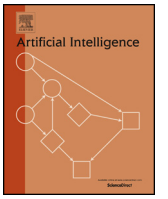
Artificial Intelligence 287 (2020) 103351

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



A technical survey on statistical modelling and design methods for crowdsourcing quality control

Yuan Jin [a,*], Mark Carman [b], Ye Zhu [c], Yong Xiang [c]

[a] *Faculty of Information Technology, Monash University, Victoria, 3800, Australia*
[b] *DEIB, Politecnico di Milano, Milan, 20133, Italy*
[c] *School of Information Technology, Deakin University, Victoria, 3125, Australia*

ARTICLE INFO

Article history:
Received 2 December 2018
Received in revised form 2 July 2019
Accepted 9 June 2020
Available online 30 June 2020

Keywords:
Crowdsourcing
Quality control
Statistical modelling and inference
Mechanism design

ABSTRACT

Online crowdsourcing provides a scalable and inexpensive means to collect knowledge (e.g. labels) about various types of data items (e.g. text, audio, video). However, it is also known to result in large variance in the quality of recorded responses which often cannot be directly used for training machine learning systems. To resolve this issue, a lot of work has been conducted to control the response quality such that low-quality responses cannot adversely affect the performance of the machine learning systems. Such work is referred to as the quality control for crowdsourcing. Past quality control research can be divided into two major branches: *quality control mechanism design* and *statistical models*. The first branch focuses on designing measures, thresholds, interfaces and workflows for payment, gamification, question assignment and other mechanisms that influence workers' behaviour. The second branch focuses on developing statistical models to perform effective aggregation of responses to infer correct responses. The two branches are connected as statistical models (i) provide parameter estimates to support the measure and threshold calculation, and (ii) encode modelling assumptions used to derive (theoretical) performance guarantees for the mechanisms. There are surveys regarding each branch but they lack technical details about the other branch. Our survey is the first to bridge the two branches by providing technical details on how they work together under frameworks that systematically unify crowdsourcing aspects modelled by both of them to determine the response quality. We are also the first to provide taxonomies of quality control papers based on the proposed frameworks. Finally, we specify the current limitations and the corresponding future directions for the quality control research.

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*

* Corresponding author.
E-mail addresses: yuan.jin@monash.edu (Y. Jin), mark.carman@polimi.it (M. Carman), ye.zhu@ieee.org (Y. Zhu), yong.xiang@deakin.edu.au (Y. Xiang).

https://doi.org/10.1016/j.artint.2020.103351
0004-3702/

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 [2–4].

Over the past decade, online human intelligence marketplaces have been thriving, providing organized and billed crowdsourcing services to requesters all over the world. Online crowds are registered with the marketplaces as *crowd-workers* who are autonomous and generally receive a small payment, typically a few cents [5], for finishing each question. Two popular crowdsourcing marketplaces are Amazon Mechanical Turk[1] (AMT) and CrowdFlower.[2]

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 [6], machine translation systems [7] and general supervised learning systems [8–10].

As these platforms continue to grow by attracting more label collection tasks, they allow people with *varied abilities* and *motivations* to join them to share the labelling workloads. This results in a *deterioration* in the *quality of labels* and varieties of cheating behaviour prevalent on the platforms [11].

1.1. Crowdsourcing quality control

To deal with the above issues, quality control for crowdsourcing (QCC) is required by which the influence of *high-quality* labels is guaranteed to outweigh that of *low-quality* ones.

Worker filtering is the main QCC method used by crowdsourcing platforms to deal with label quality deterioration and cheating behaviour of crowd-workers. It removes two types of workers: *unqualified* workers and *low-performing* workers. The *unqualified* workers are removed using a *quiz* before a task commences. The quiz contains only *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. The worker filtering mechanism typically removes all the responses given by each low-performing crowd-worker. Other qualified workers then have to redo all the questions answered by these workers. This results in more budget and time consumption.

The wisdom of the crowd (WoC) [12] is an alternative to worker filtering that does not discard any response from low-performing workers. It retains the influence of their correct responses and tries to *smooth out* the influence of the incorrect ones. It centres on the observation that proper *aggregation* of *multiple answers* given by different people to the same question is able to yield a better answer. For example, websites, such as Rotten Tomatoes,[3] Netflix[4] and Last.fm,[5] utilize WoC methods (which aggregate reviews from their users) to provide summary reviews and overall recommendation scores about their items. These summary reviews and scores have turned out to be very accurate in depicting the underlying quality of the items.

The efficacy of the WoC approach relies on two crucial aspects. The first aspect is the *redundancy* of the responses to the same question. 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. The second aspect is the *aggregation* of the redundant responses for eliciting an accurate final answer. Two prerequisites need to be satisfied for the aggregation to work properly. First, the majority of the crowd are reliable in deriving their answers. Second, there are sufficient responses collected for each question. The WoC aggregation and the worker filtering can be applied together. The filtering is typically applied prior to or during the crowdsourcing, while the WoC aggregation is typically applied during or after the crowdsourcing.

The simplest WoC approach is the *majority vote* (MV). It considers the correct answer to a question to be the one endorsed by the majority of the crowd-workers. It assumes that each response is independent to one another and has the same quality irrespective of workers' abilities. Consequently, its performance is usually limited, especially when responses collected for each question are scarce.

1.2. Statistical models for quality control

Crowdsourced responses are fundamentally *not independent* and *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). The latter implies that responses with higher quality should be modelled to have greater influence in the WoC aggregation and vice-versa.

Statistical models provide a means of encoding the dependency of response quality on relevant aspects of crowdsourcing (e.g. crowd-workers who gave the responses). *Inference* procedures can be applied to these models to estimate the response quality and important attributes (e.g. worker ability/expertise) of its dependent aspects. Aggregation for obtaining the correct response to each question is also carried out by the inference procedures.

[1] https://www.mturk.com/.
[2] https://www.crowdflower.com/.
[3] https://www.rottentomatoes.com/.
[4] https://www.netflix.com/.
[5] https://www.last.fm/.

Table 1
A summary of current QCC surveys.

Survey	Areas	Details	Pros	Cons
[13] [14] [15]	Mechanism Design	Review of designs for tasks/ measures/interfaces that allow: - Assessment of response quality (e.g. by quiz, expert/peer review); - Assurance of response quality (e.g. by worker filtering, selection, training, team work)	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.
[16] [17] [18]	Statistical Modelling Methods	Review of statistical models for QCC which estimate: - Worker ability/expertise; - Question difficulty; - Question true answer; - Response quality;	Technical details about a variety of statistical models are specified. They include model assumptions, variables parameter estimation, etc.	Lack of design information on crowdsourcing applications. Ignoring aspects other than worker and question which also affect response quality: - Context (in which workers are situated, e.g. time, location); - Answer options (their semantic relationships).

1.3. Quality control mechanism design

A *quality control mechanism* is a program which runs to (reactively or proactively) control modules of a crowdsourcing task which interact with crowd-workers to improve their performance. The *design* of such a mechanism specifies how the control is conducted. For instance, a payment (and bonus) module is common to crowdsourcing tasks. In this case, a payment mechanism can be designed to manipulate this module on the time and amount workers get rewarded such that they are constantly motivated to do their best.

