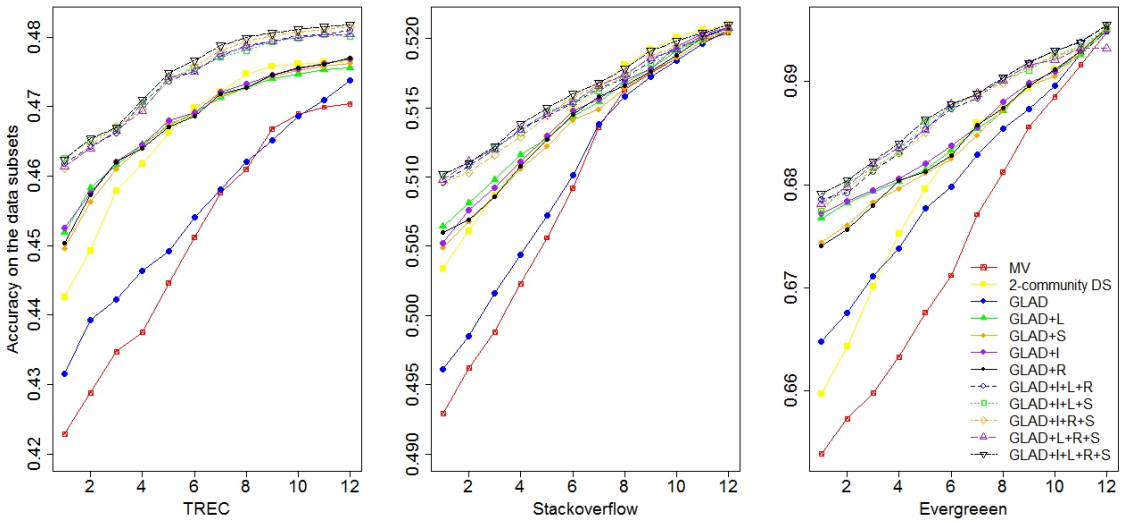


Figure 5.4: Changes of the true label/answer predictive accuracy by varying the number of responses subsampled from each worker across the three datasets.

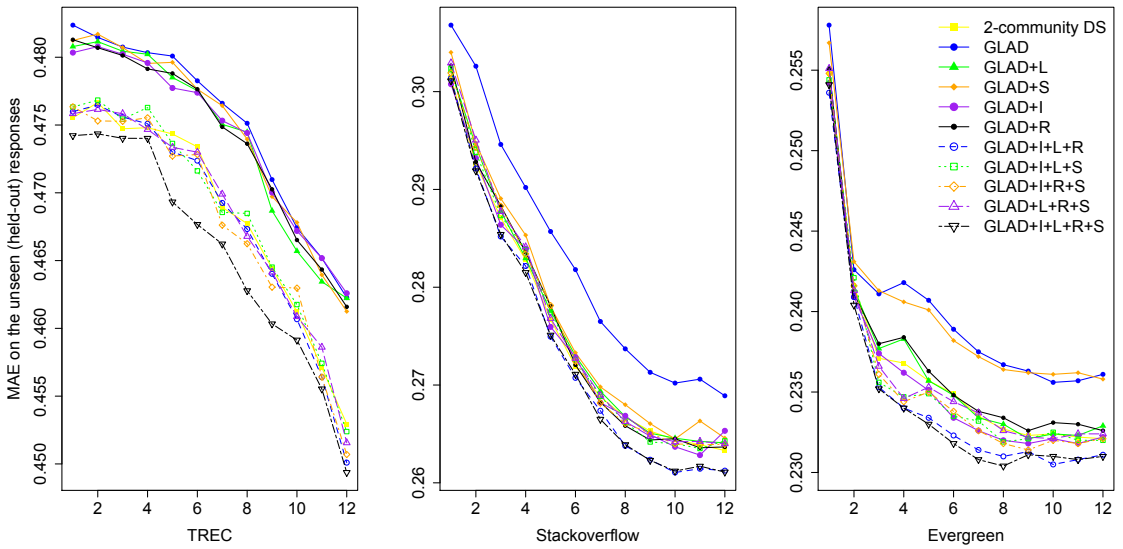


Figure 5.5: Changes of the unseen (held-out) response predictive accuracy by varying the number of responses subsampled from each worker across the three datasets.

5.5.4 Unseen Held-out Response Prediction with Subsampled Responses

When applied to the task of predicting unseen (held-out) responses of workers as shown in Figure 5.5, our framework still holds clear advantages over the GLAD model by leveraging only single types of side information, and over the community-based DS model (with its optimal number of communities being 2) by leveraging multiple (in our experiment at least 3) types of side information.

	Top 5 features ranked according to:	
	Supervised Prediction (P-value)	Unsupervised Prediction (weight)
I	1. num_alphaNumeric_chars 2. num_links 3. domain_business 4. content_length 5. frame_tag_ratio	1. num_links 2. num_alphaNumeric_chars 3. content_length 4. frame_tag_ratio 5. domain_business
L	1. searchOnline_sometimes 2. artist_agreeModerately 3. revisit_topic_diversity_level3 4. education_Bachelor 5. search_topic_diversity_level3	1. searchOnline_veryOften 2. lazy_slightlyAgree 3. revisit_topic_diversity_level3 4. thorough_stronglyAgree 5. age_45_to_49
S	1. use_mobilePhone, 2. use_PC 3. use_tablet, 4. weekends, 5. latenight	1. use_mobilePhone, 2. use_tablet 3. use_PC, 4. weekends, 5. latenight

Table 5.5: Comparison between top 5 most predictive features for supervised and unsupervised (actual) setting.

Overall, our framework makes far more significant improvements over the baselines when the questions are fewer and the response data is more scarce, compared to its performance in the previous experiments based on the entire set of responses.

5.5.5 Statistical Analysis of Feature Importance

To get a clear idea of which features might be important for predicting the accuracy of the responses, we conducted one-way ANOVA where the resulting P-values indicate the significance of the correlation between the features and the correct responses [188, 189]. We compared the P-values with the feature weight estimates from our framework with all the types of side information. We list the Top-5 features with respect to the P-values and the absolute feature weight values inferred from the Evergreen dataset in Table 5.5. We also obtain similar results with Stack-Overflow and TREC datasets. Although our framework works in a fully unsupervised manner whereas one-way ANOVA is supervised, the results show that our framework is equally capable of identifying salient features for predicting the accuracy of each response.

5.6 Conclusion

In this chapter, we developed a probabilistic framework for improving the quality control of crowdsourced responses by leveraging side information from questions, crowd-workers, working sessions and responses. The respective source features of the side information include the genres

and the content features of the questions, the demographics and the personality traits of the crowdworkers, the user access devices and the working time periods for the sessions, and the durations and the orders of the responses. The efficacy of the framework has been demonstrated on three new crowdsourcing datasets, where we have observed overall consistent improvements in both the accuracy and the logarithmic loss for true answer prediction. The improvements have also been observed on unseen (held-out) response prediction. All these improvements are achieved under the situation where the responses are scarce across both the workers and the questions. Moreover, response-level information was found particularly useful for helping the framework to account for worker-specific biases. In addition, our framework was found to be promising at identifying salient source features without any supervised information.

Based on the work in this chapter, we are able to answer research question *RQ*2 with all its sub-questions from *RQ*2.1 to *RQ*2.4:

- **Answer to *RQ*2.1**: The basic framework is chosen to be the GLAD model due to the fact that it decomposes the response quality into two factors: the *worker*-specific *expertise* factor and the *question*-specific *difficulty* factor. The two factors allow for straightforward and elegant incorporation of their respective side information and the *contextual information* by *linearly* adding regressions over the corresponding observed features.
- **Answer to *RQ*2.2**: Each type of side information is incorporated into the basic framework as the observed features of the corresponding crowdsourcing aspects. Worker and question features are mapped by linear regressions parametrized by global coefficients into some scalars which are added linearly to their respective factors in the framework. Session features (whose values are specific to each worker's individual working session) are mapped and added in the same way as the worker features. Response features are preferably mapped by linear regressions specific to individual workers with their respective coefficients to account for their response biases [25].
- **Answer to *RQ*2.3**: Our experiment results show that when all the available types of side information are incorporated into the basic framework, the resulting full-information framework is able to outperform the basic framework: the GLAD model, the community-aware DS model which is a state-of-the-art approach in handling response sparsity problem, and the basic majority vote by the largest performance margins respectively. Thus, it is the most useful in alleviating the response sparsity problem both on the worker and question sides.

- **Answer to *RQ*2.4**: Our experiment results also show that when the number of responses from each worker is only one (and thus the scarcest possible scenario), the full-information framework is able to achieve the largest performance margins against the above three baselines. In other words, considering all types of side information available is most useful for handling the most extreme response sparsity that can happen on the side of workers.
- **Overall Answer to *RQ*2**: Extending the GLAD model allows us to build a *unified scalable* quality control framework by *linearly* adding:
  - *global* regressions over features regarding *workers*, *their sessions* and *questions* into the *worker expertise* and *question difficulty* factors of GLAD;
  - *local* regressions specific to *individual workers* over their respective response feature vectors into the corresponding response correctness probabilities to account for their *response biases*.

  Our framework is able to leverage the worker, question and contextual side information to improve the response quality estimates and further, the true answer estimates. According to our experiments, the improvement is overall maximized when all types of side information available are leveraged by our framework.

There is an interesting possibility for future work based on this chapter:

- investigation of the possibility of extending our framework to handle the case of multi-dimensional worker expertise and question difficulty [25];

Chapter 6

Distinguishing Question Subjectivity from Difficulty for Improved Quality Control

In this chapter, we endeavour to answer the third research question (i.e. *RQ*3) with all its sub-questions (i.e. *RQ*3.1 to *RQ*3.4) by investigating the possibility of separately modelling and estimating the *difficulty* and the *subjectivity* of questions to better account for the quality of responses given to them.

According to our literature review, more recent (and more advanced) quality control methods model the influence that individual questions exert on the quality of responses [10, 31, 87, 88], apart from modelling workers' abilities as done by the earlier work [56, 83]. In particular, two key attributes of questions have drawn the modelling attention:

- **Difficulty.** The modelling of question difficulty is founded on the assumption that greater agreement among workers' responses to a question indicates less difficulty for them in determining the correct response. Quality control methods often encode this assumption using a function in which worker ability counteracts question difficulty for predicting the probability of a correct response. The probability is known as the *quality of the response*: the more difficult the question, the lower the quality of a response, and vice versa. In addition, some methods [10] also consider the existence of *deceptive* questions which are so difficult that the assumption that the majority of responses are correct no longer holds on them.

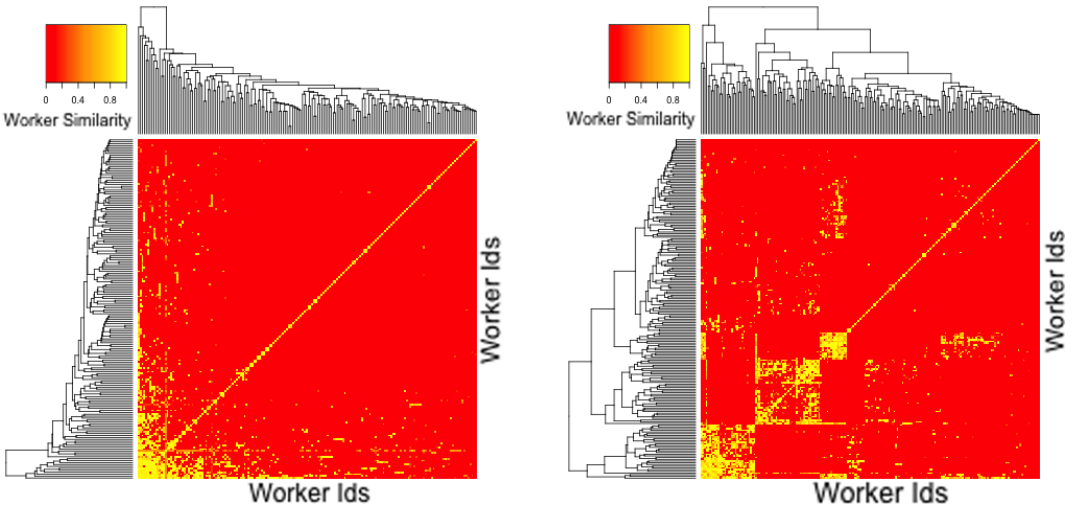
- **Subjectivity.** In crowdsourcing, there exist tasks that contain (*purely* or *partially*) *subjective* questions [26]. Intuitively, the degree of subjectivity of a question depends on the number of response options that are correct. Being purely subjective means all of the options are correct, while being partially subjective means more than one but not all of them are correct. It is widely recognised that even expert assessors can disagree with each other on the correct responses to questions in tasks, such as relevance judgement, which are considered to be partially subjective [182]. In this case, the objectivity assumption on the questions does not hold and most of the quality control methods based on this assumption cannot distinguish the *ability* of workers from their *preferences* for the different response options.

Note that question subjectivity is different from the *domain* difficulty of *objective* questions which involves different groups of workers finding different types of objective questions more or less challenging than others. Handling such domain difficulty would require the modelling of *domain-level expertise*. For example, geography students might find physics questions more difficult than geography questions and vice versa. Modelling domain-level expertise and (objective) difficulty has been studied by Welinder et al. [25]. We did not study this subject in this thesis but extending our work to account for domain expertise and difficulty should be relatively straightforward.

When dealing with subjective questions, collaborative filtering for recommendation [112] considers users who share similar preferences to form *groups*. Those within the same group respond similarly towards subjective questions (e.g. rating movies) which share certain characteristics. This grouping effect can also be observed in crowdsourcing when crowd-workers respond to partially subjective questions. Figure 6.1 illustrates this effect by providing heat maps of pairwise worker similarity for two tasks: (a) a relatively objective task and (b) a more subjective one. The objective task required workers to judge whether a pair of products were the same based on their names, descriptions and prices, while the subjective task asked workers to judge whether an image contained "fashion related items"[1]. The similarity between pairs of workers is calculated as the percentage of agreement across the jointly answered questions[2] and hierarchical clustering was performed to group similar workers together. The three yellow boxes along the diagonal for the more subjective task (b) indicates the three distinct groups of worker response behaviour for this task, which was absent in the more objective task (a).

[1]The datasets for the two tasks have been listed in section 6.4.

[2]Pairs of crowd-workers not sharing any questions had their similarity set to be .00001. We also investigated the effects of smoothing to the similarity calculation by adding $\frac{1}{|\mathcal{K}|}$ and 1 to the nominator and the denominator of the agreement ratio respectively, but the resulting changes to the clusters in the heat maps were negligible.



(a) (b)

Figure 6.1: Heatmaps showing inter-worker response similarity (% of response agreement) for two different tasks: (a) a *relatively objective* product matching task and (b) a *more subjective* fashion judging task, both involving binary worker responses. Hierarchical clustering was performed to order workers such that similar workers are close together. The three yellow blocks in (b) indicate three groups of response behaviour and higher subjectivity for task (b).

We can see from Figure 6.1a, there is no real cluster of crowd-workers, while from Figure 6.1b, there are three distinct clusters. Workers belonging to the same cluster exhibit similar labelling behaviour while there are differences between the clusters. Since the workers were mostly reliable on both tasks[3], we conjecture that the grouping of workers in the fashion judgement task reflects tendency to provide different (yet equally correct) responses to the same question (due to its subjectivity).

To facilitate response quality control for the above tasks and generally, any crowdsourcing task that exhibits arbitrary degrees of question subjectivity and difficulty, we are motivated to develop a statistical model encoding both these properties. More specifically, question subjectivity causes groups of crowd-workers to emerge. A group specifies a particular correlation between the crowd-workers' responses and indicates that they agree on the latent correct responses to different (partially or purely) subjective questions. We model such a correlation by factorising it into the latent preferences of the workers and the latent features of those questions. The assumption is that workers with similar latent preferences tend to perceive the same correct response when responding to partially or purely subjective questions with similar characteristics. Moreover, for partially subjective questions, which possess varying degrees of inherent difficulty, this difficulty corrupts the crowd-workers' ability to provide the correct responses to the question (as determined by their

[3]We estimated for the two tasks the accuracy of workers based on their agreement with the majority vote, and found that most workers had an accuracy greater than 0.5.

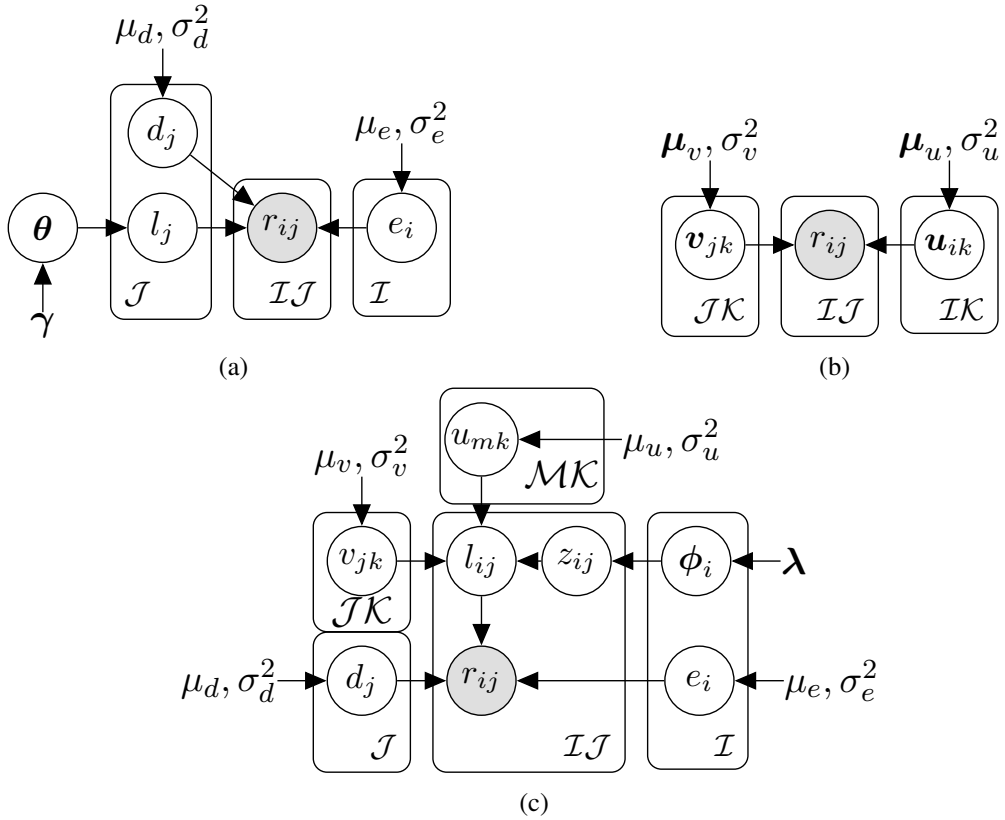


Figure 6.2: (a) shows GLAD with a latent variable l_j for each objective truth, (b) shows a collaborative filtering model without objective truths, and (c) is the proposed *subjectivity-and-difficulty response* (SDR) model for partially-subjective questions that is able to distinguish question difficulty from subjectivity.

subjective preferences) to a certain extent depending on the worker's level of expertise. A greater extent of the corruption is captured as a lower probability that the worker's response is equal to the subjective (worker-specific) correct response to the question, thereby the lower quality of the worker's response.

Our proposed model *considers both the subjective (i.e. worker-specific) truths regarding the correct answers to individual questions and also the difficulty-dependent probability that a worker's response to a question will equal her perceived subjective truth*. It encodes the question difficulty explicitly and the question subjectivity implicitly via latent variables for worker preferences and corresponding question features. Based on this model, we provide a Monte Carlo simulation approach for quantifying question subjectivity as the expected number of subjective truths perceived by different groupings of crowd-workers with respect to their preferences. Finally, we derive from the model a ranking of questions in terms of either difficulty or subjectivity which is found to be coherent with human assessment.

6.1 Related Work

6.1.1 Latent Variable Modelling in Crowdsourcing

Most state-of-the-art response quality control methods in crowdsourcing have operated under the assumption that each question is purely objective. These methods are primarily based on statistical modelling of the interactions between crowd-workers and questions which determine either the marginal probabilities of the workers' responses being equal to the corresponding correct answers [25, 87] or the conditional probabilities of the responses given the correct answers [56, 73]. In comparison, the marginal probabilistic modelling is simpler than the conditional modelling, and also better at mitigating the response sparsity problem in crowdsourcing. The basic marginal probabilistic modelling is the GLAD model from Whitehill et al. [87], which models the correctness of each response as a logistic function where the question difficulty counteracts the expertise of the responding worker. Its graphical representation is shown by Figure 6.2a with the following generative scheme for a response r_{ij} of worker i given to question j: $\boldsymbol{\theta} \sim Dir(\boldsymbol{\gamma})$; $l_j \sim$ Categorical($\boldsymbol{\theta}$); $r_{ij} \sim$ Categorical($\boldsymbol{\psi}_{ijl_j}$). This means a true answer l_j is drawn for question j from a discrete distribution parametrised by $\boldsymbol{\theta}$, which was previously drawn from a Dirichlet distribution parametrised by $\boldsymbol{\gamma}$. Then, a response r_{ij} is drawn, conditioned on l_j, from a discrete distribution over the set of response options $\mathcal{K}$. The parameters of this conditional distribution is a probability vector $\boldsymbol{\psi}_{ijl_j}$. Its k-th component ψ_{ijl_jk}, $k \in \mathcal{K}$ is calculated as follows:

$$\psi_{ijl_jk} = \begin{cases} f(e_i, d_j) & \text{If } l_j = k \\ \frac{1-f(e_i,d_j)}{K-1} & \text{Otherwise} \end{cases} \qquad \text{where} f(e_i, d_j) = \frac{1}{1 + e^{-(e_i/\exp(d_j))}} \tag{6.1}$$

The function f takes in the expertise factor e_i of worker i and the difficulty factor d_j of question j. The output of the function is the probability of the response r_{ij} being correct. When $e_i \to +\infty$ or $d_j \to 0$, this probability grows, indicating a stronger positive correlation between r_{ij} and l_j. When $e_i \to 0$ or $d_j \to +\infty$ and the question has binary options, the probability approaches 0.5, which suggests r_{ij} is arbitrarily picked. When $e_i \to -\infty$, the probability decreases to 0, indicating a stronger negative correlation.