Quality control mechanisms are usually based on statistical models which provide various estimates designed to trigger and direct the mechanisms' control of the task modules. For example, in payment mechanism designs, the ability estimate for a worker can be used to determine whether she needs to be rewarded or not. In addition, the modelling assumptions can also be used to derive theoretical guarantees for the mechanisms that they inspire desirable worker behaviour (e.g. being honest in their responses). We will discuss these in details later in the survey.

1.4. Related surveys

A variety of surveys in the area of crowdsourcing have been published in the past. The subjects of these reviews include general overviews of crowdsourcing [19,20], management of certain components of crowdsourcing platforms, such as the routing and recommendation of tasks [21,22], and different applications of crowdsourcing, such as information retrieval [23], software engineering [24], data mining [25], health and medicine [26], music [27] and neogeography [28].

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: *quality control mechanism design* and *statistical (modelling) methods*, which is summarized in Table 1.

Current design surveys [13–15] mainly review *general-purpose* design strategies for building QCC mechanisms, response quality measures and user interfaces deployed on various task modules. As shown in Fig. 1, their main weakness is that they barely provide any technical details about the statistical methods (and vice versa). These methods provide the designs with various estimates, and how the designs utilize these estimates to achieve their goals.

Current statistical modelling surveys [16–18] focus on the models and inference procedures that learn attributes of two crowdsourcing aspects: the crowd-worker and the question. Table 1 shows their attributes 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. For example, worker *effort* and *honesty* are attributes that have been frequently modelled by game-theoretic methods used for incentive designs [29,30]. Questions that might contain more than one correct answers (thereby suggesting *subjectivity*) have also been studied and modelled in [31]. Current surveys however ignore quality control methods that model and learn 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 is characterized by the time, location, labelling device, pay rate, Web-page, etc. It has been modelled and learned for quality control purposes in [32–34].

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 according to recent QCC research [35–37]. Our survey reviews all the QCC research that deals with these aspects.

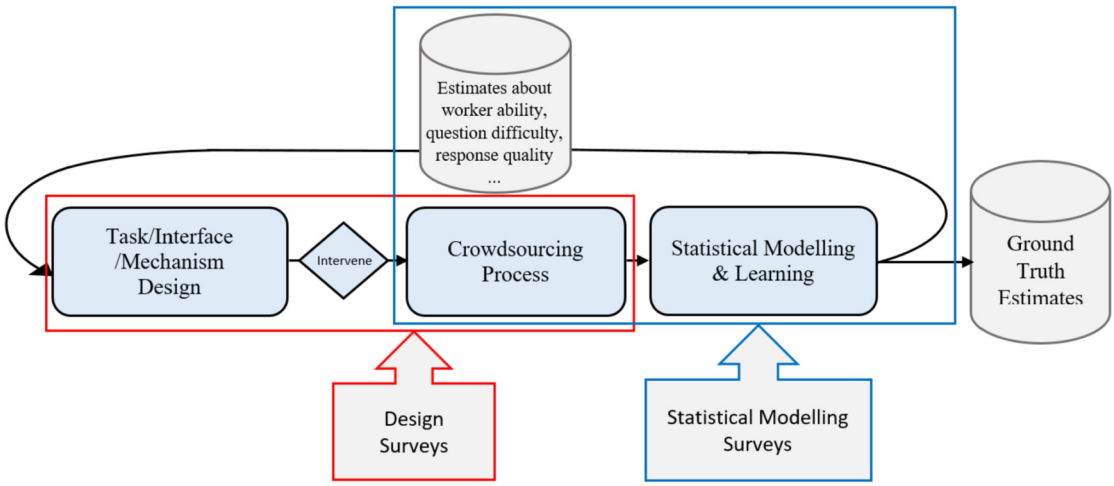


Fig. 1. A diagram that shows a quality control process for crowdsourcing and its different parts reviewed by quality control design and statistical modelling surveys. There however lacks a current survey that links the statistical models with the designs.

1.5. Contributions

In this work, we overcome the weaknesses of current QCC surveys with following contributions:

- A unified taxonomy of the key aspects of crowdsourcing considered by the QCC research to determine the response quality, along with their important attributes.
- A graph framework of all the QCC research in which the nodes represent the crowdsourcing aspects and their attributes considered by the research. The graph contains paths indicating different lines of research.
- A systematic review of all the QCC research based on the graph. The review starts from the most basic work, which considered only the crowd-worker ability, and finishes with the most sophisticated work, which considered multiple aspects and attributes.
- Hierarchical categorization of QCC papers. The hierarchies are constructed according to features of the papers (e.g. considered aspects, modelling assumptions, parameter estimation techniques, design features, etc.).
- An in-depth discussion of the QCC research and identification of current research limitations along with the proposal of future research directions.

2. Crowdsourcing aspects

The QCC research makes (explicit or implicit) assumptions on how quality of responses is correlated with certain aspects of crowdsourcing. It proposes to encode these assumptions into quality control mechanisms and statistical models to effectively control the response quality. To conduct a systematic review on this research, we start by specifying four prominent aspects of crowdsourcing it has considered. These aspects (along with their key attributes) are shown in Fig. 2.

Items/Questions. A question is the smallest unit in a crowdsourcing task. It is *objective* when it has a single correct answer. It is *purely subjective* when every answer is correct (e.g. a demographic question). In between lies *partially subjective* questions. Intuitively speaking, it has multiple correct answers but at least one incorrect answer. For instance, consider the task to judge if an image contains a person wearing fashionable clothes or not [38,39]. Workers can disagree on what the correct answers are for an image that contains a clothes-wearing person as they have different preferences for fashion styles. Nonetheless, they should agree on what the wrong answer is for that image (i.e. no person in it).

Both objective and partially subjective questions possess certain degrees of *difficulty* which obscure their correct answers from crowd-workers to various extents. A difficult question leads to variation in the responses across crowd-workers. For extremely difficult questions, workers may have to resort to random guessing.

Crowd-workers. A worker has certain *motivation* for answering the questions, and a certain level of (*domain*) *expertise* required by the subject of a task. The motivation governs the levels of *effort* exerted by the worker to answer each question, and also the *truthfulness* of the worker's responses. When two workers possess the same level of expertise (in the same domain), the one motivated to exert more effort is more likely to yield high-quality responses. The truthfulness of the worker's response determines how likely she responds with the answer she believes to be correct (for a question). If the worker is malicious, she is more likely to give a different response.

When encountering a partially subjective question (e.g. judging whether an item is fashion-related or not), a worker would exhibit *preferences* for certain *features* of the item (e.g. vintage design and fabrics). Such preferences are independent from the ability of the worker, thereby having no effect on the correctness of a response.