Few quality control methods have considered modelling the subjectivity of questions. One of the two papers that have made progress in this regard is by Tian and Zhu [111]. It assumes that a higher joint degree of difficulty and subjectivity for a crowdsourcing task will increase

(decrease) the number of groups of responses given to the questions. The expected size of each group becoming smaller (larger) indicates overall weaker (stronger) correlations of responses given to the questions. This paper makes no attempt in separating difficulty and subjectivity when only the difficulty accounts for the *quality of responses*. Moreover, this work requires every question to be answered by every worker, which is unrealistic in crowdsourcing. The other work by Nguyen et al. [26] has focused on modelling partially subjective questions with just ordinal answers. It assumes each response to a question is generated by a univariate Gaussian the mean and the variance of which are linearly regressed over the observed features of the question. This means the model fits poorly any multi-modal distribution of responses (e.g. responses distributed only on rating 1 and 5).

6.1.2 Latent Variable Modelling in Collaborative Filtering

In model-based collaborative-filtering [112], matrix factorization is applied to predicting ordinal ratings provided by users to items (e.g. movies). Its categorical version, shown in Figure 6.2b, is less commonly applied but is important for the construction of our model for the quality control of crowdsourced categorical responses. It has a generative process: $r_{ij} \sim \text{Categorical}(\boldsymbol{\rho}_{ij})$ where $\boldsymbol{\rho}_{ij} = \{\rho_{ijk}\}_{k\in\mathcal{K}}$ with its k-th component calcuated as:

$$\rho_{ijk} = P(r_{ij} = k|\{\boldsymbol{u}_{ik}, \boldsymbol{v}_{jk}\}_{k\in\mathcal{K}}) = \frac{\exp(\boldsymbol{u}_{ik}^T \boldsymbol{v}_{jk})}{\sum_{k'\in\mathcal{K}} \exp(\boldsymbol{u}_{ik'}^T \boldsymbol{v}_{jk'})} \tag{6.2}$$

Here, $\boldsymbol{\rho}_{ij}$ is also called the *soft-max* function, $\boldsymbol{u}_{ik}$ and $\boldsymbol{v}_{jk}$ are respectively the latent preferences of worker i and the latent features of item j in relation to the k-th answer option. The inner product term $\boldsymbol{u}_{ik}^T \boldsymbol{v}_{jk}$ indicates how much tendency user i responds to item j with the k-th answer option. Note that the response probabilities $\boldsymbol{\rho}_{ij}$ for collaborative filtering is different from $\boldsymbol{\rho}_{ij|l_j}$ for crowdsourcing as the former does not deal with question true answers.

6.2 Proposed Model

We propose a new model, which endeavours to combine the key characteristics of the latent variable models specified in sections 6.1.1 and 6.1.2. We call it the SDR model (Subjectivity-and-Difficulty Response model), which comprises an *upstream module* that generates a *subjective* truth for a question based on the worker's perception of the correct answer, and a *downstream module* that imposes a *difficulty*-dependent corruption on the subjective truth for generating the actual response

from the worker to the question. More specifically, in the upstream module, the latent subjective truth l_{ij} of question j as perceived by crowd-worker i is drawn from a soft-max function specified by equation 6.2 (from the Collaborative Filtering model), except that the worker's response r_ij in the equation is now replaced by the worker-specific true answer l_{ij}. This function explains how the worker's latent preferences interact with the question's latent features to generate the subjective truth behind her response to the question. In the downstream module, conditioned on the latent subjective truth l_{ij}, the response r_{ij} actually given by worker i to question j is determined by the logistic function $f(e_i, d_j)$. It encodes how the question difficulty d_j counteracts the worker expertise e_i to corrupt the subjective truth in order to produce the observed response, which will be defined later in this section. Essentially, the above perception-corruption process generalises the response corruption process for objective questions modelled in Welinder et al. [25] by additionally considering the question subjectivity.

Unfortunately the upstream+downstream model described above suffers from an over-parametrisation issue whereby *both* the upstream component (which determines the worker-specific correct answer) and the downstream component (which determines the noise resulting from worker inaccuracy) can *independently and adequately* explain the variance observed in worker responses to the same question. In other words, the varied responses from different workers to the same question could equally be due to different perceptions on what constitutes the correct answer to the question or to difficulty of the question causing low accuracy amongst the respondents. To remedy this situation we explicitly enforce a group structure over workers in order to limit the variation in the perceptions across workers. This is done by changing the upstream module to have *sparsity-inducing priors* over the latent preferences of crowd-workers.

In this chapter, we use the *Latent Dirichlet Allocation* (LDA) [107] to provide such priors by imposing a flexible and non-exclusive group structure over workers. The final graphical representation of the SDR model is shown in Figure 6.2c. The new upstream module of our model assigns a probability vector $\boldsymbol{\phi}_i$, which follows a Dirichlet with a concentration parameter $\boldsymbol{\lambda}$, to each worker i. Each component ϕ_{mi} of this probability vector reflects the worker's tendency to show a particular preference m among the set of preferences $\mathcal{M}$ she possesses when answering any question. Then, a preference assignment z_{ij} is drawn from $\boldsymbol{\phi}_i$ for determining the specific preference worker i will show for answering question j. As for preference m, it has a weight u_{mk} for each response option k to determine the extent to which it considers option k as the correct response. In this paper, we fix the dimension of u_{mk} to be simply 1. This weight is multiplied with

the latent feature v_{jk} of question j whose value (again, positive or negative) measures the tendency of each option k to be preferred as the correct response. The result of this multiplication is input to a soft-max function for drawing the subjective truth l_{ij} behind the response r_{ij}.

The above generative process can be formulated as: $\boldsymbol{\phi}_i \sim Dir(\boldsymbol{\lambda})$; $z_{ij} \sim$ Categorical$(\boldsymbol{\phi}_i)$; $l_{ij} \sim$ Categorical$(\boldsymbol{\rho}_{z_{ij}})$[4] with the k-th component of the soft-max function $\boldsymbol{\rho}_{z_{ij}}$ calculated as follows:

$$\rho_{z_{ij}k} = \frac{\exp(u_{z_{ij}k} v_{jk})}{\sum_{k' \in \mathcal{K}'} \exp(u_{z_{ij}k'} v_{jk'})} \tag{6.3}$$

Embodying the sparsity-inducing effect of LDA, the preference probabilities $\boldsymbol{\phi}_i$ are dedicated to revealing the underlying groups of crowd-workers. The soft-max specified by equation 6.3 captures the positive correlations between the correct responses to the same question perceived by the workers within the same group. When the number of preferences $|\mathcal{M}| = 1$, ϕ_i only has one element with value 1. This has a two-fold meaning that each question has one correct answer and that every worker should perceive the correct answer of any question in the same way. When $|\mathcal{M}| > 1$, this indicates there exist some underlying worker groups. We can recover these groups by applying K-means clustering to the estimated preference probabilities $\hat{\boldsymbol{\phi}}_i$ with the Elbow method [190] to determine the right number of groups.

The downstream module corrupts the correlations between the subjective truth l_{ij} and the response r_{ij}. It draws r_{ij} from a discrete probability distribution $\boldsymbol{\psi}_{ij}$ specified in equation 6.1.1 except the logistic function $f(e_i, d_j)$ has the following definition from Rasch [191]:

$$f(e_i, d_j) = \frac{1}{(1 + e^{-(e_i - d_j)})} \tag{6.4}$$

The term $(e_i - d_j)$ naturally explains the type of biases induced by *deceptive* questions when the difficulty d_j is much larger than the average value over worker expertise e_i. This is not captured in $f(e_i, d_j)$ in equation 6.1.1 as the term $\exp(d_j)$ is never smaller than 0, meaning questions never bias workers to answer incorrectly due to their difficulty.

If more responses have their corresponding $(e_i - d_j)$ estimated to be greater than zero, it means that more of them are deemed correct. This results in the number of inferred correct answers to each question to increase. As a result of this increase, the number of latent preferences $\mathcal{M}$ needs to grow, from the perspective of SDR, to fit the seemingly more diverse correlations between latent

[4]Note that this is different from $\boldsymbol{\rho}_{ij}$ in equation 6.2 which focuses on individual workers while $\boldsymbol{\rho}_{z_{ij}}$ focuses on the z_{ij}-th preference of workers.

correct answers across the questions. Thus, for the SDR model to recover the right number of latent preferences, the priors for e_i and d_j need to be set up properly, which will be elaborated more in section 6.4.2.

6.3 Estimation

6.3.1 Model Parameter Estimation

We now provide equations used for parameter estimation, using the notation $\rho_{z_{ij}k}$ from equation 6.3 and $f(e_i, d_j) = f_{ij}$ from equation 6.4 to simplify the equations. The conditional probability for the preference assignment z_{ij} to worker i when answering question j is:

$$P(z_{ij}{=}m|e_i, d_j, \boldsymbol{u}_m, \boldsymbol{v}_j, \lambda) \propto \sum_{k\in\mathcal{K}} \rho_{mk} f_{ij}^{\delta_{ijk}} \left(\frac{1-f_{ij}}{K-1}\right)^{1-\delta_{ijk}} \frac{N^m_{\backslash ji} + \lambda_m}{\sum_{m'\in\mathcal{M}} N^{m'}_{\backslash ji} + \lambda_{m'}} \tag{6.5}$$

where $\delta_{ijk} = \mathbb{1}\{r_{ij} = k\}$ and $N^m_{\backslash ji}$ denotes the number of questions excluding question j answered by worker i given her preference m. The joint distribution of the other parameters given z_{ij} is:

$$\begin{aligned} Q &= p\Big(\{e_i\}_{i\in\mathcal{I}}, \{d_j, \boldsymbol{v}_j\}_{j\in\mathcal{J}}, \{\boldsymbol{u}_m\}_{m\in\mathcal{M}} | \{\boldsymbol{z}_{ij}\}_{i\in\mathcal{I}, j\in\mathcal{J}}, \mu_{\{e,d,u,v\}}, \sigma^2_{\{e,d,u,v\}}\Big) \\ &= -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \log\left[\sum_{k\in\mathcal{K}}\left(\rho_{z_{ij}k} f_{ij}^{\delta_{ijk}}\left(\frac{1-f_{ij}}{K-1}\right)^{1-\delta_{ijk}}\right)\right] - \sum_{i\in\mathcal{I}} \log\left(p(e_i|\mu_e, \sigma_e^2)\right) - \\ &\sum_{j\in\mathcal{J}} \log\left(p(d_j|\mu_d, \sigma_d^2)\right) - \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} \log\left(p(u_{mk}|\mu_u, \sigma_u^2)\right) - \sum_{j\in\mathcal{J}} \log\left(p(v_{jk}|\mu_v, \sigma_v^2)\right) \end{aligned} \tag{6.6}$$

The partial derivatives for Q with respect to e_i and d_j are as follows:

$$\frac{\partial Q}{\partial e_i} = -\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \left(\frac{-1}{K-1}\right)^{1-\delta_{ijk}} f_{ij}(1-f_{ij})\rho_{z_{ij}k} + \frac{e_i - \mu_e}{\sigma_e^2} \tag{6.7}$$

$$\frac{\partial Q}{\partial d_j} = -\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} (-1)^{\delta_{ijk}} \left(\frac{1}{K-1}\right)^{1-\delta_{ijk}} f_{ij}(1-f_{ij})\rho_{z_{ij}k} + \frac{d_j - \mu_d}{\sigma_d^2} \tag{6.8}$$

The partial derivatives for Q with respect to the k-th component of $\{\boldsymbol{u}_{mk}, \boldsymbol{v}_{jk}\}_{k\in\mathcal{K}}$ are as follows:

$$\frac{\partial Q}{\partial u_{mk}} = -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \zeta_{ijm} v_{jk} \sum_{k'\in\mathcal{K}} f_{ij}^{\delta_{ijk'}} \left(\frac{1-f_{ij}}{K-1}\right)^{1-\delta_{ijk'}} \omega_{ijkk'} + \frac{u_{mk} - \mu_u}{\sigma_u^2} \tag{6.9}$$

$$\frac{\partial Q}{\partial v_{jk}} = -\sum_{i\in\mathcal{I}} u_{z_{ij}k} \sum_{k'\in\mathcal{K}} f_{ij}^{\delta_{ijk'}} \left(\frac{1-f_{ij}}{K-1}\right)^{1-\delta_{ijk'}} \omega_{ijkk'} + \frac{v_{jk} - \mu_v}{\sigma_v^2} \tag{6.10}$$

Here, $\zeta_{ijm} = \mathbb{1}\{z_{ij} = m\}$ and $\omega_{ijkk'} = \rho_{z_{ij}k}\big(1 - \rho_{z_{ij}k}\big)^{\mathbb{1}\{k=k'\}}\big(- \rho_{z_{ij}k'}\big)^{\mathbb{1}\{k\neq k'\}}$. The parameter estimation involves two alternating procedures: sampling z_{ij} according to equation 6.5 and optimizing Q in equation 6.6 using LBFGS [5] based on equations 6.7, 6.8, 6.9, and 6.10.

6.3.2 True Answer Estimation

A worker-specific subjective truth l_{ij} (as perceived by worker i) fails to provide overall information about the correct answers to partially subjective question j. Thus, we should gather the l_{ij} values from all workers who answer the question. However, in practice, a question is usually assigned to only a limited number (usually 3 or 5) of the workers, making the estimate of the true answer distribution poor. Our solution for improving this estimate is to first find underlying clusters of workers (across all questions) by applying K-means with the Elbow method based on 10-fold cross validation to the posterior means $\{\hat{\boldsymbol{\phi}}_i\}_{i\in\mathcal{I}}$ of the latent preference probabilities of all the workers. With the centroid $\hat{\boldsymbol{\phi}}_c$ of each resulting cluster c, we then calculate the probability that the true answer l_{cj} (as perceived by the workers in cluster c) takes the value k as follows:

$$P(l_{cj} = k|c) = \sum_{m\in\mathcal{M}} P(l_{cj} = k|m)P(m|c) = \sum_{m\in\mathcal{M}} \left(\frac{\exp(\hat{u}_{mk}\hat{v}_{jk})}{\sum_{k'=1}^{K} \exp(\hat{u}_{mk'}\hat{v}_{jk'})} \hat{\phi}_{mc} \right) \quad (6.11)$$

where $\hat{u}_{mk}$ and $\hat{v}_{jk}$ are the estimates of the weight u_{mk} for preference m and the latent feature v_{jk} of question j, both specific to option k. The best estimate regarding the correct answer l_{cj} according to the workers assigned to cluster c is then:

$$\hat{l}_{cj} = \arg\max_{k\in\mathcal{K}} P(l_{cj} = k|c) \quad (6.12)$$

Now we have a set of correct answer estimates $\hat{\mathcal{L}}_j = \{\hat{l}_{cj}|c \in \mathcal{C}\}$ for question j from all the worker clusters (with $\mathcal{C}$ being the set of the clusters). For the task of (single) true answer prediction, we choose one label from $\hat{\mathcal{L}}_j$ as predicted true answer l_j by following certain strategies. Two simple strategies are to choose $\hat{l}_{cj}$ from the cluster c with the highest average expertise over its workers, or from the cluster with the largest proportion of workers assigned to it. The first strategy states that the answer perceived by on average the most expert group of workers is the most appropriate, while the second assumes it to be the one perceived by the largest group of workers which should

[5] Two of the most popular iterative optimization algorithms used by quality control models are: *steepest gradient descent* [25, 87, 94, 169] and *BFGS* [26]. Compared to BFGS, L-BFGS is faster and more stable for handling large numbers of parameters. L-BFGS is also similar to the conjugate gradient algorithm, thereby having similar performance which is typically better than steepest gradient descent [192].

represent the mainstream school-of-thought. In this chapter, we apply the second strategy because most crowdsourcing datasets used in the experiments correspond to relatively simple tasks for which the provided correct answers we believe are more likely to be mainstream opinions. As for the first strategy, it can be more useful for revealing a minority group of expert workers who show distinct preferences from the public.

Algorithm 1: Subjectivity estimation for question j

Input: $\hat{v}_j$; $\{\hat{\boldsymbol{u}}_m\}_{m\in\mathcal{M}}$; $\hat{\boldsymbol{\Phi}}_c = \{\hat{\boldsymbol{\phi}}_c\}_{c\in\mathcal{C}}$; $T = 50,000$ (maximum number of iterations).

Output: $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ /* Expected number of correct answers for question j */

$n_j \leftarrow 0$ /* Initialise number of correct answers for question j to zero */

for $t \leq T$ *and the sample standard deviation of estimates for* $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ *over* $(t-100, t] \geq 10^{-4}$ **do**

  $\hat{\mathcal{L}}_j \leftarrow \{\}$ /* Initialise set of correct answers to be sampled at iteration t. */

  for *each* $c \in \mathcal{C}$ **do**

    $z_{cj} \sim \text{Categorical}(\hat{\boldsymbol{\phi}}_c)$ /* Sample group preference z_{cj}. */

    $\hat{l}_{cj} \sim \boldsymbol{\rho}_{z_{cj}}$, where $\rho_{z_{cj}k} = \frac{\exp(\hat{u}_{z_{cj}k}\hat{v}_{jk})}{\sum_{k'\in\mathcal{K}} \exp(\hat{u}_{z_{cj}k'}\hat{v}_{jk'})}$ /* Sample correct answer $\hat{l}_{cj}$ perceived by worker cluster c. */

    $\hat{\mathcal{L}}_j \leftarrow \hat{\mathcal{L}}_j \cup \hat{l}_{cj}$ only if $\hat{l}_{cj} \notin \hat{\mathcal{L}}_j$ /* Add sampled $\hat{l}_{cj}$ to $\hat{\mathcal{L}}_j$ when it first appears. */

  $n_j^{(t)} \leftarrow n_j^{(t-1)} + |\hat{\mathcal{L}}_j|$ /* Increase n_j by number of distinct correct answers sampled at t.*/

  $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|] \approx \frac{n_j^{(t)}}{t}$ /* Divide n_j by t as estimate of question j's subjectivity at t. */

6.3.3 Subjectivity Estimation

Despite not being directly estimated in the model, question subjectivity can still be quantified and estimated after the model has been learned. This is achieved based on the reasonable assumption that the subjectivity of each question is proportional to the number of correct answers it has. Despite not knowing the actual number of correct answers $|\mathcal{L}_j|$ to question j, we can estimate the value by taking its expectation with respect to the clusters of workers derived in section 6.3.2. More precisely, the expected number of correct answers to question j with respect to worker clusters $\mathcal{C}$ is: $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|] = \sum_{n=1}^{|\mathcal{K}|} nP_{\mathcal{C}}(|\mathcal{L}_j| = n)$. In this equation, n iterates over the possible number of correct answers (from 1 to the size of $\mathcal{K}$). The probability $P_{\mathcal{C}}(|\mathcal{L}_j| = n)$ denotes how likely it is that the number of correct answers to question j equals n, with respect to the worker clusters $\mathcal{C}$. When $\mathcal{C}$ and $\mathcal{K}$ are large, it is difficult to calculate this probability due to a combinatorial explosion. Thus we apply Monte Carlo simulation to estimate (a measure of) the subjectivity of question j

as $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ using Algorithm 1. To test the convergence of this algorithm, we monitor the *sample standard deviation* of estimates for $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ over every 100 consecutive iterations (i.e. $(t-100, t]$). The reason for monitoring this specific quantity is that it is always large at the beginning of the sampling process. As the process proceeds, the sample standard deviation becomes smaller and smaller. When the process converges, it stabilises at a very small value ϵ. We observed $\epsilon = 10^{-4}$ to be a reasonable threshold value across the partially subjective datasets used in the experiments. Finally, the algorithm outputs $n_j^{(t)}/t$ as the estimate of $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ where t is either the iteration that satisfies the 10^{-4} threshold or the maximum number of iterations T.

6.4 Experiments and Results

The evaluation of our proposed model consists of four parts. The first part is its sensitivity to various degrees of subjectivity in different crowdsourcing tasks. The second and the third parts are its performance in predicting the provided correct answers for questions and the responses to be given by crowd-workers to unseen questions, respectively. The last part is its consistency with human assessors in assessing the difficulty and the subjectivity of questions.

6.4.1 Datasets

We have used 10 crowdsourcing datasets to evaluate the performance of our model across the four parts. Table 6.1 summarises these datasets as being either (primarily) objective or partially subjective. The identification tasks of *event time ordering* (Time), *dog* and *duck breeds*, and *identical products* (Product) concern objective factual knowledge. The judgement tasks of *image beauty* (Image), *document relevance* (Rel 1&2), *fashionable items* (Fashion), *facial expression* (Face) and *adult content* (Adult) contain certain degrees of subjectivity.

Objective datasets	**($\|\mathcal{I}\|$, $\|\mathcal{J}\|$, $\|\mathcal{R}\|$)**	***Partially subjective datasets***	($\|\mathcal{I}\|$, $\|\mathcal{J}\|$, $\|\mathcal{R}\|$)
Time [57]	(76, 462, 4620)	**Image [111]**	(402, 60, 24120)
Dog [89]	(109, 807, 8070)	**Rel1 [193]**	(642, 1787, 13310)
Duck [25]	(53, 240, 9600)	**Rel2 [194]**	(83, 585, 1755)
Product [195]	(176, 8315, 24945)	**Fashion [196]**	(199, 3837, 11511)
		Face [197]	(27, 584, 5242)
		Adult [198]	(269, 333, 3324)

Table 6.1: The objective and the partially subjective datasets used in this paper. The headers correspond to the notation introduced in Table 1.1.

The details of the datasets in Table 6.1 are as follows:

Temporal Ordering [57] (**Time**). This task asks workers to judge the temporal relations between pairs of events in documents. The answers can be: "strictly before" & "strictly after".

Dog Breed Identification [89] (**Dog**). The task asks workers to identify the correct dog breed out of 4 possible breeds for each dog image.

Duck Identification [25] (**Duck**). The task asks workers to judge whether there is a duck in each image.

Product Matching [195] (**Product**). This task asks workers to judge whether pairs of products are the same or not based on their names, descriptions and prices.

Object Identification & Image Beauty Judgement [111] (**Image**). This task contains 5 sub-tasks (i.e. **Beauty 1 & 2, Sky, Building, Computer**). Each sub-task asks workers to either judge whether a particular object (i.e. Sky, Building, Computer) is present in each image or pick images they deem beautiful (i.e. Beauty 1 & 2).

Relevance Judgement 1 [193] (**Rel1**). The task asks workers to judge the relevance of documents to some queries. Workers can choose from one of three relevance levels: "highly relevant", "relevant", and "non-relevant".

Relevance Judgement 2 [194] (**Rel2**). The questions of this task come from the part of TREC 2011 crowdsourcing track [194] that does not contain the questions of relevance judgement task 1. We collected crowdsourced judgements for this task from CrowdFlower.

Fashion Judgement [196] (**Fashion**). The task asks workers to judge whether each image contains any "fashion-related" object or not. The answers can be "Yes", "No" and "Not sure".

Facial Expression Judgement [197] (**Face**). The task asks workers to label the facial expressions of different people with different orientation of their faces. The answers can be "Sad", "Angry", "Neutral", and "Happy".

Adult Content Judgement [198] (**Adult**). The task asks workers to classify the adult levels of different websites. There are four levels available: "G" (General Audience), PG (Parental Guidance), "R" (Restricted) and "X" (Porn).

6.4.2 SDR Hyper-parameter Setup

As discussed at the end of section 6.2, to find the right number of latent preferences, the hyper-parameters of the expertise e_i and the difficulty d_j in the SDR model need to be set properly. This is achieved by *held-out validation* which leverages *noise* within worker responses for detecting signs that SDR might be overfitting the responses by introducing more latent preferences than necessary.

We construct a held-out validation dataset by randomly sampling a response from each worker. Thus, the size of such a dataset equals the number of workers participating in a task. Then, given a certain setting of the hyper-parameters, we learn our model based on the remaining responses and use the parameter estimates from the learned model to calculate the prediction accuracy: $1-MAE$ (Mean Absolute Error) over the held-out dataset. We repeat the model learning process with each hyper-parameter setting over the same 100 random held-out validation data subsets. We then obtain the average prediction accuracy for our model across these subsets for each hyper-parameter setting. Finally, we choose the setting (including the number for latent preferences) that yields the highest average prediction accuracy for use in the experiments.

6.4.3 Sensitivity Analysis

We first verify whether SDR is sensitive to various degrees of subjectivity in different crowdsourcing tasks. If a task is (almost entirely) objective, the optimal size of the latent preference set $\mathcal{M}$ should be 1, meaning that every worker perceives the correct answers in the same way. Consequently, the probabilities of latent preferences $\boldsymbol{\phi}_i$ for worker i collapse to $\phi_i = 1$, and the set of true answers $\mathcal{L}_j$ for question j collapses to a single true answer l_j. We conduct the held-out validation on our model across the objective datasets each with the 100 randomly sampled data subsets described in section 6.4.2. We expect that the average held-out prediction accuracy for SDR across these data subsets will decrease when the number of latent preferences it has increased from 1 to 2, since in this case the model starts to overfit by learning the noise in the training responses. If a task is sufficiently subjective, our model should uncover the right number of underlying groups of workers along with the right number of latent preferences. We conduct the experiment in the same way to see the difference in average prediction accuracy on held-out unseen responses with the number of preferences increasing from 1 to 3 over the partially subjective datasets. We expect the average prediction accuracy to be higher when the number of preferences is greater than 1. Since Tian and Zhu [111] has provided us with the number of worker clusters emerging respectively from the five sub-tasks which constitute the *Image* data in Table 6.1, we thus compare the corresponding numbers of clusters derived from our model with theirs.

The results of the sensitivity analysis are shown in Tables 6.2 and 6.3. We can see from Table 6.2 that the average prediction accuracy of *SDR with 1 latent preference* is constantly higher than that of *SDR with 2 preferences* over all the objective datasets. This result indicates there is just

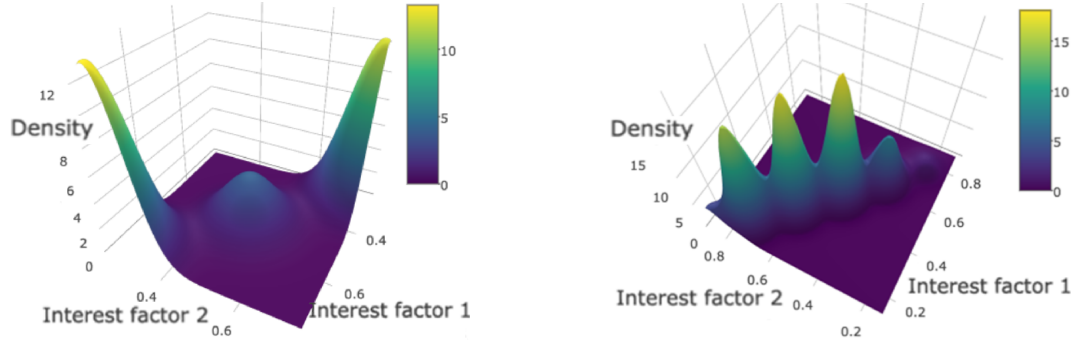
Dataset	The SDR model	
	m = 1	m = 2
Time	**0.8967**	0.8915
Dog	**0.6970**	0.6625
Duck	**0.8427**	0.8388
Product	**0.8396**	0.8291

Table 6.2: Average accuracy of our model with 1 and 2 latent preferences on predicting the held-out validation response of each worker over 4 objective tasks.

Dataset	The SDR model		
	m = 1	m = 2	m = 3
Beauty 1	0.6736	**0.6944**	0.6924
Beauty 2	0.6914	**0.6998**	0.6937
Sky	0.8889	**0.8962**	0.8862
Building	0.8997	**0.9026**	0.9007
Computer	0.8098	**0.8117**	0.8074
Rel1	0.3956	**0.3985**	0.3983
Rel2	0.4426	**0.4481**	0.4481
Fashion	0.7517	**0.7589**	0.7522
Face	0.7181	**0.7203**	0.7123
Adult	0.7469	**0.7494**	0.7446

Table 6.3: Average accuracy of our model with 1, 2 and 3 latent preferences on the held-out validation prediction over 10 partially subjective tasks the first 5 of which are sub-tasks of the *Image* task.

one underlying group of workers for each of the tasks. This reflects that even though the difficulty-dependent corruption introduced noises to the objective truths to form the actual responses, SDR was still able to recover the right number of underlying group of workers. From Table 6.3, *SDR with 2 preferences* clearly outperforms *itself with 1 preference* across all the partially subjective tasks. This means multiple groups of workers have emerged due to the sufficient subjectivity of these tasks. The table also shows that further increasing the number of latent preferences to 3 no longer improves the performance. This was most likely caused by over-fitting, and also suggests a two-dimensional latent space is accurate enough to explain the worker clustering effects emerged from these tasks. We now show the density of the workers' latent preference probabilities $\hat{\phi}_i$ estimated by SDR from the *Image* data. We found our results were consistent with those from Tian and Zhu [111]. We show two of the *Image* sub-tasks in Figure 6.3 to illustrate this consistency. According to Tian and Zhu [111], the sub-task of judging whether images are beautiful is more subjective than the sub-task of identifying skies in images. This is re-confirmed by SDR with the inferred number of worker clusters for the former sub-task greater than that for the latter as shown in Figures 6.3a and 6.3b.



(a) (b)

Figure 6.3: (a) shows the 3 worker clusters on identifying sky from images and (b) shows the 4 worker clusters on judging beautiful images.

Dataset	Question true answer prediction					
	SDR	MV	GLAD	DS	CDS	MdWC
Rel1	**0.4998**	0.4522	0.4457	0.4309	0.4697	0.4674
Rel2	**0.4752**	0.4544	0.4567	0.4512	0.4604	0.4586
Fashion	**0.8733**	0.8580	0.8689	0.8415	0.8463	0.8700
Face	**0.6423**	0.6404	0.6130	0.5924	0.5986	0.6079
Adult	**0.7598**	0.7568	0.7587	0.7534	0.7582	0.7556

Table 6.4: Accuracy of all the models on predicting the true answers of the four partially subjective datasets (the results for the *Image* task are not included as the number of items in this task is too small to show any difference in the performance of different models).

6.4.4 Question True Answer Prediction

To verify the ability of the SDR model to predict the question true answers, we compare it with the Majority Vote (MV) and several frequently applied quality control methods including GLAD, *Multi-dimensional Wisdom of Crowds* (MdWC) [25], *Dawid&Skene* (DS) [56] and its variant *Community* DS (CDS) [73]. All of these methods assume that each question has a single correct answer. Among them, the MdWC model is a new baseline introduced in this chapter. This model endows both crowd-workers and questions with multi-dimensional latent factors, and provides the workers with additional variables to account for their response biases. The performance measure *true answer prediction accuracy* is calculated as: $\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\mathbb{1}\{l_j=\hat{l}_j\}$, where $\hat{l}_j$ is inferred from the respective baselines. For SDR, it is: $\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\mathbb{1}\{l_j=\hat{l}_{cj}\}$, where $c=\arg\max_{c\in\mathcal{C}} N_c$ with N_c the number of workers assigned to cluster c after Elbow K-means, and $\hat{l}_{cj}$ calculated by equation 6.12. The hyper-parameters for each baseline except MV are optimised using the held-out validation specified in section 6.4.2 on the exact same random held-out validation subsets of each dataset in Table 6.1.

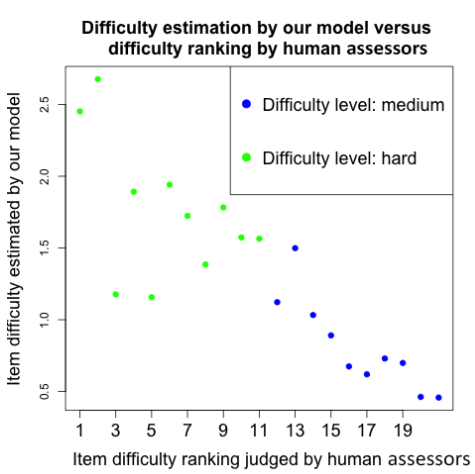
Dataset	Unseen worker response prediction				
	SDR	GLAD	DS	CDS	MdWC
Beauty 1	**0.6974**	0.6884	0.6256	0.6927	0.6912
Beauty 2	0.7006	**0.7011**	0.6796	0.6842	0.6998
Sky	**0.9014**	0.8772	0.8801	0.8862	0.8903
Building	0.8987	0.8912	0.8956	**0.9006**	0.8976
Computer	0.8284	0.8139	0.8115	0.8196	**0.8336**
Rel1	**0.4067**	0.4035	0.3654	0.3972	0.3987
Rel2	**0.4386**	0.4312	0.4257	0.4304	0.4340
Fashion	**0.7659**	0.7593	0.6977	0.7621	0.7633
Face	**0.7224**	0.7193	0.6625	0.7081	0.7148
Adult	**0.7386**	0.7347	0.6767	0.7312	0.7354

Table 6.5: Average accuracy of all the models on predicting the unseen held-out test response of each worker across all the partially subjective datasets.

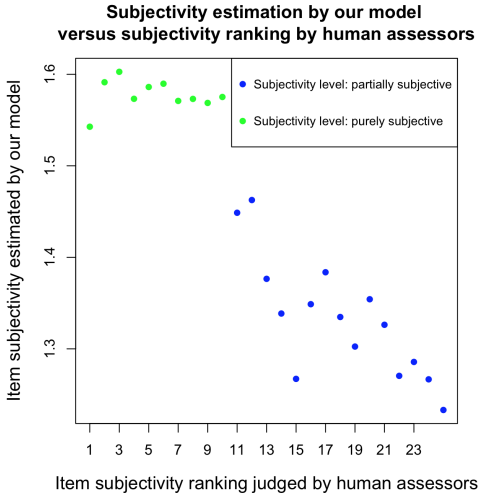
The results of the question true answer prediction are listed in Table 6.4. Across all the partially subjective datasets except the Image data, the SDR model, based on the *largest-group* strategy for choosing the best worker clusters, is superior to the other 5 baselines. Especially, for the tasks of relevance judgement 1&2 and fashion judgement, SDR is able to outperform the best baselines by 3%, 1.5% and 0.3% with almost 54, 9 and 13 more correctly predicted question answers respectively. Since SDR reduces to a model being very similar to GLAD when dealing with objective datasets, it has achieved very similar results to GLAD in true answer prediction over all the objective datasets except for the Duck data. In this task, SDR is superior to GLAD (0.69 versus 0.62 from GLAD) and very close to the accuracy achieved by MdWC. This suggests that SDR is at least as robust as GLAD when predicting true answers for objective tasks.

6.4.5 Worker Response Prediction

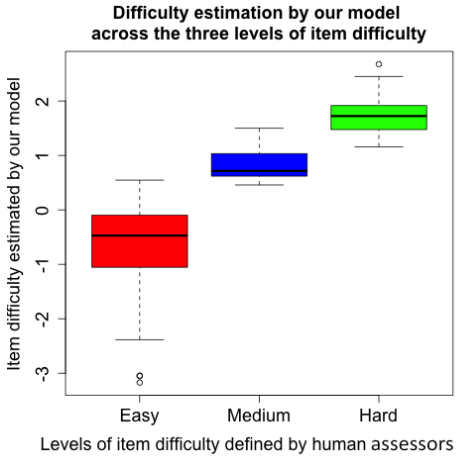
Predicting the individual crowd-worker responses to previously unseen questions is a much more significant and common use-case for (partially) subjective crowdsourcing tasks than it is for the objective tasks. In this experiment, we evaluate the performance of all the models except MV on predicting the *next response* from each worker. We first sample one response from each worker to create a *held-out test* dataset, and then learn all the models from the rest of the data with their hyper-parameters optimised as described in section 6.4.2 using the exact same random validation data subsets. Finally, we evaluate the prediction performance of the models on the held-out test data with the evaluation function being: (1 - MAE). Due to a limited computing budget, in this experiment, we reduce the number of *held-out validation* iterations for each model to be 15 before



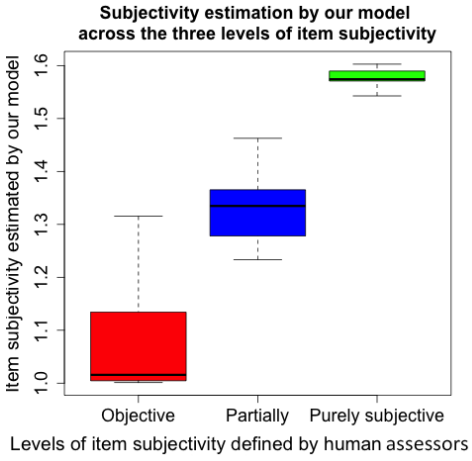
(a)



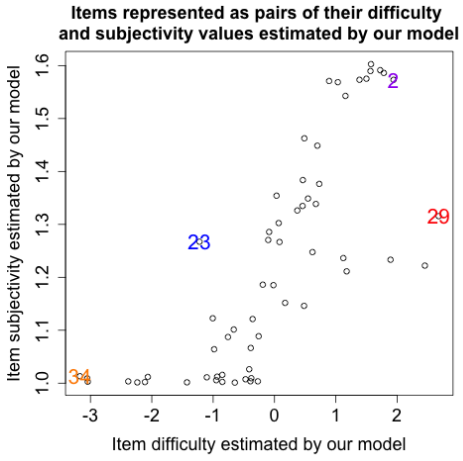
(b)



(c)



(d)



(e)

(f)

Figure 6.4: (a) and (b) show the correlations of both the difficulty and the subjectivity estimates with the corresponding rankings judged by human assessors, while (c) and (d) show the correlations respectively with the levels to which the images were categorized by the assessors. (e) shows the images as points with coordinates being the difficulty and the subjectivity estimates, and highlights some points, while (f) shows their corresponding images.

a single iteration of *held-out test* is conducted. We performed 15 such random tests before the average performance of each model was elicited.

The results of the worker response prediction are shown in Table 6.5. We can see that SDR is not the best on 3 out of the 10 partially subjective datasets, topped by different baselines. Despite that, SDR has still performed adequately well (being second best on those datasets). We conjecture that this is because all these 3 datasets have binary answer options which intrinsically constrain the response behaviour of workers. This results in overall weaker correlations both in the worker responses and in the underlying correct responses across the questions. For the other 7 datasets, 5 of them have more than two answer options per question, and thus contain stronger correct response correlations for SDR to exploit to achieve better performance. To see if the difference in the prediction accuracy between any two algorithms is significant, we conducted the Nemenyi post-hoc test [199], with its parameter $\alpha = 0.1$, over performance ranks of models derived from Table 6.5. The result reveals that the performance difference between SDR and either CDS, GLAD or DS is statistically significant.

6.4.6 Subjectivity and Difficulty Estimate Consistency with Human Assessment

In this experiment, we investigate whether the estimates of the difficulty and the subjectivity[6] of questions derived from the SDR model are *consistent* with the judgements of five human assessors. We focused on the object identification & image beauty task[7] from Tian and Zhu [111] as the total number of its questions is 60, a manageable workload for the assessors to provide high-quality judgements. The assessors are either PhD or Master students, three of whom are avid photographers with (what we believe to be) adequate knowledge about what constitutes beautiful images, while the other two are novices who, during the group discussion, provided suggestions as to how novices might react to different images. We asked them to rank the images with respect to (i) difficulty and (ii) subjectivity. The respective instructions we gave to them were:*"rank all these images by how hard they are for crowd-workers to judge correctly by avoiding possible incorrect answers"* and *"rank them this time by how subjective they are for crowd-workers to judge"*. The assessors first developed their two rankings independently of one another and without any time limit. During this process, they could revise their rankings as many times as they wished until they felt confident

[6]In this experiment, we also ran the Algorithm 1 across all the partially subjective datasets to test its average convergence time for the subjectivity estimation of each question. The result is 0.6 second (on a 16 GB 2.6 GHZ dual-core laptop). Algorithm 1 spent the longest total running time (33 minutes) on the largest dataset "Fashion" which contains 3,837 questions.

[7]Crowd-workers are asked whether an image is beautiful or not.

to submit. The assessors then worked together to merge their rankings into single rankings (for both difficulty and subjectivity) through group discussion and majority vote. The resulting rankings were then compared with the corresponding rankings based on the estimates from the learned SDR model. The assessors were also asked to categorise each image into one of the three levels of difficulty (i.e. *easy*, *medium*, and *hard*), and into one of the three levels of subjectivity (i.e. *objective*, *partially subjective*, and *purely subjective*). We did this to see whether there existed any correlation between the difficulty or subjectivity levels to which images were categorised, and their corresponding estimates from SDR.

The results of the subjectivity and difficulty coherence evaluation have been summarised in Figure 6.4. Figures 6.4a and 6.4b show overall there is a strong negative correlation between the model estimates and the rankings judged by human assessors. The larger the estimate for either the difficulty or the subjectivity of an image, the higher it tends to be ranked by human assessors. Figures 6.4c and 6.4d show that there exist clear positive correlations between the levels of difficulty and subjectivity into which the images get categorised by the human assessors, and the estimated values of these two properties inferred by SDR. To support our argument regarding the efficacy of SDR in revealing the two key properties of images, we have selected four images highlighted in different colours in Figure 6.4e with their image ids. We can see that image 34 is inferred by our model to be both easy and objective as both of its estimates shown in Figure 6.4e are small. This can be re-confirmed by visual inspection of the image in Figure 6.4f. It is easy to see that there is no sky in the image 34. Image 29 has been identified by SDR to be hard with low subjectivity according to its estimates shown in Figure 6.4e. This is reasonable as the image indeed contains an extraterrestrial sky which is hard for novice workers to realise, while expert workers are able to realise and find the image objective. Images 2 and 23 both belong to the image beauty judgement task which requires workers to select 6 most beautiful images from 12 images. Our model has identified that image 2 is more subjective and harder to judge. This is probably because image 2 delivers a view of scenery which is more likely to resonate with certain subgroups of workers while image 23 is merely showing an object. As a result, workers tend to show more varied feelings and opinions towards image 2. On the other hand, image 23 does have better image quality and thus is easier for workers to make their decisions on whether it is beautiful or not.

6.5 Conclusion

In this chapter, we have proposed the SDR (Subjectivity-and-Difficulty Response) model, a novel quality-control framework for crowdsourcing that is able to jointly model and *distinguish* question *subjectivity*, which causes *worker-specific truth* for individual questions, from question *difficulty*, which determines the probability that a worker's response to each question equals her perceived subjective truth.