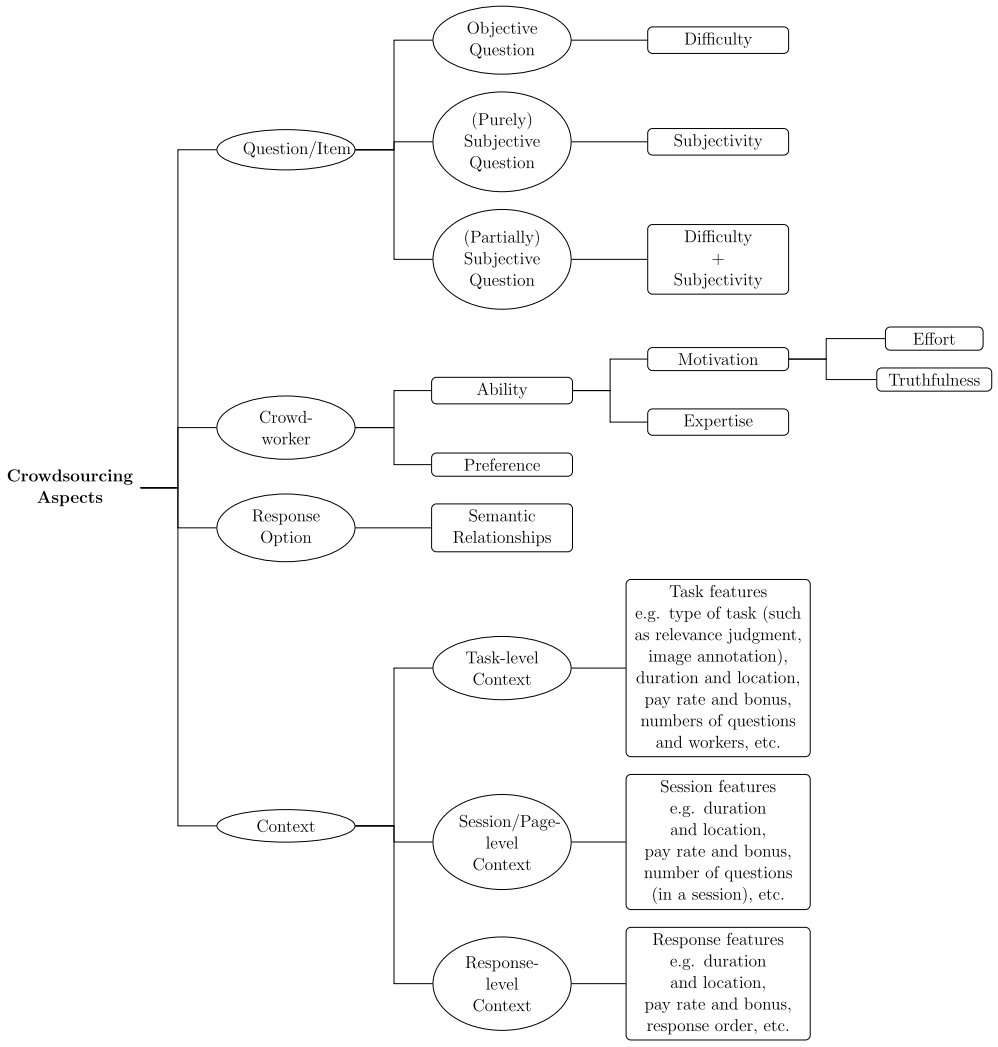


Fig. 2. Aspects of crowdsourcing considered by the QCC research. Each ellipse node denotes a particular crowdsourcing aspect. Question and context have finer definitions at level 2. Rectangle nodes denote attributes of aspects exploited by the QCC methods.

Response options. It is possible for a crowdsourcing task to have a large (but finite) set of response options for its questions. A typical example is the crowdsourcing for the ImageNet database [40] which stores millions of images according to tens of thousands of categories that are connected in a *semantic relational* graph. In this case, the semantic relationships between these categories (as response options) would influence the workers' responses. For instance, consider the classification task of 120 dog breeds from the ImageNet [41]. A worker is more likely to confuse the correct breed (e.g. golden retriever) with a breed that is more related to it (e.g. Labrador) than with a less related one (e.g. Chihuahua).

Contexts. A context in crowdsourcing is an *environment* in which crowd-workers are situated. Changing or intervening in the context can affect the motivation of workers which further affects the quality of their responses. From the literature, we found that different QCC methods control the context at different *levels of granularity*. We thus propose to refine the definition of the context according to the following three levels:

- *Task* level: a task-level context is characterized by 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 took multiple tasks such as micro-averaged and macro-averaged worker response time for each task.
- *Session* level: a session-level context within a task corresponds to a question page of the task that a 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 can thus be described by features regarding a task page (e.g. the pay rate for answering each question on the page, their topics and total word counts, etc.). They also concern the worker's responses

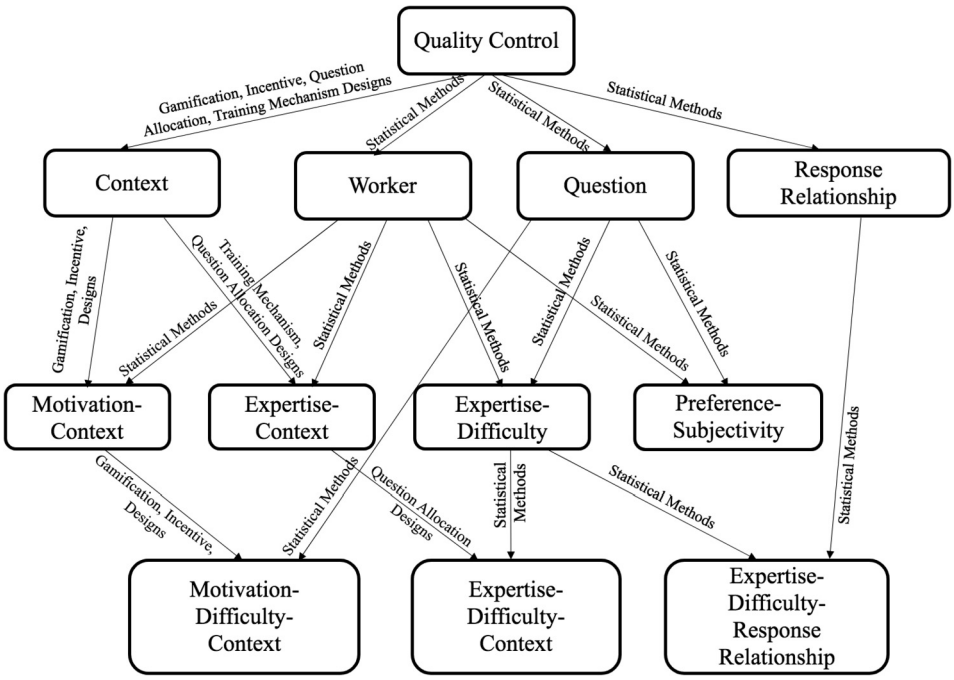


Fig. 3. A graph framework of the past QCC research.

within a working session (e.g. the average response time on each question, the total response time, the device used for the response, 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 worker to a question, and is described by features regarding this response (e.g. its payment, duration, location and position in the sequence of all the responses given by the same worker).

3. A graph framework of the QCC research

This survey is organized according to the graph shown in Fig. 3. 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. The root node "Quality Control" has outgoing edges to the four major crowdsourcing aspects specified in Fig. 2. Nodes at the second level correspond to *pairs* of attributes which have been jointly considered by some of the 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 developed by the QCC research that exploited the aspects. For example, *gamification, payment, question allocation*, and *training* mechanisms all make use of the worker context to control response quality. Edges labelled "Statistical Methods", indicate that the corresponding research focuses on statistical modelling and inference of the attributes.

A path in the graph denotes a line of research. For example, the path from node "Context" to node "Expertise-Difficulty-Context" denote the line of research that designed question allocation mechanisms by considering more aspects (i.e. worker expertise and question difficulty).

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

We carry out the rest of the survey according to the proposed graph. Each section corresponds to a node in the graph and the QCC methods reviewed in the section are given by the labels.

4. Modelling worker ability

Modelling the effect that individual crowd-workers have on the quality of responses (QoR) is most widely adopted by the QCC research. The effect is assumed to be determined by the ability/expertise of each worker. We have developed a taxonomy (see Fig. 4) to summarize the corresponding research work.

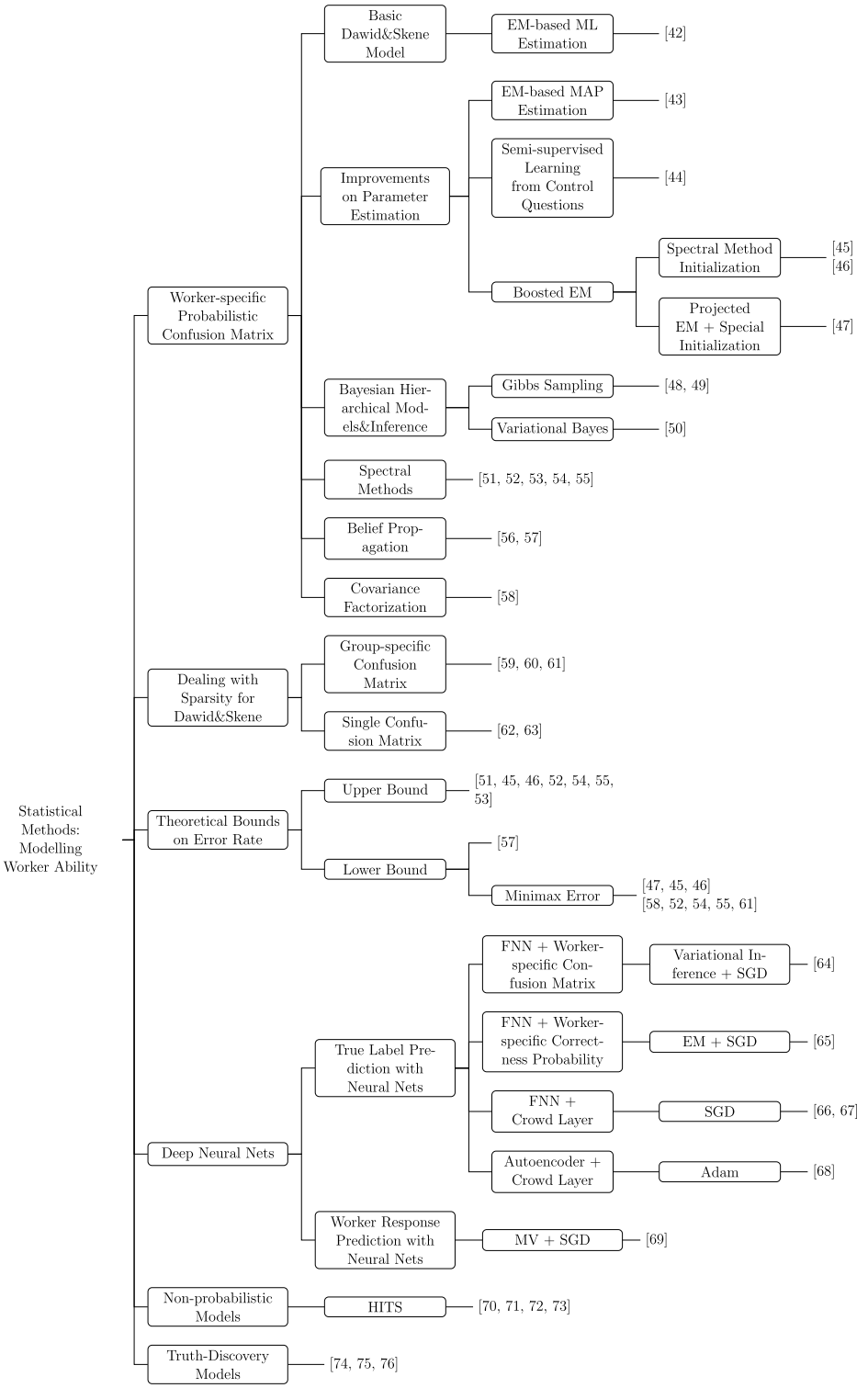


Fig. 4. A taxonomy of QCC papers that only considered worker ability/expertise to determine response quality. They focused on statistical modelling and inference.

4.1. The Dawid & Skene model

The work that first introduced a model for estimating worker ability was that of Dawid & Skene [42], which we refer to as the DS model. It deals with the scenario where a worker i reads a question j which has underlying true answer l_j, and then gives her 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. 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: $\pi_{ikk} = P(r_{ij} = k | l_j = k), k \in \mathcal{K}$. Each off-diagonal entry $\pi_{ikk'}$ records the probability of the response being incorrect: $\pi_{ikk'} = P(r_{ij} = k' | l_j = k), k \neq k'$. Since the k-th row of the confusion matrix stores conditional probabilities, the entries must sum to one $\sum_{k' \in \mathcal{K}} \pi_{ikk'} = 1$.

The DS model adopts the expectation-maximization (EM) algorithm to perform maximum likelihood (ML) estimation over all the worker responses. The ML estimation finds the locally optimal estimates for the model parameters: both the probabilities in the worker-specific confusion matrices and the true answers for the questions. In this case, the aggregation for inferring the true answers 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 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:

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

In Eq. (1), $\hat{l}_j$ is the estimate of the correct answer l_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 $\{\hat{\rho}_{jk}\}_{k \in \mathcal{K}}$:

$$\hat{\pi}_{ikk'} = \frac{\sum_{j \in \mathcal{J}_i} \hat{\rho}_{jk} \mathbb{1}\{r_{ij} = k\}}{\sum_{k' \in \mathcal{K}} \sum_{j \in \mathcal{J}_i} \hat{\rho}_{jk'} \mathbb{1}\{r_{ij} = k'\}} \tag{2}$$

$$P(\hat{l}_j = k) = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \hat{\rho}_{jk} \tag{3}$$

In Eq. (2), $\mathcal{J}_i = \{j | (j \in \mathcal{J}) \wedge (r_{ij} \neq ?)\}$ is the set of questions answered by worker i. In Eq. (3), $|\mathcal{J}|$ is the total number of questions.

There has been simplification in some of the subsequent work to only consider binary response options. This means the confusion matrix specific to each worker is reduced to only two free parameters (i.e. the diagonal entries). Such a simplified model is called a *two-coin* DS model [77]. If the two diagonal entries 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 [45].

4.2. Improvements on parameter estimation

As an improvement on the maximum likelihood EM estimation for the DS model, Snow et al. [43] employed the maximum a posteriori (MAP) estimation for the parameters. Later, Tang and Lease [44] leveraged control questions for improving the ML estimation. This was achieved via semi-supervised learning based on the true answers of the control questions to refine the parameter estimation for the DS model. Zhang et al. [45,46] proposed to use spectral methods to initialize the EM algorithm to escape local optimum in the search for the optimal true answer probability estimates and confusion matrix estimates. Gao and Zhou [47] 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 customised the initialization procedures for the projected EM to avoid local optimum.