Experiment results show that our model improves both the correct answer prediction for questions and the unseen held-out response prediction for crowd-workers compared to five baselines across numerous partially subjective crowdsourcing datasets. Moreover, our model shows robustness to both objective and partially subjective datasets by discovering the right numbers of underlying worker groups for each. Finally, our model is able to provide estimates of the difficulty and the subjectivity of questions that are consistent with the judgements from human assessors.

With respect to the research question *RQ*3 and its sub-questions *RQ*3.1 to *RQ*3.4, we have the following answers:

- **Answer to *RQ*3.1**: The subjectivity controls the number of correct answers for a question. It gives rise to the interaction between workers' *personal preferences* and each response option's *preference/likeability factors*. This interaction results in subjective truths for the same question that are preferred by different groups of workers.

  The difficulty controls the extent to which workers are confused (or tricked) by a question to give random responses or responses different from the subjective truths they perceived for the question. Thus question difficulty counteracts workers' abilities to provide their subjective truths in their responses.

- **Answer to *RQ*3.2**: The difficulty-dependent interaction that generates the actual responses of the workers depends on the subjectivity-dependent interaction that generates the subjective-truths perceived by the workers. From the generative process point of view, the former interaction can be viewed as a *noise and bias addition* process conditioned on the subjective truth signals generated from the latter interaction.

- **Answer to *RQ*3.3**: Our proposed model encodes the above dependency with explicit variables separate for question difficulty and worker preferences. A grouping structure is enforced over the worker preferences by an *LDA-based sparsity-inducing prior* that allows the model

to *implicitly learn the subjectivity of each question*. Empirical results have shown that our model overall improves both true answer prediction and unseen held-out response prediction on partially subjective questions.

- **Answer to *RQ*3.4**: With the LDA-based structure, our model is able to *group* worker preferences while at the same time inferring these preferences from the responses. The group-specific preferences are then used to effectively compute the *subjectivity estimate* of each question based on a Monte Carlo simulation procedure we proposed. In this case, the grouping efficacy of our model directly affects the subjectivity estimation. Meanwhile, our model also directly estimates the difficulty of each question. The efficacy of both estimates has been verified by their agreement with human assessments.

- **Answer to *RQ*3**: Separately modelling and estimating question difficulty and subjectivity can improve the quality control performance in terms of true answer prediction for partially subjective questions. This is achieved based on the assumption that the subjectivity induces groups of workers with different subjective-truth preferences and the difficulty confuses the workers to make them deviate from their subjective-truths. Our model effectively encodes this assumption using a difficulty-expertise dependent logistic regression model that is combined with a group-inducing LDA mechanism over a latent preference space.

The following future work warrants investigation:

- It might be possible to improve the efficiency of Algorithm 1 for faster convergence and better scalability of subjectivity estimation for datasets with very large numbers of questions.

- Instead of applying the K-means algorithm with the Elbow method to the estimated preferences of workers, an alternative would be to integrate this clustering process with the SDR model into some non-parametric Bayesian model (e.g. a Dirichlet process mixture model). This model would allow techniques such as Chinese restaurant process to automatically infer the number of preference clusters among workers and then the subjectivity of each question while also learning their difficulty.

- In the SDR model, although the difficulty-dependent logistic regression depends on the preference-modelling LDA via latent subjective truths, the difficulty and the preferences are treated as independent variables. One could allow for dependency for these two variables by making the priors of the difficulty parameter dependent on the size of the set of preferences

$\mathcal{M}$. It would be interesting to see whether such a subjectivity-dependent difficulty model could further improve the quality control performance in crowd-sourcing (thereby indicating that such a dependency does indeed exist).

Chapter 7

Leveraging Response Semantic Relationships for Improved Quality Control in Multi-class Crowdsourcing

In this chapter, we answer the research question *RQ*4 with all its sub-questions (i.e. *RQ*4.1 to *RQ*4.4) by investigating how to effectively model and learn the joint effects of *worker ability*, *question difficulty* and *semantic relationships between response options* on response quality. The answers to these questions will shed light on how to make use of the underlying semantic relationships (as well as any external knowledge about them) to improve the quality control performance in multi-class crowdsourcing.

According to our literature review in Chapter 2, current quality control methods for crowdsourcing largely account for variations in responses through modeling the interactions between *worker expertise* and *question difficulty*. These methods have been reported to achieve overall superior performance over the conventional majority vote approach. However, assuming individual crowdsourcing tasks contain *small numbers of uncorrelated response options*, these methods inevitably ignore the impact that the *semantic relationships* between response options have on the quality of workers' responses.

In practice, it is not unusual for crowdsourcing tasks to involve response options/categories [1] that are correlated with one another in terms of *large structural semantic relationships*. A typical example of such a crowdsourcing task is the classification of objects in images (to build datasets

[1] In this chapter, we use the terms (response) "category" and (response) "option" interchangeably.

like ImageNet)[2] where a large number of answer categories are related through hierarchical class relationships described in semantic resources such as the WordNet[3]. Other examples include the classification of Webpages for the Open Directory Project (called DMOZ)[4], or for DBpedia[5], where the large numbers of categories are connected through semantic relationships maintained by their respective online volunteer communities.

This chapter focuses on leveraging semantic relationships between response options for improving quality control performance in crowdsourcing problems, especially those involving *highly multi-class* labels. Semantic relationships are inherent in highly multi-class labeling problems, and their identification and evaluation traditionally relies on human knowledge and assessment which features prominently in crowdsourcing. Conversely, knowing the semantic relationships between response options should contribute to accurate inference about how responses are formed in highly multi-class crowdsourcing.

When semantic relationships between response options exist in crowdsourcing, crowd-workers with greater expertise tend to respond to the questions either with the correct answer, or with options more related to the true answer of that question. Moreover, questions with greater difficulty tend to receive responses with options less related to their correct answers.

To be more specific, consider a simple measure of relatedness between two response options k and k' shown below:

$$\text{Relatedness}(k, k') = \frac{1}{|\text{shortest\_path}(k, k')| + 1} \tag{7.1}$$

where $|\text{shortest\_path}(\cdot, \cdot)|$ is the length of the shortest path between any pair of response (label) categories in some known semantic structure (e.g. a class hierarchy). Using this relatedness measure, Figure 7.1 shows the relationship between the *relatedness* of the response category to the true answer, and three *summary statistics* (namely the maximum, mean and minimum values) for the *response accuracy* of workers and the *question difficulty* (in terms of response error). The crowdsourcing task involved in this case is identifying breeds of dogs in images from ImageNet [200].

Every coloured "violin" area in each sub-figure of Figure 7.1 represents the distribution of a particular summary statistic about either the response accuracy of workers or the response errors

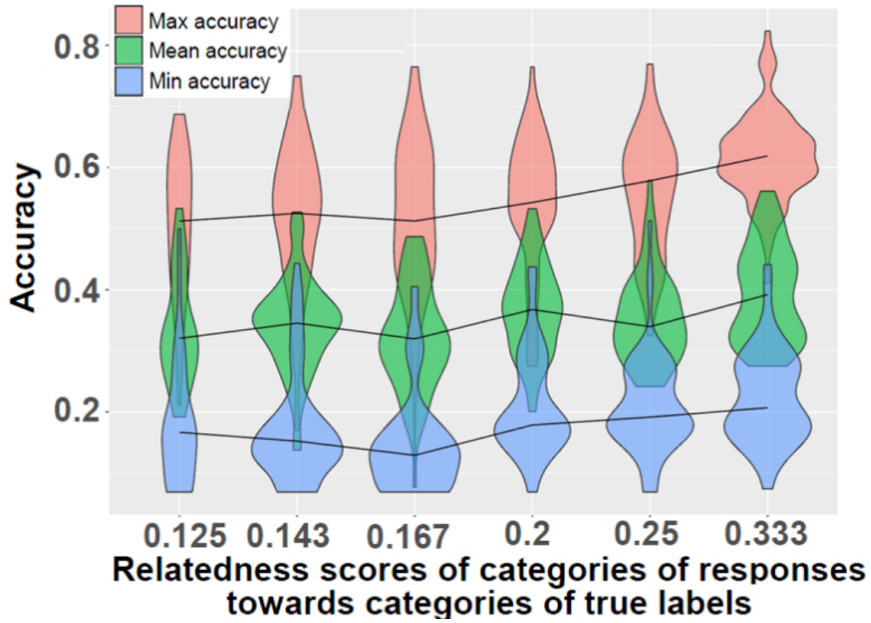
[2] http://www.image-net.org/
[3] https://wordnet.princeton.edu/
[4] http://www.dmoz.org/
[5] http://wiki.dbpedia.org/

for the questions given the true answers. The medians of the areas with the same colours (i.e. the same summary statistics across different relatedness scores) are connected by straight lines in each sub-figure. We observe from Figure 7.1 the following:

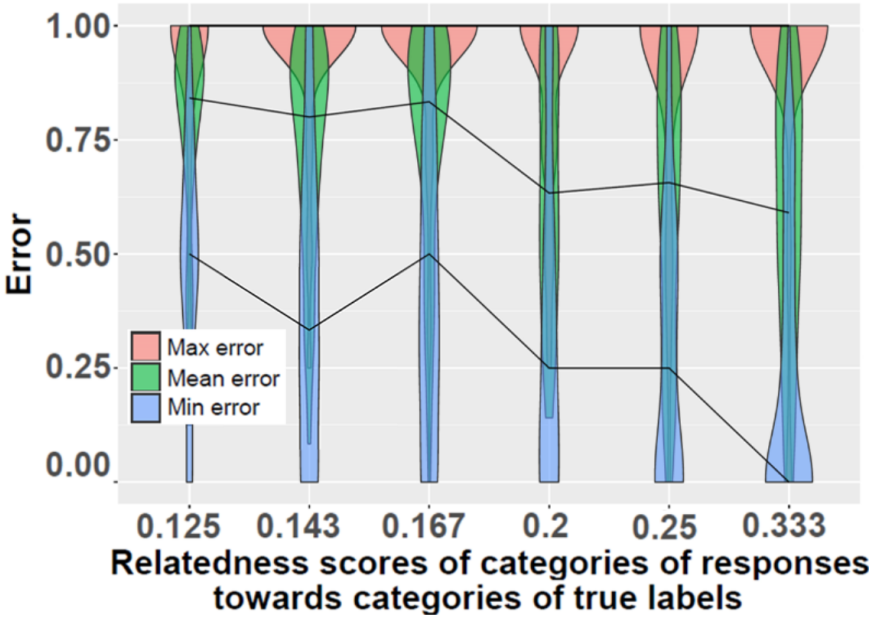
- According to Figure 7.1a, more related response categories (with higher relatedness scores) to the correct answer tend to be chosen more often by workers with higher response accuracy;
- According to Figure 7.1b, less related categories (with lower relatedness scores) to the correct answer tend to be given more often to more difficult questions (i.e. ones with larger response errors in Figure 7.1b).

In this chapter, we leverage the above observed relationship between response category relatedness and worker accuracy/question difficulty for improving the quality control of crowdsourced responses. This is done by encoding the correlations between the categories into the conditional probability of a worker giving a particular response to a question given its true answer. Such an encoding can help to refine the estimation of correctness of crowdsourced responses (which is modelled using the above conditional probabilities in most QCC methods). The encoding is based on a *latent symmetric* relatedness matrix where each off-diagonal entry is a real-valued score representing how related the response categories are to one another. In this case, each correct response category is associated with a continuous scale accommodating the latent relatedness scores of all the other categories as possible worker responses. We also model expertise of workers and difficulty of questions on the same scale. According to Figure 7.1a, a worker with greater expertise and a category more related to the correct answer should have the estimated values for their respective variables reside further down the positive infinite end of the scale once learned from responses. Likewise from Figure 7.1b, a question with greater difficulty and a less related category should have their estimated values situated towards the negative infinite end of the scale. The interactions between these variables on the scale are captured and transformed into the aforementioned conditional probabilities through an *ordered logit* model.

In this model, the difference between question difficulty and worker expertise serves as the *response-specific slope*. The off-diagonal scores in a particular row of the latent relatedness matrix (corresponding to a correct response category) serve as the *intercepts* specific to the response categories. In this case, a slope value larger than the largest intercept on the scale means that the response is more likely to be correct, while a slope falling between smaller intercepts means the response is likely to be a less related category to the correct response category. In addition, the off-diagonal scores in the matrix share the same Normal prior which can incorporate external prior



(a)



(b)

Figure 7.1: (a) Worker response accuracy versus category relatedness. (b) question difficulty (in terms of response error) versus category relatedness.

knowledge (e.g. semantic distances extracted from Wordnet and DMOZ) to better calibrate the estimated scores.

7.1 Related Work

Two papers have considered leveraging relationships between response options [54, 55] for improving quality control of crowdsourced responses. In Han et al. [55], a model called **SEEK** was

proposed in which the conditional probability of any possible response category given by a worker to a question given its true answer is output from a *soft-max* function. The function takes in the observed relatedness scores of all the response categories along with the question's difficulty and the worker's expertise.

Inside the soft-max function, the difference between the difficulty and expertise parameters is multiplied by the relatedness score of the response category in order to compute the corresponding conditional probabilities. Since the difference between the difficulty and expertise parameters is the same for all response categories, the conditional probabilities are thus proportional to the relatedness scores. The larger a score is, the higher the conditional probability of the corresponding response given the true answer. In comparison, our model allows the conditional probabilities to be proportional to the joint interaction between the difficulty-expertise difference and the relatedness scores.

In Han et al. [55], each relatedness score between a pair of response categories can vary from 0 to 1. The score is 1 when the two categories are identical and ranges between 0 and 1 only when one of the categories is a hypernym of the other, otherwise, the score is always 0. Clearly, this method of pre-computing the relatedness scores between response categories constrains the quality control performance of SEEK in crowdsourcing tasks where most of the categories are not hypernyms.

In Fang et al. [54], a model called **DASM** is proposed which share the same idea as SEEK except that the relatedness scores are pre-computed as the inverse of the Euclidean distances between response categories in terms of their observed features. Both of these models rely on the availability of the external knowledge about the response category relatedness, while our model is able to infer such relatedness directly from the responses themselves.

7.2 Problem Formulation

ssume a large but finite set of response options $\mathcal{K}$ and a set of questions $\mathcal{J}$, for which a set of workers $\mathcal{I}$ have provided a set of responses $\mathcal{R}$ to $\mathcal{J}$. A question $j \in \mathcal{J}$ has one unknown true answer $l_j = k$, where $l_j \in \mathcal{L}$, the set of corresponding true answers of individual questions in $\mathcal{J}$, and k is a particular option in $\mathcal{K}$. For the set of options $\mathcal{K}$, there exists a tree structure organizing them in terms of their semantic relationships. The relationships are quantified into an observed real-valued relatedness matrix $\boldsymbol{X} \in \mathbb{R}^{|\mathcal{K}|\times|\mathcal{K}|}$. Each off-diagonal entry $x_{kk'}$ expresses how related an option k' (as a response to a question) is to the true answer k for that question, and is calculated using

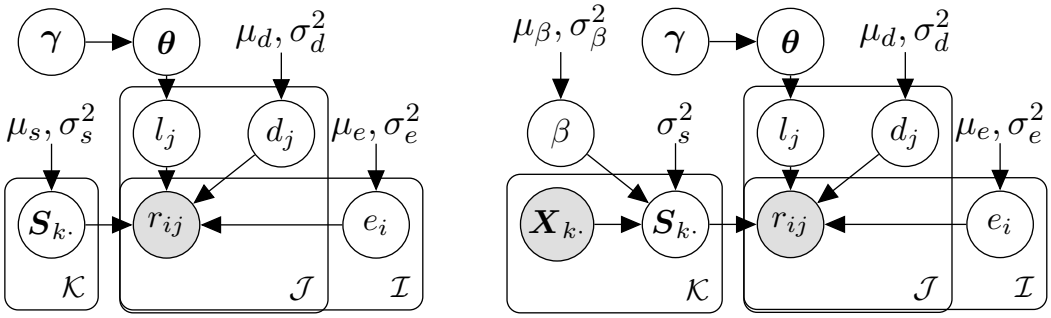


Figure 7.2: The DELRA model with and without encoding observed knowledge matrix $\boldsymbol{X}$ specifying relationships between categories are shown in Figure 7.2a and 7.2b.

equation 7.1. Based on these inputs, our model should output a corresponding set of predictions $\hat{\mathcal{L}}$ for the latent true answers $\mathcal{L}$ such that the overall difference between the former and the latter sets across their corresponding elements is as small as possible.

7.3 Proposed Model

In this chapter, we propose the Difficulty-Expertise-Label-Relationship-Aware (**DELRA**) model, characterized by a latent relatedness matrix $\boldsymbol{S} \in \mathbb{R}^{|\mathcal{K}|\times|\mathcal{K}|}$. The matrix element $s_{kk'}$ specifies how related a response option k' (as a response to a question) is to another option k (as the true answer of that question) in crowdsourcing. We assume that $\boldsymbol{S}$ is symmetric so that $s_{kk'} = s_{k'k}$ where $s_{kk'}, s_{k'k} \in \boldsymbol{S}$. This assumption is reasonable since if crowd-workers perceive option k' to be related to category k overall to a certain degree, they should also perceive the relatedness of category k to category k' to the same degree. Based on the assumption, the DELRA model is shown in Figure 7.2a, and has the following generative process:

1. Draw true answer category proportions $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\gamma})$;

2. For each pair of categories (k,k') where $k \neq k'$:

   (a) if $k' > k$ then draw relatedness $s_{kk'} \sim \text{Normal}(\mu_s, \sigma_s^2)$;
   else $s_{kk'} \leftarrow s_{k'k}$[6];

3. For each question $i \in \mathcal{J}$:

   (a) Draw its true answer $l_j \sim \text{Categorical}(\boldsymbol{\theta})$;

   (b) Draw its difficulty $d_j \sim \text{Normal}(\mu_d, \sigma_d)$;

[6]The expression "$a \leftarrow b$" stands for assigning b to a or equivalently replacing a with b.

4. For each worker $i \in \mathcal{I}$:

   (a) Draw her expertise $e_i \sim \text{Normal}(\mu_e, \sigma_e^2)$;

5. For each worker-question pair (i, j):

   (a) Draw response $r_{ij} \sim \text{Categorical}(\boldsymbol{\psi}_{ijl_j})$ where $\boldsymbol{\psi}_{ijl_j}$ is a $|\mathcal{K}|$-dimensional vector with each element $\psi_{ijl_jk} = P(r_{ij} = k|l_j)$ specified as the difference between consecutive sigmoid functions as follows:

$$\psi_{ijl_jk} = f_{ijl_jk} - \max_{z:f<f_{ijl_jk}} f_{ijl_jz} \quad \text{where } f_{ijl_jl_j} = 1, f_{ijl_j0} = 0 \tag{7.2}$$

   Here f_{ijl_jk} is a sigmoid function relating the odds of observing response $r_{ij} = k$ given true answer l_j to a linear combination of the relatedness score s_{l_jk}, the worker expertise e_i and the question difficulty d_j:

$$f_{ijl_jk} = \frac{1}{1 + \exp(-(s_{l_jk} - e_i + d_j))} \tag{7.3}$$

From equation 7.3, the difference between the term (e_i - d_j) and the relatedness score variables determine the conditional probabilities of different categories selected as the response r_{ij} given the true answer l_j. Table 7.2 shows two examples of how the conditional probability $P(r_{ij} = k|l_j)$ specified by equation 7.2 varies according to different values for (e_i - d_j) and the relatedness score variables.

In example 1, suppose that response categories k' and k reside consecutively on the relatedness scale specific to the true answer category l_j with their respective relatedness scores being 0.0 and 0.5. The value of (e_i - d_j) resides within the interval bounded by the two scores. When $l_j \neq k$, and (e_i - d_j) decreases from 4.5 to 0.25, the conditional probability $P(r_{ij} = k|l_j \neq k)$ increases from 0.007 to 0.124. In other words, when a worker is less expert or a question is more difficult (or both at the same time), a less related category k is more likely to be selected.

In example 2, suppose that k'' is the most related category to the true answer category l_j and the value of (e_i - d_j) is always 4.5. When $l_j = k$, and the relatedness score $s_{l_jk''}$ of category k'' to category k increases from 0.0 to 5.0, the conditional probability $P(r_{ij} = k|l_j = k)$ decreases from 0.989 to 0.378. This means that responses given by workers to a question are less likely to be correct when the category of the question's true answer has very related or similar categories that intrinsically bias the workers.

Example 1			$P(r_{ij} = k \mid l_j \neq k) =$
$s_{l_j k}$	$s_{l_j k'}$	$e_i - d_j$	$f_{ijs_{l_j k}} - f_{ijs_{l_j k'}}$
0.5	**0.0**	**4.5**	**0.007**
0.5	**0.0**	**0.25**	**0.124**

Table 7.1: Example 1 of equation 7.2

Example 2		$P(r_{ij} = k \mid l_j = k) =$
$s_{l_j k''}$	$e_i - d_j$	$1 - f_{ijkk''}$ where $k'' = \text{argmax}_{z: f < f_{ijkk}} f_{ijkz}$
0.0	**4.5**	**0.989**
5.0	**4.5**	**0.378**

Table 7.2: Example 2 of equation 7.2

Apart from inferring the relatedness matrix $\boldsymbol{S}$ from responses, our model also allows for the encoding of useful prior knowledge about the entries in each row of the matrix corresponding to a particular category as the true answer to help calibrate the inference. In Figure 7.2b, the Normal prior Normal(μ_s, σ_s^2) in Figure 7.2a shared by all the entries of the matrix $\boldsymbol{S}$ is now replaced by individual Normal priors. These priors are centered on the product results between a global coefficient β and the off-diagonal elements in the observed matrix $\boldsymbol{X}$ after it is *log-transformed* followed by *standardization*. The priors share the same variance σ_s^2. Correspondingly, step 2(a) of the above generative process of the DELRA model is now changed to:

2. For each pair of categories (k,k') where $k \neq k'$:

   (a) if $k' > k$ then draw $s_{kk'} \sim$ Normal$(\beta x_{kk'}, \sigma_s^2)$;

   else $s_{kk'} \leftarrow s_{k'k}$;

The global term $\beta \sim$ Normal$(\mu_\beta, \sigma_\beta^2)$. The term $x_{kk'}$ goes through the transformation:

$$x_{kk'} \leftarrow \frac{\log(x_{kk'}) - \hat{\mu}_{\log(\boldsymbol{X})}}{\hat{\sigma}_{\log(\boldsymbol{X})}} \tag{7.4}$$

where $\hat{\mu}_{\log(\boldsymbol{X})}$ and $\hat{\sigma}_{\log(\boldsymbol{X})}$ are respectively the sample mean and the sample standard deviation of the logarithm of all the original terms in $\boldsymbol{X}$. The reason for the introduction of the logarithmic transformation is that the outputs from the relatedness function specified by equation 7.1 are highly skewed and we do not want this skew to impact the estimation of the relatedness matrix. The reason for the standardization operation is that every log-transformed $x_{kk'}$ is negative, thus having a negative mean. We want to adjust them to be centered on zero with scale one to allow for easier setup of priors for other model parameters. After the transformation using equation 7.4, the prior

mean $\beta x_{kk'}$ for the score $s_{kk'}$ suggests how the relatedness between options k and k' according to the external knowledge correlates with their latent relatedness in crowdsourcing a priori.

7.4 Parameter Estimation

In this section, we describe how the model parameters are estimated. More specifically, in each iteration of the estimation procedure, we alternate between: (i) *Collapsed Gibbs sampling* for inferring the true answers for questions $\mathcal{L}$ given the current estimates of the other model parameters including the worker expertise e_i, the question difficulty d_j and the relatedness matrix $\boldsymbol{S}$, and (ii) gradient descent using the *LBFGS-B* for updating these parameters given the current assignment of $\mathcal{L}$.

Collapsed Gibbs Sampling for $\mathcal{L}$: At this stage, we obtain posterior samples for question true answers $\mathcal{L}$ given the current estimates of all the other parameters. The conditional probabilities of true answer l_j for question j is obtained by marginalizing out the multinomial probability vector $\boldsymbol{\theta}$, which ends up being:

$$P(l_j = k|\mathcal{L}_{\setminus j}, \mathcal{R}_j, \{e_i\}_{i\in\mathcal{I}_j}, d_j, \boldsymbol{s}_k, \boldsymbol{\gamma}) \propto \frac{N_{\setminus jk} + \gamma_k}{\sum\limits_{z\in\mathcal{K}}(N_{\setminus jz} + \gamma_z)} \prod_{i\in\mathcal{I}_j} \psi_{ijkr_{ij}} \tag{7.5}$$

where $\mathcal{I}_j$ is the set of workers who responded question j with a set of responses $\mathcal{R}_j$, $\mathcal{L}_{\setminus j}$ is the set of current true answer assignments to all the questions except j, and $N_{\setminus jk}$ is the number of questions except j whose true answers are now predicted to be k.

Gradient Descent for Other Parameters: The conditional probability distributions of the other model parameters including e_i, d_j, and S are hard to compute analytically due to the presence of the sigmoid function. Instead, we run the LBFGS-B until convergence of the following objective function Q:

$$Q = -\log\left(p(\{e_i\}_{i\in\mathcal{I}}, \{d_j\}_{j\in\mathcal{J}}, \boldsymbol{S}|\mathcal{R}, \mathcal{L}, \mu_{\{e,d,s\}}, \sigma^2_{\{e,d,s\}})\right) = -\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}_j}\log(\psi_{ijl_jr_{ij}}) + \frac{1}{2}\left[\sum_{i\in\mathcal{I}}\frac{(e_i-\mu_e)^2}{\sigma_e^2} + \sum_{j\in\mathcal{J}}\frac{(d_j-\mu_d)^2}{\sigma_d^2} + \sum_{k\in\mathcal{K}}\sum_{k'\in\mathcal{K}\&k'>k}\frac{(s_{kk'}-\mu_s)^2}{\sigma_s^2}\right] \tag{7.6}$$

The complicated part of the gradient calculation[7] lies in the partial derivative of $\mathcal{Q}$ with respect to the category-relatedness score $s_{kk'}$, which is computed as:

$$\frac{\partial Q}{\partial s_{kk'}} = -\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}_j}\frac{\partial \log(\psi_{ijl_jr_{ij}})}{\partial s_{kk'}} + \frac{s_{kk'}}{\sigma_s^2} \tag{7.7}$$

Where for the true answer $l_j = k$ and observed response $r_{ij} = k'$ we have:

$$\frac{\partial \log(\psi_{ijl_jr_{ij}})}{\partial s_{kk'}} = \frac{f_{ijkk'}(1 - f_{ijkk'})}{\psi_{ijkk'}} \tag{7.8}$$

And for other responses $r_{ij} \neq k'$ we have:

$$\frac{\partial \log(\psi_{ijl_jr_{ij}})}{\partial s_{kk'}} = \begin{cases} \frac{-f_{ijkk'}(1-f_{ijkk'})}{\psi_{ijkr_{ij}}}, & \text{If } k' = \text{argmax}_{z:f<f_{ijkr_{ij}}} f_{ijkz} \\ 0, & \text{Otherwise} \end{cases} \tag{7.9}$$

Note that we also impose symmetry on the category relatedness terms $s_{kk'} = s_{k'k}$.

When observed matrix $\boldsymbol{X}$ is introduced into the model, the coefficient β is updated by maximum a posteriori estimation for a linear regression over $\boldsymbol{X}$ as follows:

$$\beta = \frac{\mu_\beta/\sigma_\beta^2 + \text{vec}(\boldsymbol{X})^T\text{vec}(\boldsymbol{S})}{1/\sigma_\beta^2 + \text{vec}(\boldsymbol{X})^T\text{vec}(\boldsymbol{X})} \tag{7.10}$$

where $\text{vec}(\cdot)$ is the vectorization function for matrices.

7.5 Experiments and Results

In this section, we present experiments that evaluate the performance of our model with and without leveraging external knowledge regarding semantic relationships between response options for improving the quality control of crowd-sourced responses.

7.5.1 Datasets

We have collected four new crowd-sourcing datasets from CrowdFlower for our experiments. Table 7.3 summarizes these datasets.

[7] Since the gradient calculation for e_i and d_j is very straightforward, we skip it without affecting the reading of the rest of the chapter. More importantly, without dealing with more equations, we can present equations 7.7, 7.8 and 7.9 succinctly.

Dataset	$\|\mathcal{I}\|$	$\|\mathcal{J}\|$	$\|\mathcal{K}\|$	$\|\mathcal{R}\|$
Dog	136	1,200	120	6,000
Bird	428	707	72	5,660
Instrument	334	1,323	193	7,233
Movie	169	737	148	3,539

Table 7.3: Dataset Summary. The headers correspond to the notation introduced in Table 1.1.

- **Dog Breed Identification** [200] (**Dog**). The images and the set of categories used in this task originate from the Stanford Dog dataset [200]. There are 120 breeds of dogs involved in the task with 10 images for each dog breed randomly sampled from the Stanford dataset. There are 5 answers collected for each image regarding the breed that the crowd-workers think appears in that image. The 120 breeds are organized under the subtree "Dog" in WordNet.

- **Bird Species Identification** [201] (**Bird**). The categories involved are species of birds from the Caltech-UCSD Birds 200 dataset [201]. Originally, there were 200 bird species in this dataset, but only 72 of them were present in the WordNet. As a result, we have only used these categories for the experiments and randomly sampled 10 images for each of them from the Caltech-UCSD dataset. Since this task is quite hard overall, we collect on average 8 answers for each of the images.

- **Classification of Webpages about String Instruments** (**Instrument**). This task asks for judgements regarding the sub-directories under which Webpages about String instruments should be placed. All the sub-directories share one root directory "Arts/Music/Instruments/String Instruments" from DMOZ. We have collected 5 judgements for each of the 1,323 Webpages across the 193 sub-directories corresponding to different aspects of String instruments.

- **Classification of Webpages about Movies** (**Movie**). The judgements collected in this case are about the sub-directories under which Webpages about various movies should be placed. Also being part of the DMOZ directories, all the sub-directories involved share the root directory "Arts/Movies". We have collected 5 judgements for each of the 737 Webpages across the 148 sub-directories corresponding to different aspects of movies.

7.5.2 True Answer Prediction

To verify the capability of our model on predicting question true answers, we compare it with several state-of-the-art crowdsourcing quality control methods. They include the majority vote

Methods	Datasets			
	Dog	Bird	Instrument	Movie
DELRA	0.4803	0.4278	0.4489	0.3367
DELRA+X	**0.4833**	**0.4331**	**0.4561**	**0.3433**
SEEK	0.4688	0.4046	0.4406	0.3217
SEEK+X	0.4752	0.4256	0.4453	0.3342
DASM	0.4720	0.4229	0.4448	0.3274
MV	0.4742	0.4170	0.4414	0.3256
GLAD	0.4675	0.4017	0.4450	0.3229
DS	0.4341	0.3219	0.3900	0.2931
MdWC	0.4742	0.4041	0.4409	0.3311
PM	0.4367	0.3621	0.4002	0.2999
Minimax	0.4770	0.4224	0.4456	0.3202

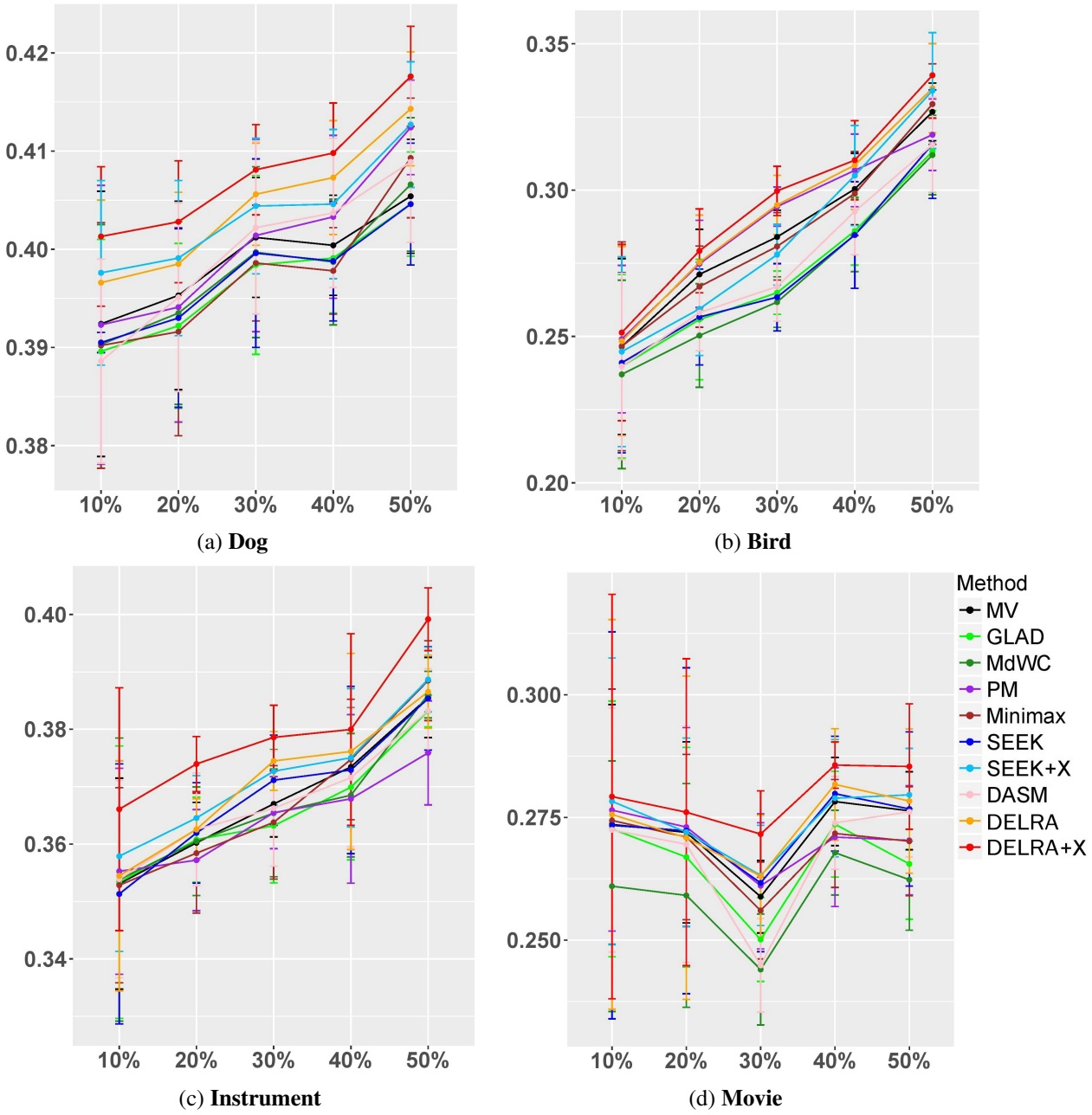
Table 7.4: The accuracy of different models on inferring the true (answer) responses of the (question) items across the four datasets.

(MV), GLAD, MdWC and DS models which were used as baselines in the previous chapters, and also the following two new baselines:

- **Minimax entropy (Minimax)**[89]. In this model, the conditional probabilities for every option with which a worker can respond given each true answer are estimated. In this case, the total entropy of the conditional probabilities over all the categories as the responses to the questions given their true answers is optimized according to the minimax principle with constraints.

- **Participant-Mine voting** (**PM**)[76]. The ability/accuracy of each worker and the true answer of each question are inferred together using the HITS [86] algorithm. Each question is treated as a Webpage for HITS with the total accuracy of the workers responding to it as its authority level and the total difference between the true answer estimate of the question and its received worker answers as its hub level.

Apart from these baselines, we also compare our model with the original SEEK model discussed in section 7.1. Moreover, we also experiment with changing the external knowledge matrix input to the SEEK model to be the same matrix $\boldsymbol{X}$ input to the DELRA model (called **DELRA+$\boldsymbol{X}$**) with each entry transformed by equation 7.4 in both cases. We call this model **SEEK+$\boldsymbol{X}$** and use it as another baseline. Likewise, we adapt the DASM model by calculating the distance between any pair of response categories using our distance definition specified in equation 7.1 rather than theirs since we do not have any observed feature about response categories. To measure the performance of our model and all the baselines, we use the *true answer prediction accuracy*, defined as $\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\mathbb{1}\{l_j = \hat{l}_j\}$.



(a) **Dog** (b) **Bird**

(c) **Instrument** (d) **Movie**

Figure 7.3: The accuracy of different models on inferring the true answers of the items from 10% to 50% of the total responses across the four datasets. Note that x-axis and y-axis in each figure are respectively the sampling proportions of responses and the average true answer prediction accuracy over 10 runs.

Table 7.4 shows the results of the true answer prediction of both DELRA and all the baselines. We can see that *with* and *without* the knowledge matrix $\boldsymbol{X}$ incorporated, the DELRA model respectively outperforms all the baselines by at least 0.6% and 0.3% over the Dog dataset, 0.8% and 0.22% over the Bird dataset, 1.1% and 0.36% over the Instrument dataset and 0.9% and 0.25% over the Movie dataset. Especially, SEEK+$\boldsymbol{X}$ has the exact same knowledge matrix input as DELRA+$\boldsymbol{X}$, but has yielded lower performance even compared to the DELRA model without

incorporating $\boldsymbol{X}$. This suggests that not only our model is able to better leverage the external knowledge about semantic relationships between response categories, but also it is a better model in explaining how responses are generated from the interactions among the expertise of workers, the difficulty of items and the relationships between response categories in crowdsourcing.

7.5.3 True Answer Prediction Under Response Sparsity

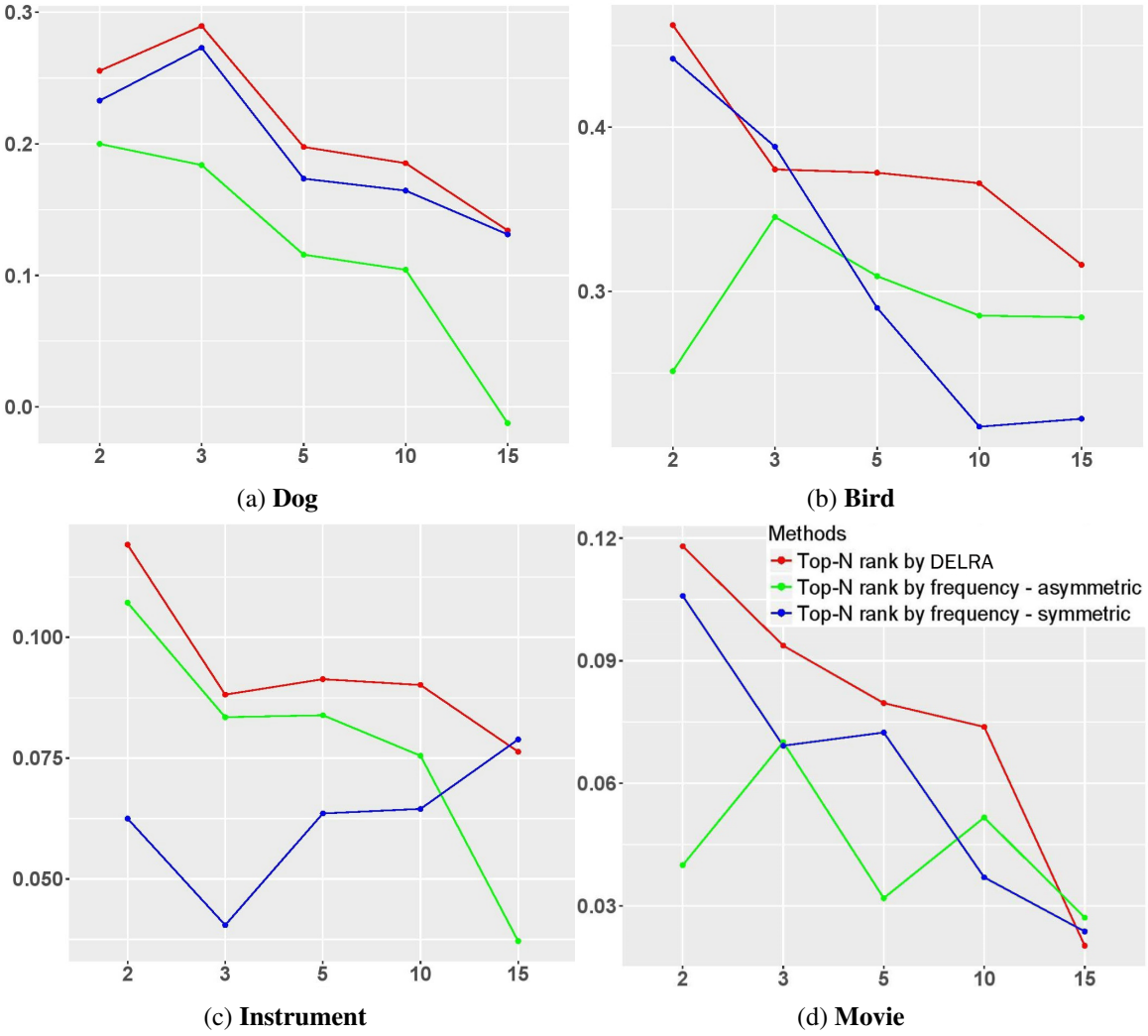
We now proceed to investigate how DELRA performs under various degrees of sparsity in crowdsourced responses. To do this, we randomly sample different proportions (between 10% and 50%) of the responses from each of the datasets and average the performance over 10 runs for each model (on each proportion). Figures 7.3a to 7.3d show the results of the true answer prediction of all the models under varying degrees of response sparsity across the four datasets. The DELRA model incorporating the knowledge matrix $\boldsymbol{X}$ clearly beats all the baselines with convincing margins across 10% to 50% of the total responses from each dataset. Moreover, even without access to the external knowledge $\boldsymbol{X}$, DELRA still performs close to that of SEEK+$\boldsymbol{X}$ and outperforms the other baselines when the sampling proportion is greater than 10%. When the sampling proportion is only 10%, DELRA without $\boldsymbol{X}$ seems to suffer from the same response sparsity as any other baseline that does not leverage $\boldsymbol{X}$.

7.5.4 Consistency between Estimated Relatedness and Ground-Truth Relatedness

In this experiment, we evaluate how consistent the estimates of the relatedness between categories from DELRA *without* $\boldsymbol{X}$ are with the relatedness scores in $\boldsymbol{X}$ pre-computed using equation 7.1 followed by the transformation in equation 7.4. More specifically, for each response category, we calculate the *Pearson correlation coefficient* between two rankings. The first ranking is the Top-N most related categories in terms of the model's estimate of their relatedness to the target category. The second ranking is the Top-N rank of the most related categories in terms of the pre-computed relatedness scores in the row of $\boldsymbol{X}$ that corresponds to the target category.