Instead of point estimation based on EM, Kim and Ghahramani [49] and Carpenter [48] 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. [50] applied variational Bayesian inference to efficiently estimate the joint probabilities of the worker confusion matrices and the question true answers. The authors further extended the frameworks with dynamic worker confusion matrices and adapted the variational lower bound accordingly.

Ghosh et al. [51] proposed a spectral algorithm that decomposes a question-question matrix capturing response correlations across questions 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. [53] proposed spectral methods that focus on decomposing worker-worker matrices capturing response correlations across workers. Karger et al. [52] applied spectral decomposition explicitly to the worker-question response matrix. Recently, Bonald and Combes [58] factorized the response covariance matrix into one-coin worker abilities and developed a non-iterative algorithm that allows for real-time ability estimation. The algorithm leverages the current most informative pair of workers with the highest covariance and their respective covariance with a target worker to estimate the target worker's current ability.

Karger et al. [54,55] combined the previous spectral analysis with belief propagation algorithms which achieved nearer optimality for true answer estimation. Liu et al. [56] applied full belief propagation (BP) to both one-coin and two-coin DS models. They found that the efficacy of BP depends heavily on the choice of prior distributions over worker ability variables. In comparison, Ok et al. [57] 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.

4.3. Dealing with sparsity in DS

The main drawback of the DS model is its vulnerability to sparsity in the responses from workers. When the number of responses per worker is small, the confusion matrix for each worker cannot be estimated reliably. 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 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. [59] 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 via Bayes' rule based on the group confusion matrix for the cluster to which the worker belongs. In a similar setting, Imamura et al. [61] proposed to group crowd-workers' responses to handle a special case of the response sparsity problem in crowdsourcing, which they referred to as *laissez-faire* crowdsourcing. In it, only a few workers answer a large number of questions while most workers answer very few questions. Instead of using the Bayes' rule to smooth worker confusion matrices, the authors directly replaced them with the corresponding group confusion matrices to form the log-likelihood function to be maximized over the responses. Both the models are parametric in the sense that they require the number of communities to be set up in advance.

In [60], 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 [62] and Kamar et al. [63] have developed statistical models for an extreme case where 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.

4.4. Theoretical bounds on error-rate of DS estimation techniques

A 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 responses increases. Among them, Ghosh et al. [51] first derived the upper bound for the error rate of a spectral inference method for true answer prediction. The technique considered binary responses under the one-coin DS model and assumed that each crowd-worker has answered a large number of questions. With the same setting, Gao and Zhou [47] showed the global maximum likelihood estimator follows a minimax lower bound with respect to the error rate, and their projected EM algorithm theoretically can achieve nearly that rate. Imamura et al. [61] further relaxed the uniform prior distribution assumption on the true answer probabilities in [47] and they derived a generalized minimax lower bound that can be applied to any type of the DS models.

Changing the setting by allowing each worker to answer just a few (rather than many) questions, Zhang et al. [45,46] proved their proposed EM with spectral method initialization yielded a tighter upper bound than that of [51] and was faster to achieve the minimax error rate than [47]. Later, Bonald and Combes [58] showed that their non-iterative algorithm can match an even stricter lower bound on the minimax error rate than the previous work. Karger et al. [52] 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, they [54,55] showed their framework based on belief propagation methods yields a tighter upper bound than [52] 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 in [53]. In [57], 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. [78] established both the lower and the upper bounds of the error rates that match exactly the exponential rates under the setting in [54,55].

Despite their theoretical soundness and empirical feasibility, current 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 [79]. Its inference framework adopted the same setting as [54,55] but extended the binary response options to multiple response options. They 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. [56] 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 provided that certain parameter estimation methods are used for true answer estimation.

4.5. Neural network approaches

Deep Learning approaches [80] have become very popular over the last few years in machine learning applications. Recently, research work on leveraging deep neural networks for quality control in crowdsourcing starts to emerge. The main idea of some of these works is to use deep neural networks to predict the true labels of items (e.g. images). This prediction can be either imposed as a prior distribution/regularization term in the MAP estimation of the log-likelihood over worker responses [64,65], or integrated into a larger deep neural framework which learns a complex mapping between the true labels and their associated worker responses [66,67]. In both cases, worker ability was considered, and each item was encoded as an input feature vector to the network.

In [64], the Bayesian hierarchical DS model [50] was extended by replacing the Dirichlet prior distribution on the true label probabilities with a feedforward neural network (FNN). This network takes in the feature vector (e.g. image pixel values) of each item and predicts their true labels through a softmax operator. Then, according to the DS model, responses of crowd-workers are generated from the categorical distributions conditioned on the true label predictions and the confusion matrices of the workers. The estimation of the model parameters alternates between two stages: variational inference for the posterior distributions of the true labels and the confusion matrices, and stochastic gradient descent (SGD) for the network parameters given the current true label estimates.

The same FNN-based prior setting for the true labels was adopted in [65] as well. The difference lies in the likelihood setting for which this work uses the one-coin DS model to only consider the correctness probability of each worker response. This probability is computed as a logistic function which takes in a dot product between a worker-specific expertise vector and a latent representation vector of an item. The latent representation vector is a low-dimensional embedding of the item's raw feature vector. The model parameter estimation in this work uses the EM algorithm to obtain the MAP estimates for the true labels, the workers' expertise vectors and the embedding parameter matrix. The FNN parameters were estimated through RMSprop minimization [81] with the SGD.

In [66], the authors proposed a deep neural framework which consists of three types of layers: the neural net (NN) layer, the ground-truth layer, and the crowd layer. In the paper, the NN layers include convolutional layers followed by fully connected dense layers. They perform a series of non-linear transformation on the input feature vectors of items and the outcomes are fed into a ground-truth layer to predict each item's true label probabilities. These probabilities are then input into a single crowd layer. This layer consists of many *parallel sub-layers* corresponding to each crowd-worker. A sub-layer is fully connected to the previous ground-truth layer and learns a worker-specific mapping from an item's true label probabilities to the corresponding worker's responses. Therefore, the number of sets of parameters that need to be optimized for the crowd layer equals the number of crowd-workers. Each set of weight parameters captures the abilities of individual workers. The entire network was trained using the SGD which minimizes the cross-entropy loss.

In [67], a similar deep neural framework was adopted which contains all the three types of layers. However, in this work, the ground-truth layer no longer connects to the crowd layer but the other way around. More specifically, the NN layers and the crowd layer both connect to their own ground-truth layers. A crowd layer takes in the one-hot vectors that encode the labels given by each worker to an item. The ability of a worker is now characterized by a dedicated weight matrix (equivalent to a real-valued confusion matrix). A one-hot vector gets dot product with the weight vector of the associated worker that corresponds to its encoded label. All the dot-product results are summed together before fed into a softmax function to obtain the aggregated true label probabilities. A key idea in the paper is to make the worker response aggregation using the crowd layer agrees with the data classifier using the NN layers on the true label prediction for the same item. The agreement is measured by the mutual information between the two counterpart ground-truth layers. This mutual information was maximized by tuning the weights between the layers using the SGD.