We set N to be 2, 3, 5, 10 and 15 to obtain the respective average Pearson correlation coefficients across all the response categories. We also implement two *supervised* baselines (in terms of knowing the true answers) for obtaining the Top$-N$ rank of the most related categories:

- **Top-N rank by frequency - asymmetric**. For each response category, the relatedness of a second response category to it is the frequency of the second category as the responses to questions with the first category as their true answers.



(a) **Dog** (b) **Bird**

(c) **Instrument** (d) **Movie**

Figure 7.4: Average Pearson correlation coefficients between the Top-N most related category rank yielded by different methods, and the ground-truth Top-N rank yielded by the pre-computed related scores based on equation 7.1. Note that x-axis and y-axis in each figure are N and average correlation, respectively.

- **Top-N rank by frequency - symmetric**. For each response category, the relatedness of a second response category is the total frequency of the second category as either the responses to questions with the first category as their true answers, or the true answers for questions which receive the first category as responses.

Figures 7.4a to 7.4d show how the Top-N most related category ranks by both the DELRA model without $\boldsymbol{X}$ and the two baselines are correlated with the Top-N ranks by the ground-truth relatedness scores calculated using equation 7.1. Showing overall higher average correlation with the ground-truth relatedness scores across the four datasets, our model clearly yields more consistent category relatedness estimates than the two baselines even though it is unsupervised. We conjecture this is due to the ability of our model to distinguish responses of different quality by accounting

for the interaction between worker expertise and item difficulty, while the baselines treat all the responses as the same in terms of their quality.

7.6 Conclusion

In this chapter we proposed the DELRA model, a quality control framework for crowdsourcing that leverages the *semantic relationships* between response options. The model features a latent real-valued matrix that captures the *relatedness* between response options alongside variables for worker expertise, question difficulty and question true answers. DELRA encodes the joint interaction among these variables to refine estimation of conditional probabilities of different response categories given different true answer categories. This leads DELRA to outperform numerous state-of-the-art quality control methods in terms of true answer prediction.

Moreover, DELRA allows for elegant encoding of *external prior knowledge* regarding the semantic relationships between response options for calibrating the estimation of the latent relatedness matrix. This leads to its further improvements in the true answer prediction even under much less response data by by limiting or preventing over-fitting. Finally, the relatedness matrix learned solely from the response data by DELRA shows good consistency with the relatedness matrix pre-computed from external semantic hierarchies organizing the categories.

With respect to the research question *RQ*4 and its sub-questions from *RQ*4.1 to *RQ*4.4, we have the following answers:

- **Answer to *RQ*4.1**: The semantic relationships between response options can be captured by a relatedness/similarity matrix with rows and columns corresponding to individual response options. The entries of the matrix accommodate real values that quantify the semantic similarities between each pair of response options.

  Another interpretation for this matrix is the confusion matrix that captures the extents to which crowd-workers confuse correct response categories with the other categories.

  In both interpretations, the larger the entry values in the matrix, the more similar or confusing the corresponding pair of response options are.

  The DELRA model we proposed in this chapter possesses a latent relatedness matrix which can be inferred directly from responses.

- **Answer to *RQ*4.2**: The previous assumption about the role of the response semantic relationships in determining correlations within crowdsourced responses is that response categories more related to the true answer categories are more likely to be given as responses regardless of worker ability and question difficulty.

  A more sophisticated assumption which we also observed in reality is that (i) more related categories are more likely to be given by more capable workers, and (ii) less related categories are more likely to be given to more difficulty questions.

  The DELRA model we proposed in this chapter captures the above assumption on how response semantic relationships interact with the worker expertise and the question difficulty using a ordered logit model. This model possesses a (response-specific) slope being the expertise-difficulty difference, and ordinal intercepts specific to individual response categories.

- **Answer to *RQ*4.3**: Since the number of response options considered here is very large, the latent relatedness matrix can end up being massive. Inferring such a massive matrix is difficult especially when the number of responses available for each pair of options is small. In this case, estimation of the corresponding entries in the matrix becomes unreliable.

  External knowledge about the response semantic relationships can be useful for indicating the possible values of these entries. An elegant way to make use of it is to encode it into the parameters of the prior distribution assumed over the entries, which smooths their unreliable estimates. We showed that with proper preprocessing steps for the external knowledge matrix, it can be incorporated into the prior mean of the Normal distribution which generates the entry values via a linear regression.

- **Answer to *RQ*4.4**: The DELRA model was able to achieve overall superior true answer prediction performance over several state-of-the-art quality control methods two of which had leveraged external knowledge about the semantic relationships. When incorporating the external knowledge itself, the DELRA model achieved even better performance in both full-response and sparse-response scenarios.

- **Answer to *RQ*4**: Our study in this chapter confirmed that we can build a statistical model which can effectively capture the joint interaction between workers, questions and response semantic relationships for improved quality control performance.

This model is based on a sophisticated assumption (as per our answer to *RQ*4.2) regarding how responses are formed under worker expertise, question difficulty and semantic relatedness between responses and true answers. It is able to not only infer a semantic relatedness matrix between response options directly from responses but also utilize external prior knowledge to refine the inference. The efficacy of the model with and without the external knowledge has been confirmed by defeating two state-of-the-art methods in the same area in terms of true answer prediction performance.

There are several opportunities to further advance this line of work in the future:

- It would be interesting to investigate the effect of an *asymmetric* latent relatedness matrix between response options on the quality control performance of DELRA.

- We used the inverse of the shortest paths between pairs of response options to pre-compute the observed relatedness matrix used in the Normal prior mean over the corresponding latent relatedness matrix. It would be interesting to see how sensitive the quality control performance of DELRA is to incorporating different observed relatedness matrices pre-computed by different similarity methods (e.g. Leacock-Chodorow Similarity [202], Wu-Palmer Similarity [203], etc.).

- Both the latent and the observed relatedness matrix are quadratic in the number of response options and therefore will be very large when the number of response options is large. In this case, using low-rank decomposition techniques could be useful for (i) lowering the complexity of the DELRA model to prevent over-fitting and (ii) learning the underlying factors that construct the observed relatedness matrix. In both cases, it is desirable to extend DELRA with a low-rank decomposition component (possibly using Bayesian hierarchical modelling techniques).

Chapter 8

Conclusion

Crowdsourcing has become a principal tool that allows research communities and companies to collect data for system development and business analysis, with significant cost savings and fast turnaround. However, cost-effective crowdsourcing is often elusive in terms of the quality of the response data (and the associated costs) due to various aspects that affect the response quality.

The most widely recognized aspect of crowdsourcing affecting the response quality is the crowd-workers themselves. Crowd-workers can have varying *expertise* for the tasks, and varying *motivations* to solve them which will determine the *effort* they exert and the level of *honesty* in their responses.

Traditional techniques applied to crowd-workers which ensure that crowdsourcing can produce the highest quality responses possible are: *worker filtering* which removes *unqualified* or *low-performing* workers and *majority vote* which *aggregates responses* from multiple workers to obtain the most voted (and thus the most trusted) response.

Worker filtering is an instance of crowdsourcing task design which specifies how to set up or adapt modules of crowdsourcing platforms to positively affect the question-answering behaviour of crowd-workers and thereby improve the quality of their responses.

Majority vote is an instance of a *wisdom-of-the-crowd* technique which aggregates multiple responses to the same question in order to produce a more accurate label. More advanced wisdom-of-the-crowd techniques focus on *statistical* aggregation and consider questions to have varying difficulty. Such techniques make use of *statistical models* that encode the latent attributes of crowd-workers (namely their level of ability) and of questions (namely their difficulty), and the *statistical dependency* of the observed responses on those attributes.

Both the task design and wisdom-of-the-crowd techniques aim to control the response quality such that low-quality/error-prone responses are removed, penalized, or discouraged while high-quality responses are treated in an opposite fashion. These techniques are thus referred to as the *quality control* methods.

This thesis focuses on advancing both task design and the wisdom-of-the-crowd techniques for achieving more cost-effective crowdsourcing. By identifying overlooked aspects that affect the response quality, we develop design and modelling techniques that incorporate these various effects to improve either the response quality itself or the estimation of the response quality to allow for more accurate aggregation for identifying the correct responses.

In Chapter 1, we started by introducing the concepts of *paid* and *unpaid* crowdsourcing. Confirming paid crowdsourcing to be the main focus of this thesis, we proceeded to discuss online platforms and their popularity for data collection. We emphasized that on these platforms, crowd-workers are largely *motivated by monetary rewards* while tasks usually offer low and fixed pay rates to these workers. This results in the workers solving the tasks quickly and carelessly with minimum effort and non-total honesty. We noted also that different workers have different skills and expertise which inherently prevents them from exhibiting the same performance when solving the same task.

The above problems cause the quality of crowdsourced responses to vary and become unreliable. As a solution, quality control methods aim to improve the quality of the overall response to each question. We pointed out that compared to worker filtering for quality control, the-wisdom-of-the-crowd approach preserves all the response data. This is a better alternative because each response carries non-zero information about the true answer and is indicative of the ability of the corresponding worker. By calculating proper statistics from the repsonses, one is able to estimate the workers' abilities. Then, by aggregating the responses weighted by the ability estimates of their corresponding workers one can compute a more accurate final response to each question (compared to simply using majority vote). Furthermore, both the worker ability estimation and the weighted response aggregation can be integrated into one statistical model in which they are conducted in alternate inference procedures.

8.1 Identification of Important Crowdsourcing Aspects

Following the introduction of quality control for paid crowdsourcing, we identified, from the literature, aspects of crowdsourcing and their attributes which either affect or indicate the quality of

the responses (See Figure 1.2). One important aspect is the crowd-worker. Its important attributes include the *motivation*, which determines the attributes *effort* and *truthfulness*, and the *expertise*. Other important aspects include the *questions*, the *response options* and the *contexts*.

We discussed the fact that questions in crowdsourcing can be categorized into being either *objective*, *purely subjective* or *partially subjective*. Objective questions can possess different levels of difficulty, while purely subjective questions are not characterised by any difficulty but only subjectivity (meaning that the correct response depends entirely on the individual responding). Meanwhile, partially subjective questions exhibit both subjectivity and difficulty respectively to certain extents.

We discussed also the fact that it is not unusual for crowdsourcing questions to involve large (but finite) sets of response options for crowd-workers to choose from. In this case, the underlying *semantic relationships* between these options become particularly useful for indicating the quality of responses and further, the worker expertise and the question difficulty. These relationships might be already organized by human assessors into explicit semantic structures (e.g. WordNet), or inferred as part of the statistical (wisdom-of-the-crowd) models.

Finally, we discussed the fact that workers can answer questions within different contexts (/environments) and that their individual motivations and response accuracy can vary across them. The context includes attributes like the question assignment mechanism used, the reward/payment mechanism, the training provided to workers, and the quality assurance systems. Changing a context can affect workers' motivations, and further the quality of their responses. For example, lowering the pay rate (in the middle of a task) demotivates workers, and makes their responses less accurate.

8.2 Definition of Granularities of Worker Contexts

A contribution we have made is to refine the crowdsoucing context into *different granularities* to shed light on the different extents to which the setting for a component of the context can be modified. According to our refinement, a context can be at either the *task level*, the *session/page level* or the *response level*. For example, the payment rate per question could vary for each task on the same platform, for each page of a task or even for each response. Moreover, contextual attributes can be categorized corresponding to the different levels (See Figure 1.2). These (observed) attributes capture the *side information* regarding the respective levels of contexts.

Assuming the above aspects and their attributes to be correlated with the response quality, we continued to find out *what quality control methods had been built by considering them (either individually or jointly)*. From these methods, *what research gaps are there in quality control for crowdsourcing*.

In chapter 2, we answered both these questions by conducting a comprehensive and in-depth literature review for quality control methods. Our review improves on the previous surveys first by providing a full anatomy of the quality control methods with respect to not only to the crowdsourcing aspects and their attributes (that affect the response quality) but also which interactions between them the methods have considered. In the previous surveys, the aspects and attributes involved are incomplete, and a characterisation of how different aspects and attributes are jointly considered by different quality control methods is missing. Providing this information systematically, our review allows researchers a better understanding of how quality control methods have evolved to consider ever more complex interactions.

8.3 A Quality Control Research Graph Visualization

We provided a *graph visualization* of the quality control methods. According to this graph, the methods can be broadly categorized into two types: *quality control design* and *statistical (wisdom-of-the-crowd) methods*. Quality control design deals with mechanisms that change the contexts to affect workers' behaviour. The task design methods make use of statistical methods to provide estimates of various attributes (e.g. worker expertise) based on which the mechanisms can perform their actions. The graph also shows that currently, the most sophisticated quality control methods have considered three aspects and their attributes which definitely include worker expertise and question difficulty. Chapter 2 was carried out based on the graph with each section corresponding to a particular part of the graph and arranged by following a top-down traversal across its different levels.

8.4 Technical Taxonomies of Quality Control Papers

Another novel contribution from our review are the *technical taxonomies* of the quality control methods. They are placed in front of each section of chapter 2. Thus, each of them corresponds to a particular type of quality control methods (either the designs or the statistical models) which considered different (combinations of) aspects and attributes. A taxonomy is characterized by the

considered aspects and attributes, modelling assumptions, parameter estimation techniques, and design features of different quality control methods of the same type. The papers that proposed them are specified at the bottom of the taxonomy. These taxonomies provide straightforward and informative guidance for researchers to identify new research directions.

8.5 Identification of Design Issues for Gamifying Crowdsourcing

At the end of the review, we summarized all the major issues we have found as the research gaps that exist in current quality control techniques for crowdsourcing. There are two broad issues respectively for quality control design and statistical (wisdom-of-the-crowd) models. The broad issue for quality control design is the *general lack of empirically justified methodology for the gamification of paid crowdsourcing tasks*. The current designs are in general ad-hoc and based on designers' subjective impressions on what game elements might be useful for particular crowdsourcing applications. Under this broad issue, a more specific issue is that an empirical study of what combinations of game elements are the most useful for improving the response quality in a common crowdsourcing application is still missing. Addressing this specific issue allows us to provide a proper solution to the broad quality control design issue.

8.6 Identification of Modelling and Inference Issues for Statistical Response Aggregation

The broad issue identified for quality control using statistical models is that the current modelling techniques and inference procedures are still incomplete and limited. This is due to: (i) the impractical assumption that responses are always sufficient for reliable model inference and (ii) ignorance of critical aspects and attributes responsible for generating responses under important scenarios. The first issue is common in statistical inference, but is a more pressing problem in quality control for crowdsourcing where few responses are usually collected for each question in a task. This issue remains to be solved and a corresponding practical solution is to collect and use *side information* regarding the various aspects of crowdsourcing that are useful to predict the values of the attributes being estimated. There are several aspects for which side information can be collected (in the form of observed features). Current statistical models are *ad-hoc* in the sense that can only encode side information of a single aspect at a time and thus cannot be generalized to take

account of all information that is available at once. A more reasonable and yet thus-far unstudied solution is to build a *unified scalable* statistical framework that allows for *arbitrary types of side information* to be incorporated.

The second issue is not a general statistical inference problem but is central to the statistical modelling for crowdsourcing. It consists of two specific issues we found during the literature review, each corresponding to an important but previously overlooked scenario in crowdsourcing. The first one is that when tasks contain *partially subjective* questions (e.g. relevance judgements), current statistical models fail to distinguish the difficulty of a question from its subjectivity. This resulted in the models confusing relatively subjective questions with difficult questions, and confusing expert workers who possess distinct preferences with adversarial workers. The second issue is that when a task contains a *large finite* set of response options, the inherent correlation between responses due to the *semantic relationships* between the options has been overlooked by past quality control research.

In chapter 3, we proposed research questions corresponding to the research issues found in the literature review. We answered these research questions separately in chapters 4, 5, 6, and 7. By answering these questions, we provide valuable and promising findings for the future work to succeed under the corresponding research directions.

8.7 Empirical Study of Gamified Paid Crowdsourcing

According to the identified quality control design issue, no formal empirical study had been conducted on how to perform the gamification or more precisely, what game elements to put together (alongside elements of a paid crowdsourcing platform) to maximize the quality of responses. To resolve this issue, we focused on the most important game element: competition, which has rarely used to gamify paid crowdsourcing. The first research question thus concerned how to best gamify general crowdsourcing applications with *competitions* among crowd-workers. It asked:

***RQ*1**: *In order to maximise response quality for a given crowdsourcing budget, which competition elements should be leveraged for gamifying crowdsourcing tasks: a real-time performance score for each crowd-worker, a performance leaderboard accessible by any worker in the task, or them combined together?*

In chapter 4, we answered this question by conducting three controlled A/B/n tests which incrementally added game elements and crowdsourcing elements to the same task. The crowdsourcing

elements included the paying of bonuses and the use of control questions for both pre-task worker vetting and in-task worker filtering.

The first test involves a control group of workers provided with default settings (a non-zero pay rate) and three treatment groups provided with extended settings which incrementally added real-time performance feedback and a task leaderboard (with two different scoring functions). The second experiment advertised a bonus for the top-10 crowd-workers. The third experiment introduced control questions to filter unreliable workers (before and during the tasks).

Our findings from the three experiments were as follows:

- For *paid* crowdsourcing, scoring-based real-time performance feedback for incentivising crowd-workers has little effect on the workers' response accuracy. This finding is different from that of Ipeirotis and Gabrilovich [11] which showed that for *unpaid* crowdsourcing the total number of correct responses as the performance feedback can significantly improve the response accuracy. Apart from the obvious difference between paid and unpaid crowdsourcing, the different findings could also be because the application handled by Ipeirotis and Gabrilovich [11] required very specific expertise (i.e. medical knowledge) and was advertised publicly online, which attracted a population of users very different from crowd-workers.

- Additionally providing a task leaderboard to crowd-workers has little effect on their response accuracy (compared to providing no feedback). This finding is close to that of Ipeirotis and Gabrilovich [11] although their experiments showed that the leaderboard brings a negative effect. This means that regardless of whether workers are paid or voluntary, there is no guarantee that they will work harder if put in competition with others.

- Combining the above two game elements with a bonus payment to the top workers on the leaderboard appears necessary to motivate all the workers to work extra hard to win the competition. This finding is inspiring as it shows that combining a small bonus [1] with a global-leaderboard competition can already improve the response quality. More importantly, the total amount of bonus paid does not (necessarily) depend on the number of questions in a task as other work does [139, 204].

- Using control questions to remove unqualified or low-performing workers does not adversely affect the above usefulness of a task leaderboard under the payment of a bonus reward. We also note that with the presence of the control questions (before and during a task), the

[1] In the second test, a total of $10 bonus was rewarded with $1 for each top 10 worker.

overall response accuracy of crowd-workers is improved in all cases. This means that most crowd-workers under paid crowdsourcing are so money-driven that they need to be monitored and (informed that they can be) filtered as early as they take the qualification quiz.

The overall message delivered by chapter 4 is that for paid crowdsourcing applications which require human judgement, crowd-workers are difficult to motivate via a global competition with a task leaderboard unless a bonus prize is rewarded to a small number of top workers on that leaderboard. Additionally, most of the workers are so money-driven that control questions need to be in place at all time to ensure a higher average response accuracy (compared to them not being in place).

8.8 A Unified Scalable Framework to Leverage Side Information for Sparse Response Aggregation

One of the issues regarding the modelling and the inference procedures for statistical wisdom-of-the-crowd aggregation is the vulnerability of the current models in handling response sparsity. In practice, the number of responses given to each question is usually kept very small due to limited budgets and large numbers of questions that need to be answered. The number of responses given by each worker can vary greatly with some being very small due to either the *cold-start* or the *early drop-out* of crowd-workers. As a result, the estimation of worker expertise and question difficulty can be very unreliable.

To solve the above issue, we leveraged side information regarding different crowdsourcing aspects that can inform estimates of either the worker expertise or the question difficulty. To *simultaneously* utilize the various types of side information, we need a *unified scalable statistical modelling and inference* framework that encodes all the information types and their respective influences on the expertise and difficulty. The second research question thus asks about:

RQ2: *Can we build a unified scalable statistical modelling and inference framework that is able to integrate and utilize arbitrary types of side information about different crowdsourcing aspects for better controlling the quality of crowdsourced responses, thereby improving the true answer prediction under response sparsity?*

In chapter 5, we built such a framework based on the GLAD model [87] by linearly adding regressions over features that capture the different types of side information into the corresponding factors of workers and questions. The side information features we have leveraged include the

demographic and *personality trait* features for crowd-workers, the *content* and *topic* features for questions, and the *contextual* features for both the workers' sessions and the individual responses (See Table 5.2).

From the experiment for true answer prediction under increasing degrees of response sparsity, we have the following finding:

- When responses are few per worker, our framework leveraging any type of side information achieves much better *true answer prediction* performance than smoothing estimation of worker confusion matrices using the community-based Dawid&Skene model [73] (See Figure 5.4).

Our framework extends from GLAD which models the correctness probability of a response while the community-based DS extends from the DS model which models the conditional probabilities in the worker confusion matrices. Estimating a response correctness probability is less likely to overfit the scarce responses compared to estimating a confusion matrix for each worker. The overfitting is even more easily prevented when side information is incorporated into the expertise-difficulty factorization of the correctness probability.

Another finding from the unseen response prediction experiment is the following:

- At the same degree of response sparsity, our framework leveraging all the four types of side information outperforms the community-based DS in unseen response prediction for both the TREC and the Stackoverflow tasks. For the Evergreen task, our framework leveraging three different types of side information clearly beats the community-based DS (See Figure 5.5).

A worker confusion matrix is better at capturing the response probabilities of a worker compared to the response correctness probability. This is more obvious for non-binary tasks (i.e. TREC relevance judgements) for which GLAD only assumes a uniform distribution for the incorrect responses. In this case, incrementally incorporating more types of side information into the estimation of correctness probabilities helps amplify the signals of incorrect responses to the extent that they can be captured by not only the community-based DS but also our framework.

The biggest advantage of our framework is its practicality in quality control for crowdsourcing. This practicality comes not only from its neat incorporation of various types of side information but also the heuristics we derived from the experiments for setting up both the frameworks' hyper-parameters and its optimization parameters. These heuristics were discussed in 5.4.4 and summarized below:

- The Normal prior mean for the question difficulty in the framework needs to be calibrated at 0 as in GLAD [87] (to ensure the model identifiability). The only difference is that the prior variance is recommended to be set at 0.1 (as opposed to 1 in GLAD) to ensure a strong regularization for battling scarce responses per question.

- As for the prior mean and variance of the worker expertise, the *uninformative* setting used by GLAD is 1 for both values. An *informative* alternative is to make use of external knowledge about the expertise of workers from the crowdsourcing platform being used.

  For example, in CrowdFlower, crowd-workers are ordered into different *levels* based on their performance in previous tasks. They range from the *level-1* workers to the *level-3* workers (i.e. from being the least to the most professional and reliable). In our tasks, only level-3 (i.e. the most elite) workers were allowed to participate. Given this information, we adjusted the prior mean and variance accordingly, setting them to 2 and 0.1 respectively (to reflect the eliteness) compared to both 1 in GLAD (which covers a wider variety of workers).

- Both the regularization terms for the coefficients corresponding to the side information features and their step sizes in the gradient descent need to be set up relative to the data quantities they deal with. More specifically, the regularization terms need to be the scaled in accordance with the inverse of the number of associated side information features. Their step sizes need to be the scaled with the inverse of the number of associated data points. Such settings guarantee the stability of the parameter estimation procedures.

8.9 A Statistical Model Distinguishing Subjectivity from Difficulty for Partially Subjective Questions

The second issue regarding the modelling and the inference procedures for statistical aggregation is handling partially subjective questions. The current techniques are unable (i) to distinguish the *difficulty* of a question from its *subjectivity* and (ii) to distinguish expert workers with distinct *subjective opinions* regarding the true answer from either misled or adversarial workers who give objectively incorrect answers. The third research question thus asks about:

RQ3: *Can we better control the quality of worker responses, thereby improving true answer prediction for partially subjective questions, by distinguishing question difficulty from question*

subjectivity in the modelling and inference of the worker-question interaction which generates the responses?

In chapter 6, we proposed a model which *considers both the worker-specific truths regarding the correct answers to each question and also the difficulty-dependent probability that a worker's response will equal her perceived subjective truth*. This model allows us to provide a measure of subjectivity for each question. It also allows us to derive a ranking of questions in terms of either difficulty or subjectivity that is coherent with human assessment.

The important findings from chapter 6 are listed below:

- Unseen (held-out) response prediction can be used to determine the number of preferences which exist among crowd-workers and cause them to respond with particular correct responses to each partially subjective question. More specifically, if an improvement is observed in the unseen response prediction of the model by introducing an additional preference variable, it means that there likely does exist such a preference among crowd-workers. This additional preference (variable) allows our model to account for extra meaningful correlation between correct responses to the questions. If the addition of an extra preference variable results in no improvement, it indicates that only random noise remains which can be readily captured by the expertise-difficulty dependent logistic function of our model. In this case, the random noise corresponds to the arbitrary responses of crowd-workers due to their inexpertise and the difficulty of the questions.

- Both conducting a *clustering* operation on the responses across questions, our model achieves overall superior performance over the MdWC model [25] in terms of both true answer prediction and the unseen response prediction for partially subjective questions. This is because MdWC focuses on grouping correct responses assuming there exists only *one correct response* to each question even though the questions can be partially subjective. As a result, worker expertise tends to be under-estimated while question difficulty tends to be over-estimated for such partially subjective questions. On the other hand, our model focuses on grouping correct responses assuming that *multiple correct responses* to each question could exist. This results in the expertise and the difficulty being properly estimated, which has led to the improvements in the two types of prediction.

- Special attention is required for designing the human evaluation of questions' subjectivity and difficulty. First of all, we found that asking crowd-workers to perform self-appraisals

for these two properties of the questions is very challenging. This is because most of the workers are not experts and thus cannot necessarily tell how many correct answers a question has and how hard it actually is for them to give the correct answers. Investigation of how to effectively crowdsource question subjectivity and difficulty is needed in the future.

A better albeit more costly alternative is to ask for the help from experts, which we did in chapter 6 for the object identification and image beauty task. We compared model estimates of question subjectivity and difficulty with the corresponding expert assessments. More specifically, we asked the assessors to do the following:

- Rank all the images by how hard/confusing they are for crowd-workers to judge correctly by avoiding possible incorrect answers.
- Rank them this time by how subjective they are for crowd-workers to judge.

The object identification sub-task consists of 36 questions with low subjectivity and varied difficulty, while the image beauty sub-task consists of 24 questions with varied subjectivity and difficulty. According to the feedback from the experts, the object identification sub-task is easy and clear to judge. For the image beauty sub-task, however, had a meta question not been asked, it could have been much harder for the experts to conduct the difficulty judgement.

In the original paper [111], the meta question asks crowd-workers to select the 12 most beautiful images from a set of 24 images each of which corresponds to a binary question (i.e. whether an image is beautiful or not). According to the experts, this meta question helps simplify the difficulty judgement. All the experts did was first placing images with low subjectivity (i.e. those which are clearly beautiful or not beautiful) either into the list or off the list. Then, for those subjective images (i.e. those which can be either beautiful or not beautiful depending on the workers' own opinions), the common strategy adopted by the experts was to group them based on their compositions, topics and delivered feelings. Each image group corresponds to a potential group of crowd-workers in terms of their preferences.

Given the size restriction for the list, the experts then worked out which group, should it be preferred by crowd-workers and included into the list, had more beautiful images than what was required, or needed more images to fill out the list. The experts assumed that in the first case, each worker had to remove some images they preferred and in the second case, they had to include some images they did not think were beautiful. The images involved in both

cases are thus more difficult for the workers (to decide where to place them, inside or outside the list) than the other images.

We are conservative about whether the above assessment procedures can be applied to another application (e.g. relevance judgement) especially in *large scale*. Nevertheless, the message we have got by talking to the expert assessors in this application is that for them to accurately assess the difficulty of (highly) subjective questions, they need strong signals indicating the difficulty. In the image beauty judgement application, restriction on the number of beautiful images sends a strong signal that subjective images can cause difficulty for the crowd-workers to decide where to place them.

8.10 A Statistical Model Leveraging Semantic Relationships between Response Options

The third issue regarding the modelling and the inference procedures for statistical aggregation is handling semantic relationships between response options. Few techniques have been developed that consider the semantic relatedness between categories of correct and incorrect responses to explain the response variation. Moreover, these techniques rely on external knowledge to pre-compute relatedness scores between the categories rather than estimate the scores directly from responses along with worker expertise and question difficulty. The fourth research question thus asks about:

RQ4: *Can we build a statistical model that can effectively learn the joint interactions between workers, questions and response semantic relationships for better controlling the quality of responses, thereby improving the true answer prediction, when the number of response options is sufficiently large?*

In chpater 7, we proposed a statistical model which encodes the semantic relationships between response options into a latent symmetric real-value matrix where each entry representing the extent of relatedness between the correct response and incorrect responses. The model also encodes the expertise of workers and the difficulty of questions. More importantly, it captures the interaction among the response semantic relationships, the expertise and the difficulty by leveraging the following two findings we obtained from the preliminary study.

- More capable crowd-workers tend to choose response options which are more closely related to the correct response to a question.

- More difficult questions tend to receive responses which are less semantically related to their correct responses.

The above findings are encoded as an *ordinal regression* in our model. It contains a *response-specific slope* as the difference between the question difficulty and the worker expertise. It also contains ordinal *intercepts* corresponding to each off-diagonal entry in a particular row of the relatedness matrix (for a particular correct response option). Comparing our model with several state-of-the-art quality control models on question true answer prediction using either full responses or sub-sampled responses yields the following finding:

- Our model which incorporates and infers a relatedness matrix between response options, and uses it to refine the estimation of response probabilities of crowd-workers is superior to the other models in terms of true answer prediction. This includes outperforming two state-of-the-art models which directly incorporate external knowledge about the relatedness between response options. This is attributed to our model's encoding of the above findings on how the relatedness between options is correlated with worker expertise and question difficulty, which is missing in those models.

  For true answer prediction with sub-sampled responses, our model is less dominant but still achieves similar performance to the two models that considered the external knowledge, while outperforming other models. The reason we conjecture is that it becomes harder for our model to infer the relatedness matrix as reliably as before due to the fewer responses.

We completed our model by leveraging the external knowledge about the relatedness between response options to improve its estimation under sparse responses. This was achieved by incorporating the pre-computed relatedness scores (from the external knowledge hierarchies) into the Normal prior mean over the relatedness variables in the model using a linear regression. We evaluated the augmented model with the same experiments and found that:

- With the external knowledge incorporated, our model clearly outperforms all other models with respect to true answer prediction using either full responses or sub-sampled responses. This also confirms the effectiveness of the augmentation in adjusting the model estimation to make it more reliable under response sparsity.

How exactly one should maximize the effectiveness of the augmentation remains an open question. It requires studies on different methods that can be used to pre-compute the relatedness

scores, and different pre-processing methods (e.g. normalization, logarithm) on the scores before they are fed to the linear regression for the prior mean.

Research communities have been using crowdsourcing to evaluate and reconstruct semantic relationships between categories/objects. Frequencies of different response categories given to questions with a particular true answer can indicate the relatedness between those categories and the true answer category. In the last experiment of chapter 7, we investigated whether the relatedness estimates from our model could provide a better alternative to the frequency-based methods. The finding was promising with our model yielding the Top-N most related categories (to each true answer category) that are more consistent with the ground-truths compared to two frequency-based methods. This shows that our model improves quality control for crowdsourced relatedness between response options by taking into account their correlation with worker expertise and question difficulty.

8.11 Future Directions

Based on the future work proposed at the end of each chapter, we summarize some important future directions for quality control in crowdsourcing:

- More research attention is required for *competition-based gamification* of crowdsourcing tasks which currently lack proper investigations. Apart from the systematic studies we have performed on leaderboard competition in chapter 4, there are other types of competitions worth to investigate.

  A basic one is the *duel* competition model between a pair of crowd-workers. Past gamification research [122] has focused on utilizing the *agreement* between the pair of workers rather than their direct competition. Investigation of leveraging pair-wise competition between workers in crowdsourcing involves how to design the pairing mechanism, the competition content, and the reward mechanism. The pairing mechanism specifies how to select a pair of crowd-workers to compete. The competition content concerns a certain number of questions that need to be answered by both workers. The reward mechanism specifies the criteria of being a winner and how to reward the winner (and possibly the loser) in the duel.

  An advanced type of competitions is the *tournament* where crowd-workers are paired in a number of rounds to compete until one winner emerges at the last round. The tournament competition can be viewed as an extension of the duel competition (with one tournament

consisting of multiple duel competitions). Thus, their investigation focuses should be similar with the reward design for the tournament being more complex.

Investigation is required on which type of competitions (i.e. duels, tournaments and leaderboards) is better at improving the response quality in general crowdsourcing applications. We envision a series of A/B/n (response quality) testing on each type of competition under different design strategies (e.g. for reward incentives, worker pairing, etc.) with a wide variety of crowdsourcing tasks.

- It would be interesting to develop a toolkit which implements and integrates all the statistical response aggregation models developed in chapters 5, 6 and 7. The integration is feasible as these models are all based on the GLAD model [87]. Under the object-oriented programming paradigm, it means that these models *inherit* from the GLAD model. The inheritances include the (collections of) variables for workers' expertise, questions' difficulty and true answers, and responses' (correctness) probabilities. They also include the function that calculates the correctness probabilities using the variables. Apart from the (shared) inheritances, each of our models also add in its own variables and functions that serve to address different scenarios in crowdsourcing that were previously not covered by GLAD or any other aggregation toolkit.

  In these scenarios, data requesters can use our toolkit to perform response aggregation to obtain more accurate estimates of true labels (compared to the other toolkits). The scenarios include those addressed by the different chapters of this thesis:

  - *Scenario 1:* A task has few responses collected from each worker or for each question or both, and has side information features collected. In this case, the data requester can choose to use the side-information-aware model which we developed in chapter 5 and is integrated into the toolkit.
  - *Scenario 2:* A task consists of partially subjective questions. In this case, the data requester can choose to use the subjectivity-aware model developed in chapter 6.
  - *Scenario 3:* A task has a large finite set of response options for each question among which semantic relationships exist. In this case, the data requester can choose to use the response-relatedness-aware model developed in chapter 7. If the requester has some external knowledge about the semantic relationships, this knowledge can be input to the model to boost the aggregation performance.

The toolkit will provide not only the estimates of true labels but also other estimates that are able to help data requesters with their applications. For all the above scenarios, the toolkit will provide predictions of responses from individual workers to be given to unseen questions. Specifically for scenario 1, the toolkit will provide the importance of individual side information features in predicting the correctness/accuracy of each response. For scenario 2, the toolkit will provide the estimates of questions' subjectivity and groups of crowd-workers which have different preferences due to the subjectivity. For scenario 3, the toolkit will provide the estimate of the (semantic) relatedness between every pair of response options.

In summary, with such a toolkit, data requesters are able to perform effective and interpretable statistical quality control for their data crowdsourced under the above scenarios.

Bibliography

[1] A. Marcus and A. Parameswaran, "Crowdsourced data management: Industry and academic perspectives," *Foundations and Trends in Databases*, vol. 6, no. 1-2, pp. 1–161, 2015.

[2] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 1220–1229.

[3] D. Liu, R. G. Bias, M. Lease, and R. Kuipers, "Crowdsourcing for usability testing," in *Proceedings of the Association for Information Science and Technology*, vol. 49, no. 1. Wiley Online Library, 2012, pp. 1–10.

[4] K.-J. Stol and B. Fitzgerald, "Two's company, three's a crowd: a case study of crowdsourcing software development," in *Proceedings of the 36th International Conference on Software Engineering*. ACM, 2014, pp. 187–198.

[5] M. Hossain, "Crowdsourcing: Activities, incentives and users' motivations to participate," in *2012 International Conference on Innovation Management and Technology Research*, May 2012, pp. 501–506.

[6] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu *et al.*, "Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189, 2008.

[7] A. Mao, E. Kamar, Y. Chen, E. Horvitz, M. E. Schwamb, C. J. Lintott, and A. M. Smith, "Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing," in *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[8] A. Mao, E. Kamar, and E. Horvitz, "Why stop now? predicting worker engagement in online crowdsourcing," in *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[9] E. Kamar, A. Kapoor, and E. Horvitz, "Lifelong learning for acquiring the wisdom of the crowd," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013, pp. 2313–2320.

[10] E. Kamar, A. Kapoor, and Horvitz, "Identifying and accounting for task-dependent bias in crowdsourcing," in *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[11] P. G. Ipeirotis and E. Gabrilovich, "Quizz: Targeted crowdsourcing with a billion (potential) users," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 143–154.

[12] P. G. Ipeirotis, "Analyzing the amazon mechanical turk marketplace," *XRDS: Crossroads, the ACM Magazine for Students*, vol. 17, no. 2, pp. 16–21, 2010.

[13] A. Kumar and M. Lease, "Learning to rank from a noisy crowd," in *Proceedings of the 34th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2011, pp. 1221–1222.

[14] V. Ambati, "Active learning and crowdsourcing for machine translation in low resource scenarios," Ph.D. dissertation, 2012.

[15] A. Brew, D. Greene, and P. Cunningham, "the interaction between supervised learning and crowdsourcing," in *NIPS workshop on computational social science and the wisdom of crowds*, 2010.

[16] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2013, pp. 580–587.

[17] J. Cheng and M. S. Bernstein, "Flock: Hybrid crowd-machine learning classifiers," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 600–611.

[18] C. Eickhoff and A. P. de Vries, "Increasing cheat robustness of crowdsourcing tasks," *Information Retrieval*, vol. 16, no. 2, pp. 121–137, 2013.

[19] S. M. Herzog and R. Hertwig, "the wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping," *Psychological Science*, vol. 20, no. 2, pp. 231–237, 2009.

[20] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, 2010, pp. 64–67.

[21] J. Wang, P. G. Ipeirotis, and F. Provost, "Managing crowdsourcing workers," in *2011 Winter Conference on Business Intelligence*. Citeseer, 2011.

[22] J. Surowiecki, *the Wisdom of Crowds*. Anchor, 2005.

[23] Y. Yan, R. Rosales, G. Fung, and J. G. Dy, "Active learning from crowds," in *Proceedings of the 28th International Conference on Machine Learning*. Omnipress, 2011, pp. 1161–1168.

[24] O. Alonso and S. Mizzaro, "Using crowdsourcing for trec relevance assessment," *Information Processing & Management*, vol. 48, no. 6, pp. 1053 – 1066, 2012.

[25] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "the multidimensional wisdom of crowds," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2010, pp. 2424–2432.

[26] A. T. Nguyen, M. Halpern, B. C. Wallace, and M. Lease, "Probabilistic modeling for crowdsourcing partially-subjective ratings," in *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[27] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[28] C. Eickhoff, "Crowd-powered experts: Helping surgeons interpret breast cancer images," in *Proceedings of the First International Workshop on Gamification for Information Retrieval*. ACM, 2014, pp. 53–56.

[29] G. Chatzimilioudis, A. Konstantinidis, C. Laoudias, and D. Zeinalipour-Yazti, "Crowdsourcing with smartphones," *IEEE Internet Computing*, vol. 16, no. 5, pp. 36–44, 2012.

[30] J. Bragg, W. EDU, and D. S. Weld, "Learning on the job: Optimal instruction for crowdsourcing," in *ICML Workshop on Crowdsourcing and Machine Learning*, 2015.

[31] M. Venanzi, J. Guiver, P. Kohli, and N. R. Jennings, "Time-sensitive bayesian information aggregation for crowdsourcing systems," *Journal of Artificial Intelligence Research*, vol. 56, pp. 517–545, 2016.

[32] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Computing*, vol. 17, no. 2, pp. 76–81, 2013.

[33] A. I. Chittilappilly, L. Chen, and S. Amer-Yahia, "A survey of general-purpose crowdsourcing techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2246–2266, 2016.

[34] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh, "Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions," *ACM Computing Surveys*, vol. 51, no. 1, p. 7, 2018.

[35] G. Kazai, J. Kamps, and N. Milic-Frayling, "the face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 2583–2586.

[36] J. Muhammadi, H. R. Rabiee, and A. Hosseini, "A unified statistical framework for crowd labeling," *Knowledge and Information Systems*, vol. 45, no. 2, pp. 271–294, 2015.

[37] J. Zhang, X. Wu, and V. S. Sheng, "Learning from crowdsourced labeled data: a survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 543–576, 2016.

[38] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: is the problem solved?" *Proceedings of the VLDB Endowment*, vol. 10, no. 5, pp. 541–552, 2017.

[39] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, "the future of crowd work," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, 2013, pp. 1301–1318.

[40] X. Yin, W. Liu, Y. Wang, C. Yang, and L. Lu, "What? how? where? a survey of crowdsourcing," in *Frontier and Future Development of Information Technology in Medicine and Education*. Springer, 2014, pp. 221–232.

[41] D. Geiger and M. Schader, "Personalized task recommendation in crowdsourcing information systems - current state of the art," *Decision Support Systems*, vol. 65, pp. 3–16, 2014.

[42] N. Luz, N. Silva, and P. Novais, "A survey of task-oriented crowdsourcing," *Artificial Intelligence Review*, vol. 44, no. 2, pp. 187–213, 2015.

[43] M. Lease and E. Yilmaz, "Crowdsourcing for information retrieval: introduction to the special issue," *Information Retrieval*, vol. 16, no. 2, pp. 91–100, 2013.

[44] K. Mao, L. Capra, M. Harman, and Y. Jia, "A survey of the use of crowdsourcing in software engineering," *Journal of Systems and Software*, vol. 126, pp. 57 – 84, 2017.

[45] G. Xintong, W. Hongzhi, Y. Song, and G. Hong, "Brief survey of crowdsourcing for data mining," *Expert Systems with Applications*, vol. 41, no. 17, pp. 7987–7994, 2014.

[46] B. L. Ranard, Y. P. Ha, Z. F. Meisel, D. A. Asch, S. S. Hill, L. B. Becker, A. K. Seymour, and R. M. Merchant, "Crowdsourcing - harnessing the masses to advance health and medicine, a systematic review," *Journal of General Internal Medicine*, vol. 29, no. 1, pp. 187–203, 2014.

[47] C. Gomes, D. Schneider, K. Moraes, and J. d. Souza, "Crowdsourcing for music: Survey and taxonomy," in *2012 IEEE International Conference on Systems, Man, and Cybernetics*, 2012, pp. 832–839.

[48] C. Heipke, "Crowdsourcing geospatial data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 550–557, 2010.

[49] R. M. Frongillo, Y. Chen, and I. A. Kash, "Elicitation for aggregation," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 900–906.

[50] C.-J. Ho, R. Frongillo, and Y. Chen, "Eliciting categorical data for optimal aggregation," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016, pp. 2450–2458.

[51] N. M. Villegas, C. Sánchez, J. Díaz-Cely, and G. Tamura, "Characterizing context-aware recommender systems: A systematic literature review," *Knowledge-Based Systems*, vol. 140, pp. 173–200, 2018.