In [68], the above deep neural framework was implemented as a deep denoising autoencoder [82] augmented with a crowd layer. In this case, the NN layers consist of the encoder and the decoder layers of an autoencoder. The inputs to the encoder are one-hot vectors for items (in this paper, sentences) and the decoder layers try to reconstruct these one-hot vectors as accurately as possible. The embedding layer in the middle of the autoencoder serves as the ground-truth layer which, in the paper, predicts the true sentiment of each sentence. Apart from connecting to the decoder layers, the embedding layer also connects to the crowd layer which tries to predict the sentiment judgment from individual

crowd-workers. Therefore, the autoencoder has done reconstruction for both the sentences and the workers' sentiment judgments.

All the previous work has focused on using neural networks to predict item true labels. In [69], the authors applied the feed-forward neural networks to predict responses given by individual crowd-workers rather than the true labels. In this case, the neural network takes in the feature vector of an item and predicts the multiple responses given by the individual crowd-workers who have labelled that item. The authors then trained averaging logit models to predict the ability of each worker independently from training the neural network. To predict the true label of a new item, the authors use the workers' ability estimates to weigh their corresponding responses predicted by the neural network. This forms a weighted majority vote on the item's true label.

4.6. 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 response to a question, the higher the quality of that response will be. Likewise, the more high-quality responses provided by a 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 [83]. In this case, the inference of the worker ability and the quality of responses has been conducted in similar ways to HITS in [70–73]. Inferring a true answer is done by aggregating all the responses to a question weighed by the responses' respective quality estimates. Alternatively, the weights can be the difference between each response and the true answer estimate [70].

4.7. Truth-discovery worker-ability models

Research on *truth discovery* from different (possibly unreliable) information sources [74] shares similar modelling characteristics to the QCC methods. Each source of information (equivalently a crowd-worker) is associated with a reliability variable, called a weight, which measures the quality of the claim (i.e. a response) made by the source about an object (i.e. a data item). 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 [74]. Thus, the details of these models will not be covered in our survey. A few QCC methods have adopted the idea of distances in truth discovery when handling tasks over ordinal or continuous responses such as object counting [76], and percentage annotation [75].

5. Modelling worker expertise and question difficulty

More sophisticated QCC methods consider not only the worker ability but also the question difficulty. The assumption is that some questions are intrinsically more difficult than others and thus are expected to receive less reliable responses. These methods have been summarized in the taxonomy depicted by Fig. 5.

5.1. The GLAD model

Based on the above assumption, some QCC models have taken the question difficulty into account alongside the worker expertise for estimating the quality of each response [84,89,85,90]. They model the quality of a response as the *probability of it being correct*. The most fundamental work in this area is the *GLAD* model, in which a logistic function δ_{ij} is used to represent the probability that response r_{ij} is correct:

$$\delta_{ij} = P(r_{ij} = l_j | e_i, d_j) = \frac{1}{1 + \exp(-e_i / \exp(d_j))} \tag{4}$$

where e_i is a real-valued parameter 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 worker towards certain (possibly incorrect) responses, GLAD only models the probability of the correct response and ignores any biases by assuming their corresponding probabilities to be uniform as $\frac{\delta_{ij}}{|\mathcal{K}|-1}$. Here, $|\mathcal{K}|-1$ is the number of incorrect responses.

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:

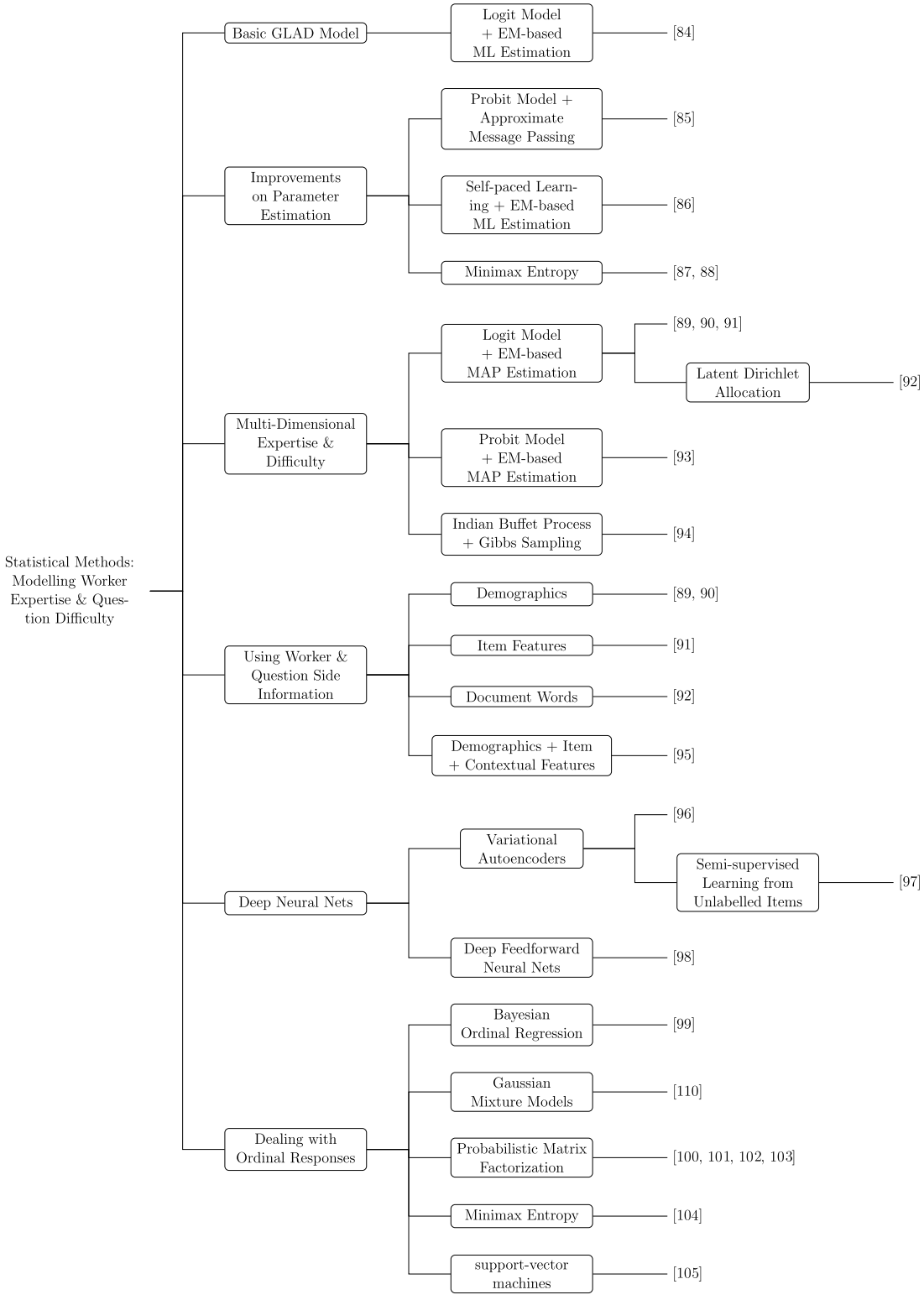


Fig. 5. A taxonomy of QCC papers that considered both worker expertise and question difficulty to explain response quality. They focused on statistical modelling and inference.