[52] H. J. Jung, Y. Park, and M. Lease, "Predicting next label quality: A time-series model of crowdwork," in *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[53] H. J. Jung and M. Lease, "Modeling temporal crowd work quality with limited supervision," in *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[54] Y.-L. Fang, H.-L. Sun, P.-P. Chen, and T. Deng, "Improving the quality of crowdsourced image labeling via label similarity," *Journal of Computer Science and Technology*, vol. 32, no. 5, pp. 877–889, 2017.

[55] T. Han, H. Sun, Y. Song, Y. Fang, and X. Liu, "Incorporating external knowledge into crowd intelligence for more specific knowledge acquisition," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 1541–1547.

[56] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied Statistics*, pp. 20–28, 1979.

[57] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[58] W. Tang and M. Lease, "Semi-supervised consensus labeling for crowdsourcing," in *SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval*, 2011, pp. 1–6.

[59] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2014, pp. 1260–1268.

[60] Y. Zhang, X. Chen, D. Zhou, and M. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3537–3580, 2016.

[61] C. Gao and D. Zhou, "Minimax optimal convergence rates for estimating ground truth from crowdsourced labels," *arXiv preprint arXiv:1310.5764*, 2013.

[62] B. Carpenter, "A hierarchical bayesian model of crowdsourced relevance coding," 01 2011.

[63] H.-C. Kim and Z. Ghahramani, "Bayesian classifier combination," in *Artificial Intelligence and Statistics*, 2012, pp. 619–627.

[64] E. Simpson, S. J. Roberts, I. Psorakis, and A. Smith, "Dynamic bayesian combination of multiple imperfect classifiers." *Decision Making and Imperfection*, vol. 474, pp. 1–35.

[65] A. Ghosh, S. Kale, and P. McAfee, "Who moderates the moderators?: crowdsourcing abuse detection in user-generated content," in *Proceedings of the 12th ACM Conference on Electronic Commerce*. ACM, 2011, pp. 167–176.

[66] D. R. Karger, S. Oh, and D. Shah, "Budget-optimal crowdsourcing using low-rank matrix approximations," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2011, pp. 284–291.

[67] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi, "Aggregating crowdsourced binary ratings," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 285–294.

[68] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Advances in Neural Information Processing Systems*, 2011, pp. 1953–1961.

[69] D. R. Karger, S. Oh, and Shah, "Budget-optimal task allocation for reliable crowdsourcing systems," *Operations Research*, vol. 62, no. 1, pp. 1–24, 2014.

[70] Q. Liu, J. Peng, and A. T. Ihler, "Variational inference for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2012, pp. 692–700.

[71] J. Ok, S. Oh, J. Shin, and Y. Yi, "Optimality of belief propagation for crowdsourced classification," in *International Conference on Machine Learning*, 2016, pp. 535–544.

[72] T. Bonald and R. Combes, "A minimax optimal algorithm for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2017.

[73] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 155–164.

[74] P. G. Moreno, A. Artes-Rodriguez, Y. W. Teh, and F. Perez-Cruz, "Bayesian nonparametric crowdsourcing," *Journal of Machine Learning Research*, 2015.

[75] C. Liu and Y.-M. Wang, "Truelabel+ confusions: a spectrum of probabilistic models in analyzing multiple ratings," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 17–24.

[76] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas, "Crowdsourcing for multiple-choice question answering." in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014, pp. 2946–2953.

[77] R. Yan, Y. Song, C.-T. Li, M. Zhang, and X. Hu, "Opportunities or risks to reduce labor in crowdsourcing translation? characterizing cost versus quality via a pagerank-hits hybrid model," in *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 1025–1032.

[78] S. Kajimura, Y. Baba, H. Kajino, and H. Kashima, "Quality control for crowdsourced poi collection," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2015, pp. 255–267.

[79] T. Sunahase, Y. Baba, and H. Kashima, "Pairwise hits: Quality estimation from pairwise comparisons in creator-evaluator crowdsourcing process." in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, pp. 977–984.

[80] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *Acm Sigkdd Explorations Newsletter*, vol. 17, no. 2, pp. 1–16, 2016.

[81] R. W. Ouyang, L. Kaplan, P. Martin, A. Toniolo, M. Srivastava, and T. J. Norman, "Debiasing crowdsourced quantitative characteristics in local businesses and services," in *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*. ACM, 2015, pp. 190–201.

[82] R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, and T. J. Norman, "Aggregating crowdsourced quantitative claims: Additive and multiplicative models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1621–1634, 2016.

[83] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1297–1322, 2010.

[84] C. Gao, Y. Lu, and D. Zhou, "Exact exponent in optimal rates for crowdsourcing," in *International Conference on Machine Learning*, 2016, pp. 603–611.

[85] D. R. Karger, S. Oh, and D. Shah, "Efficient crowdsourcing for multi-class labeling," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 1, pp. 81–92, 2013.

[86] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[87] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in Neural Information Processing Systems*, 2009, pp. 2035–2043.

[88] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver, "How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing," *arXiv preprint arXiv:1206.6386*, 2012.

[89] D. Zhou, S. Basu, Y. Mao, and J. C. Platt, "Learning from the wisdom of crowds by minimax entropy," in *Advances in Neural Information Processing Systems*, 2012, pp. 2195–2203.

[90] D. Zhou, Q. Liu, J. C. Platt, C. Meek, and N. B. Shah, "Regularized minimax conditional entropy for crowdsourcing," *arXiv preprint arXiv:1503.07240*, 2015.

[91] P. Ruvolo, J. Whitehill, and J. R. Movellan, "Exploiting structure in crowdsourcing tasks via latent factor models," Tech. Rep., 2010.

[92] P. Ruvolo, J. Whitehill, and J. Movellan, "Exploiting commonality and interaction effects in crowdsourcing tasks using latent factor models," in *Neural Information Processing Systems. Workshop on Crowdsourcing: Theory, Algorithms and Applications*, 2013.

[93] H. Kajino, Y. Tsuboi, and H. Kashima, "A convex formulation for learning from crowds." in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.

[94] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han, "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation," in *Proceedings of the 21th*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 745–754.

[95] F. L. Wauthier and M. I. Jordan, "Bayesian bias mitigation for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2011, pp. 1800–1808.

[96] L. Yin, J. Han, W. Zhang, and Y. Yu, "Aggregating crowd wisdoms with label-aware autoencoders," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 1325–1331.

[97] K. Atarashi, S. Oyama, and M. Kurihara, "Semi-supervised learning from crowds using deep generative models," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI, 2018.

[98] A. Gaunt, D. Borsa, and Y. Bachrach, "Training deep neural nets to aggregate crowdsourced responses," in *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence. AUAI Press*, 2016, p. 242251.

[99] B. Lakshminarayanan and Y. W. Teh, "Inferring ground truth from multi-annotator ordinal data: a probabilistic approach," *arXiv preprint arXiv:1305.0015*, 2013.

[100] P. Metrikov, J. Wu, J. Anderton, V. Pavlu, and J. A. Aslam, "A modification of lambdamart to handle noisy crowdsourced assessments," in *Proceedings of the 2013 Conference on the Theory of Information Retrieval*. ACM, 2013, p. 31.

[101] H. J. Jung and M. Lease, "Inferring missing relevance judgments from crowd workers via probabilistic matrix factorization," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2012, pp. 1095–1096.

[102] H. J. Jung and Lease, "Improving quality of crowdsourced labels via probabilistic matrix factorization," in *Proceedings of the 4th Human Computation Workshop at AAAI*, 2012, pp. 101–106.

[103] H. J. Jung and M. Lease, "Crowdsourced task routing via matrix factorization," *arXiv preprint arXiv:1310.5142*, 2013.

[104] H. J. Jung, "Quality assurance in crowdsourcing via matrix factorization based task routing," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 3–8.

[105] D. Zhou, Q. Liu, J. Platt, and C. Meek, "Aggregating ordinal labels from crowds by minimax conditional entropy," in *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 2014, pp. 262–270.

[106] T. L. Griffiths and Z. Ghahramani, "the indian buffet process: An introduction and review," *Journal of Machine Learning Research*, vol. 12, no. Apr, pp. 1185–1224, 2011.

[107] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[108] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[109] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations*, 2013.

[110] P. Metrikov, V. Pavlu, and J. A. Aslam, "Aggregation of crowdsourced ordinal assessments and integration with learning to rank: A latent trait model," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 1391–1400.

[111] Y. Tian and J. Zhu, "Learning from crowds in the presence of schools of thought," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge discovery and Data Mining*. ACM, 2012, pp. 226–234.

[112] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[113] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*. Springer, 2015, pp. 77–118.

[114] C. E. Antoniak, "Mixtures of dirichlet processes with applications to bayesian nonparametric problems," *the Annals of Statistics*, pp. 1152–1174, 1974.

[115] N. Kaufmann, T. Schulze, and D. Veit, "More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk." in *Americas Conference on Information Systems*, vol. 11, no. 2011, 2011, pp. 1–11.

[116] D. Chandler and A. Kapelner, "Breaking monotony with meaning: Motivation in crowdsourcing markets," *Journal of Economic Behavior & Organization*, vol. 90, pp. 123–133, 2013.

[117] N. Shah, D. Zhou, and Y. Peres, "Approval voting and incentives in crowdsourcing," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 10–19.

[118] N. B. Shah and D. Zhou, "Double or nothing: Multiplicative incentive mechanisms for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2015, pp. 1–9.

[119] N. Shah and D. Zhou, "No oops, you won't do it again: mechanisms for self-correction in crowdsourcing," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, pp. 1–10.

[120] M. Yin and Y. Chen, "Bonus or not? learn to reward in crowdsourcing," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2015.

[121] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary Educational Psychology*, vol. 25, no. 1, pp. 54–67, 2000.

[122] L. Von Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[123] B. Waggoner and Y. Chen, "Output agreement mechanisms and common knowledge," in *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[124] N. Miller, P. Resnick, and R. Zeckhauser, "Eliciting informative feedback: the peer-prediction method," *Management Science*, vol. 51, no. 9, pp. 1359–1373, 2005.

[125] R. Jurca and B. Faltings, "Collusion-resistant, incentive-compatible feedback payments," in *Proceedings of the 8th ACM Conference on Electronic Commerce*. ACM, 2007, pp. 200–209.

[126] R. Jurca, B. Faltings *et al.*, "Mechanisms for making crowds truthful," *Journal of Artificial Intelligence Research*, vol. 34, no. 1, p. 209, 2009.

[127] Y. Kong, K. Ligett, and G. Schoenebeck, "Putting peer prediction under the micro (economic) scope and making truth-telling focal," in *International Conference on Web and Internet Economics*. Springer, 2016, pp. 251–264.

[128] D. Prelec, "A bayesian truth serum for subjective data," *science*, vol. 306, no. 5695, pp. 462–466, 2004.

[129] D. C. Parkes and J. Witkowski, "A robust bayesian truth serum for small populations," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 2012.

[130] B. Faltings, R. Jurca, P. Pu, and B. D. Tran, "Incentives to counter bias in human computation," in *Second AAAI conference on human computation and crowdsourcing*, 2014.

[131] G. Radanovic, B. Faltings, and R. Jurca, "Incentives for effort in crowdsourcing using the peer truth serum," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 4, p. 48, 2016.

[132] A. Dasgupta and A. Ghosh, "Crowdsourced judgement elicitation with endogenous proficiency," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 319–330.

[133] J. Witkowski, Y. Bachrach, P. Key, and D. C. Parkes, "Dwelling on the negative: Incentivizing effort in peer prediction," in *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[134] Y. Liu and Y. Chen, "Learning to incentivize: eliciting effort via output agreement," *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016.

[135] Y. Liu and Chen, "Sequential peer prediction: Learning to elicit effort using posted prices." in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, pp. 607–613.

[136] ——, "Machine-learning aided peer prediction," in *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 2017, pp. 63–80.

[137] X. A. Gao, A. Mao, Y. Chen, and R. P. Adams, "Trick or treat: putting peer prediction to the test," in *Proceedings of the 5th ACM conference on Economics and Computation*. ACM, 2014, pp. 507–524.

[138] Y. Singer and M. Mittal, "Pricing mechanisms for crowdsourcing markets," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 1157–1166.

[139] C.-J. Ho, A. Slivkins, S. Suri, and J. W. Vaughan, "Incentivizing high quality crowdwork," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 419–429.

[140] B. Faltings and G. Radanovic, "Game theory for data science: Eliciting truthful information," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 11, no. 2, pp. 1–151, 2017.

[141] G. Radanovic and B. Faltings, "Incentives for truthful information elicitation of continuous signals," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, no. EPFL-CONF-215878, 2014, pp. 770–776.

[142] M. Brenner, N. Mirza, and E. Izquierdo, "People recognition using gamified ambiguous feedback," in *Proceedings of the 1st International Workshop on Gamification for Information Retrieval*. ACM, 2014, pp. 22–26.

[143] A. Dumitrache, L. Aroyo, C. Welty, R.-J. Sips, and A. Levas, "Dr. detective: combining gamication techniques and crowdsourcing to create a gold standard in medical text," in *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web*, 2013, pp. 16–31.

[144] C. G. Harris, "the beauty contest revisited: Measuring consensus rankings of relevance using a game," in *Proceedings of the 1st International Workshop on Gamification for Information Retrieval*. ACM, 2014, pp. 17–21.

[145] J. He, M. Bron, L. Azzopardi, and A. de Vries, "Studying user browsing behavior through gamified search tasks," in *Proceedings of the 1st International Workshop on Gamification for Information Retrieval*. ACM, 2014, pp. 49–52.

[146] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan, "Quality through flow and immersion: gamifying crowdsourced relevance assessments," in *Proceedings of the 35th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2012, pp. 871–880.

[147] S. Saito, T. Watanabe, M. Kobayashi, and H. Takagi, "Skill development framework for micro-tasking," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2014, pp. 400–409.

[148] L. Guillot, Q. Bragard, R. Smith, and A. Ventresque, "Towards a gamified system to improve translation for online meetings," in *the 3rd International Workshop on Gamification for Information Retrieval*. CEUR, 2016.

[149] J. Schlotterer, C. Seifert, L. Wagner, and M. Granitzer, "A game with a purpose to access europe's cultural treasure." in *the 2nd International Workshop Gamification for Information Retrieval*, 2015.

[150] W. Moazzam, M. Riegler, S. Sen, and M. Nygaard, "Scientific hangman: Gamifying scientific evidence for general public." in *the 2nd International Workshop Gamification for Information Retrieval*, 2015.

[151] L. C. Stanculescu, A. Bozzon, R.-J. Sips, and G.-J. Houben, "Work and play: An experiment in enterprise gamification," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 2016, pp. 346–358.

[152] J. Boyd-Graber, B. Satinoff, H. He, and H. Daume III, "Besting the quiz master: Crowdsourcing incremental classification games," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 1290–1301.

[153] M. Dontcheva, R. R. Morris, J. R. Brandt, and E. M. Gerber, "Combining crowdsourcing and learning to improve engagement and performance," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 3379–3388.

[154] T. Y. Lee, C. Dugan, W. Geyer, T. Ratchford, J. C. Rasmussen, N. S. Shami, and S. Lupushor, "Experiments on motivational feedback for crowdsourced workers." in *the 7th International AAAI Conference on Web and Social Media*, 2013, pp. 341–350.

[155] A. D. Mason, G. Michalakidis, and P. J. Krause, "Tiger nation: Empowering citizen scientists," in *the 2012 6th IEEE International Conference on Digital Ecosystems and Technologies*. IEEE, 2012, pp. 1–5.

[156] T. Itoko, S. Arita, M. Kobayashi, and H. Takagi, "Involving senior workers in crowdsourced proofreading," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2014, pp. 106–117.

[157] E. Massung, D. Coyle, K. F. Cater, M. Jay, and C. Preist, "Using crowdsourcing to support pro-environmental community activism," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 371–380.

[158] K. Jennings, *Brainiac: adventures in the curious, competitive, compulsive world of trivia buffs*. Villard Books, 2007.

[159] U. Gadiraju, B. Fetahu, and R. Kawase, "Training workers for improving performance in crowdsourcing microtasks," in *Design for Teaching and Learning in a Networked World*. Springer, 2015, pp. 100–114.

[160] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2008, pp. 614–622.

[161] P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang, "Repeated labeling using multiple noisy labelers," *Data Mining and Knowledge Discovery*, vol. 28, no. 2, pp. 402–441, 2014.

[162] J. Wang and P. Ipeirotis, "A framework for quality assurance in crowdsourcing," 2013.

[163] J. Wang, P. G. Ipeirotis, and F. Provost, "Cost-effective quality assurance in crowd labeling," *Information Systems Research*, vol. 28, no. 1, pp. 137–158, 2017.

[164] J. Zhang, X. Wu, and V. S. Shengs, "Active learning with imbalanced multiple noisy labeling," *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 1095–1107, 2015.

[165] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[166] B. Bollobás, *Random Graphs*. Cambridge University Press, 2001, no. 73.

[167] P. Donmez, J. Carbonell, and J. Schneider, "A probabilistic framework to learn from multiple annotators with time-varying accuracy," in *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM, 2010, pp. 826–837.

[168] D. Mandal, M. Leifer, D. C. Parkes, G. Pickard, and V. Shnayder, "Peer prediction with heterogeneous tasks," *arXiv preprint arXiv:1612.00928*, 2018.

[169] K. Mo, E. Zhong, and Q. Yang, "Cross-task crowdsourcing," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2013, pp. 677–685.

[170] M. Fang, J. Yin, and X. Zhu, "Knowledge transfer for multi-labeler active learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 273–288.

[171] M. Fang, J. Yin, and D. Tao, "Active learning for crowdsourcing using knowledge transfer." in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014, pp. 1809–1815.

[172] H. Zhuang and J. Young, "Leveraging in-batch annotation bias for crowdsourced active learning," in *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 243–252.

[173] H. Zhuang, A. Parameswaran, D. Roth, and J. Han, "Debiasing crowdsourced batches," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1593–1602.

[174] A. Kobren, C. H. Tan, P. Ipeirotis, and E. Gabrilovich, "Getting more for less: Optimized crowdsourcing with dynamic tasks and goals," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 592–602.

[175] H. Hu, Y. Zheng, Z. Bao, G. Li, J. Feng, and R. Cheng, "Crowdsourced poi labelling: Location-aware result inference and task assignment," in *Proceedings of the IEEE 32nd International Conference on Data Engineering*. IEEE, 2016, pp. 61–72.

[176] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[177] R. L. Plackett, "The analysis of permutations," *Applied Statistics*, pp. 193–202, 1975.

[178] C.-J. Ho, S. Jabbari, and J. W. Vaughan, "Adaptive task assignment for crowdsourced classification," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 534–542.

[179] A. Khetan and S. Oh, "Achieving budget-optimality with adaptive schemes in crowdsourcing," in *Advances in Neural Information Processing Systems*, 2016, pp. 4844–4852.

[180] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[181] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[182] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information processing & management*, vol. 36, no. 5, pp. 697–716, 2000.

[183] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: survey and practical guide," *Data mining and knowledge discovery*, vol. 18, no. 1, pp. 140–181, 2009.

[184] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," in *ACM SigIR Forum*, vol. 42, no. 2. ACM, 2008, pp. 9–15.

[185] M. Lease, "On quality control and machine learning in crowdsourcing." *Human Computation*, vol. 11, no. 11, 2011.

[186] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[187] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative big five trait taxonomy," *Handbook of personality: Theory and research*, vol. 3, no. 2, pp. 114–158, 2008.

[188] G. Kazai, J. Kamps, and N. Milic-Frayling, "An analysis of human factors and label accuracy in crowdsourcing relevance judgments," *Information retrieval*, vol. 16, no. 2, pp. 138–178, 2013.

[189] H. Li, B. Zhao, and A. Fuxman, "the wisdom of minority: discovering and targeting the right group of workers for crowdsourcing," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 165–176.

[190] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[191] G. Rasch, "Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests." 1960.

[192] T. P. Minka, "A comparison of numerical optimizers for logistic regression," in *technical report*, 2003.

[193] C. Buckley, M. Lease, and M. D. Smucker, "Overview of the TREC 2010 relevance feedback track (notebook)," in *the 19th Text Retrieval Conference Notebook.*, 2010.

[194] M. Lease and G. Kazai, "Overview of the TREC 2011 crowdsourcing track," in *Proceedings of the Text Retrieval Conference*, 2011.

[195] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1483–1494, 2012.

[196] B. Loni, M. Menendez, M. Georgescu, L. Galli, C. Massari, I. S. Altingovde, D. Martinenghi, M. Melenhorst, R. Vliegendhart, and M. Larson, "Fashion-focused creative commons social dataset," in *Proceedings of the 4th ACM Multimedia Systems Conference*, 2013, pp. 72–77.

[197] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden, "Scaling up crowd-sourcing to very large datasets: a case for active learning," *Proceedings of the VLDB Endowment*, vol. 8, no. 2, pp. 125–136, 2014.

[198] "Adult datset," https://github.com/ipeirotis/Get-Another-Label/tree/master/data, accessed: 2017-07-30.

[199] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[200] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-f. Li, "L.: Novel dataset for fine-grained image categorization," in *First Workshop on Fine-Grained Visual Categorization, CVPR (2011*. Citeseer.

[201] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," California Institute of Technology, Tech. Rep. CNS-TR-2010-001, 2010.

[202] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *WordNet: An electronic lexical database*, vol. 49, no. 2, pp. 265–283, 1998.

[203] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[204] M. Yin, Y. Chen, and Y.-A. Sun, "the effects of performance-contingent financial incentives in online labor markets." 2013.