$$\hat{\rho}_{jk} = P(\hat{l}_j = k | \mathcal{R}_j, \{\hat{\boldsymbol{e}}_i\}_{i\in\mathcal{I}_j}, \hat{d}_j) = \frac{\prod_{i\in\mathcal{I}_j} \delta_{ij}^{\mathbb{1}\{r_{ij}=k\}} \frac{1-\delta_{ij}}{|\mathcal{K}|-1}^{\mathbb{1}\{r_{ij}\neq k\}} P(\hat{l}_j = k)}{\sum_{k'\in\mathcal{K}} \prod_{i\in\mathcal{I}_j} \delta_{ij}^{\mathbb{1}\{r_{ij}=k'\}} \frac{1-\delta_{ij}}{|\mathcal{K}|-1}^{\mathbb{1}\{r_{ij}\neq k'\}} P(\hat{l}_j = k')} \tag{5}$$

In the M-step, the expected joint log-likelihood over the observed responses and unobserved true answers with respect to $\hat{\rho}_{jk}$ is maximized over the rest of the parameters. The expected joint log-likelihood Q is formulated as follows in GLAD:

$$Q(\{e_i\}_{i\in\mathcal{I}}, \{d_j\}_{j\in\mathcal{J}}; \mathcal{R}, \{\hat{l}_j\}_{j\in\mathcal{J}}) =$$
$$\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \hat{\rho}_{jk} \log\left(P(\hat{l}_j = k)\right) + \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}_j}\sum_{k\in\mathcal{K}} \hat{\rho}_{jk} \log\left(\delta_{ij}^{\mathbb{1}\{r_{ij}=k\}} \frac{1-\delta_{ij}}{|\mathcal{K}|-1}^{\mathbb{1}\{r_{ij}\neq k\}}\right)$$

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

5.2. Improvements on parameter estimation

Many later QCC methods have followed the main idea in GLAD [84]: the quality of a response being a probabilistic function over variables representing the worker ability and the question difficulty. In [85], the probit model is used where the logistic function is replaced by the Gaussian cumulative function with the mean being the expertise e_i minus a question-specific bias term, and the variance being the difficulty d_j. Additionally, instead of using the EM algorithm for the ML estimation, this work employs approximate message passing inference on the model parameters.

In [86], the authors extended the GLAD model by introducing self-paced learning [106] into the EM-based ML estimation. The authors argued that the self-paced learning can (1) prevent GLAD from being trapped in a undesirable local optimum, and (2) automatically determine which responses the EM algorithm should use to update the model parameters. For the second argument, the authors assume that the importance of a response is proportional to its associated part of the likelihood under the current model parameter estimates. This particular part of the likelihood being higher means that the response becomes more important and should be given more weights in the updates to the parameter estimates in the next round of EM. Furthermore, the self-paced learning also sets up a weight threshold which controls the pace of training by cutting off unimportant responses. The threshold decays as the training goes further, meaning that gradually, more responses will be chosen to train the GLAD model as it becomes robust enough to account for more variance in the responses.

Zhou et al. [87] 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. [88] extended their original work by regularizing the minimax optimization with relaxed constraints to prevent its response probability estimates from overfitting sparse responses.

5.3. Multi-dimensional worker expertise and side information

More recent work has extended the worker-question interaction to be *multi-dimensional*. They argue that workers can have their own areas of expertise and questions can be associated with the different areas. The authors of GLAD exploited this idea by simply converting the worker expertise and the question difficulty into vectors. Correspondingly, they converted the original scalar product into a dot product between these vectors in [89,90].

Ruvolo et al. [89] was also the first work to leverage side information of both crowd-workers (i.e. their demographics) and questions (i.e. the features of the data items) for further improving the parameter estimation. 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. Similar work was done in [93] which used the Gaussian cumulative function to represent the response quality with the mean being the dot product specified in [89] minus a worker bias term.

In [94], rather than setting up the dimension for the expertise and the 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 [107]. This work applied Gibbs sampling to infer the model parameters including the true answers.

In [91], convex optimization techniques were proposed for training worker-specific binary classifiers which took side information features about questions (indicating question difficulty) into account. To allow for the multi-dimensionality of the question features, these classifiers are endowed with weight (expertise) vectors in the following logistic function:

$$\delta_{ij} = P(r_{ij} = l_j | \boldsymbol{w}_i, \boldsymbol{x}_j) = \frac{1}{1 + \exp(\boldsymbol{w}_i^T \boldsymbol{x}_j)} \tag{6}$$

In Eq. (6), $\boldsymbol{w}_i$ 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 and the MAP estimate of the mean vector serves as the weight vector of a base classifier to estimate question true answers using Eq. (5).

In [92], the multi-dimensionality of the expertise was estimated in a topic-wise manner 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 model used Latent Dirichlet Allocation (LDA) [108] 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.

In [95], the authors proposed to extend the GLAD model by incorporating various types of side information to improve the parameter estimation when responses are scarce. The side information features concern worker demographics, question content and contextual information such as devices (e.g. PC, mobile phone), browsers, time periods of each session, time duration and orders of each response and so on. The worker and question features are incorporated into the GLAD model as:

$$\delta_{ij} = P(r_{ij} = l_j | e_i, d_j, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{1 + \exp\big(-(e_i + \boldsymbol{\alpha}^T \boldsymbol{x}_i)/\exp(d_j + \boldsymbol{\beta}^T \boldsymbol{x}_j)\big)} \tag{7}$$

The dot product $\boldsymbol{\alpha}^T \boldsymbol{x}_i$ forms expertise offsets across the workers with the global coefficients $\boldsymbol{\alpha}$ learned to bring the offsets of similar workers closer together. This helps to smooth the irregular expertise estimates that result from the sparse responses across workers. This effect is also applied to the dot product $\boldsymbol{\beta}^T \boldsymbol{x}_j$ for calibrating the difficulty d_j. The context features were further divided by the authors into the session-level features and the response-level features. The session-level features were incorporated into the expertise factor of the GLAD model same as the worker features but with different global coefficients. The response-level features are incorporated as $(\delta_{ij} + \boldsymbol{\eta}_i^T \boldsymbol{x}_{ij})$ where δ_{ij} is derived from Eq. (7), $\boldsymbol{\eta}_i$ are worker-specific coefficients, and $\boldsymbol{x}_{ij}$ are the feature values regarding response r_{ij}. The coefficients $\boldsymbol{\eta}_i$ form a local linear regression over the feature vectors of all the responses made by worker i. Such local regressions addressed worker-specific biases that the GLAD model failed to handle properly [93].

5.4. Neural network approaches

Apart from the work reviewed in Section 4.5 which only considers worker ability, there is other work additionally considering question difficulty. Yin et al. [96] used variational autoencoders [109] to map responses to each question into latent true answer distributions. 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 question. Both the encoder and the decoder are implemented as single-layer networks and the global weight vector for each layer accounts for the biases across the workers towards different response options. In addition to the weight vectors, a question-specific scalar term is incorporated at each layer to account for question difficulty. It does this by scaling the layer outputs before fed into a softmax transformation.

Atarashi et al. [97] leveraged semi-supervised learning and variational autoencoders to facilitate true answer inference. They used 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 [96]), they 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 [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. Then, the accuracy and difficulty estimates are segmented into different levels (e.g. "low" and "high" levels of accuracy/difficulty). 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 the 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. The input to the k-th network consists of the response probability of each worker. The response probability equals the output from the first stage, i.e. the correctness probability $P(r_{ij} = l_j)$, if workers' responses are the particular option k. Otherwise, the probabilities equal $\frac{1 - P(r_{ij} = l_j)}{|\mathcal{K}| - 1}$. The output from this network is normalized as $P(l_j = k)$ for each question.

5.5. 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). In this case, it is natural to measure the

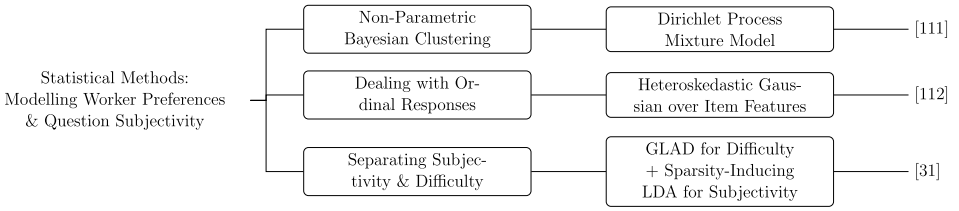


Fig. 6. A taxonomy of QCC papers that considered worker preferences and question subjectivity. They focused on statistical modelling and inference.

distance between each worker's response and the corresponding correct answer. This distance directly reflects the quality of the response, which can be modelled by a Gaussian density function. In this function, the mean 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.

In [99], the above framework was followed by 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 [110], the framework was further extended to be a worker-specific Gaussian mixture with each Gaussian component acting as a "soft" intercept in the ordinal regression. This model deals with the situation where workers show their individual biases towards different response options.

In [100] and their subsequent work [101–103], 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 variance is always set to be one. Sharing the same idea with [91], 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 mean vector over all the workers' latent vectors. Zhou et al. [104] adapted the minimax entropy principle from their previous work [87] to make it compatible with ordinal responses by modifying the constraints to account for the natural ordering in the response options.

In [105], the authors first converted the $|\mathcal{K}|$ ordinal class prediction problem into $|\mathcal{K}|-1$ binary classification problems. For each binary problem, they exploited the idea of support-vector machines which characterize the classification boundary for the binary classes using a separating width between the questions' true answer variables. The authors then defined a linear decision boundary dependent on the worker expertise and the true answer variables which approximates the underlying classification boundary. They optimized the position of the boundary in such a way that the separating width between the two classes can be maximized. They also introduced question-specific slack variables (i.e. question difficulty) to relax the optimization so that true answers for difficult questions are allowed to be misclassified to some extent.

6. Modelling worker preferences and question subjectivity

In crowdsourcing, some tasks might contain partially subjective questions. These questions possess more than one correct answer and at least one incorrect answer. To answer such a question correctly, a worker needs to avoid any incorrect answers, which depends on her ability/expertise and the difficulty of the question. Meanwhile, her subjective preferences/opinions on different (subjective) features of the question will cause her to prefer one of the correct answer options over the others.

To the best of our knowledge, very few QCC methods have endeavoured to distinguish between question difficulty and subjectivity in one unified model (see Fig. 6). In [111], the authors assumed that a higher joint degree of difficulty and subjectivity for an entire 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 group becoming smaller. The authors proposed to infer the response groups using a Dirichlet Process Mixture Model [113]. Despite attributing the response variation to both difficulty and subjectivity, the paper made no attempt to separately model the two even though they might induce different types of interactions with workers.

In [112], the authors focused on ordinal ratings given to partially subjective items with observed features. The rating r_{ij} is assumed to follow a Gaussian distribution with the mean and the variance linearly regressed on the observed features $\boldsymbol{x}_j$ of item j as $r_{ij} \sim \mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}_j, \exp(\boldsymbol{v}^T\boldsymbol{x}_j))$. Here, $\boldsymbol{w}$ and $\boldsymbol{v}$ are global coefficient vectors. The variance encodes the mixing effects of the subjectivity and difficulty of the item. Thus, this method is also not intended to separate the modelling of the two properties.

Recently, Yuan et al. [31] proposed the first QCC model that separates the difficulty and subjectivity. It replaces the single truth variable l_j for a question j with a subjective truth variable l_{ij} specific to each response r_{ij}. The subjectivity of question j is captured by factorizing l_{ij} into another latent variables that represent worker i's preferences and the question's features. The difficulty of the question is directly encoded as a variable d_j which counteracts expertise e_i in the following logistic function:

$$\delta_{ij} = P(r_{ij} = l_{ij} | e_i, d_j) = \frac{1}{1 + exp(-(e_i - d_j))} \tag{8}$$

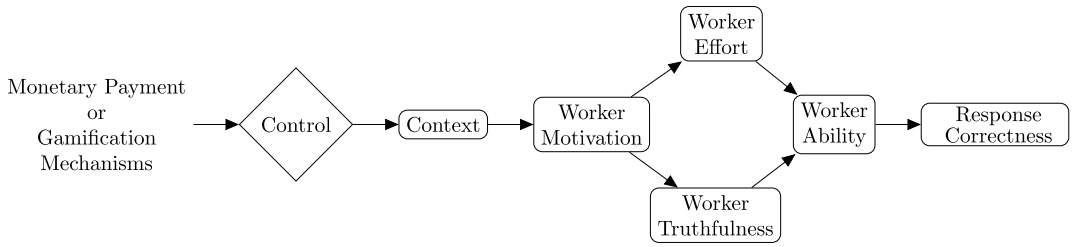


Fig. 7. A diagram shows payment and gamification mechanisms control worker contexts to affect worker motivation, which further influence worker effort and truthfulness, and eventually the correctness of responses.

To prevent the preference-feature factorization of l_{ij} from explaining all the response variation, the authors further replaced the preference vector of each worker with much sparser topic-preference vectors. These vectors are associated with each worker via an LDA. In this case, each worker has a distribution over a finite set of interested topics. A topic is associated with a topic-preference vector to be used in the preference-feature factorization for each l_{ij} via topic-response assignment. This work is also the first to propose a measure of subjectivity in crowdsourcing which is the *expected number of correct answers with respect to underlying groups of workers*. The worker groups were clustered using K-means with the Elbow method over the topic distribution of each worker.

7. Modelling worker motivation and worker contexts

Different incentive mechanisms are designed to favour different characteristics of worker motivation. According to [114], 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* mechanisms which are respectively responsible of motivating the workers extrinsically and intrinsically (see Fig. 7).

The monetary payment mechanisms are mostly theoretical mechanisms. They are extended from the DS model under different assumptions (e.g. one-coin assumption). These assumptions capture workers' efforts and their (probabilistic) beliefs on the correct answers. There are two application scenarios for these mechanisms. One is when crowd-workers are unmotivated and try to carelessly rush through all the questions to get the rewards. The other is when crowd-workers can lie about their responses or collude with one another in order to obtain higher rewards. The payment mechanisms we are about to review can theoretically guarantee that these scenarios will not happen under certain conditions.

The gamification mechanisms are mostly practical mechanisms. Despite their heavy reliance on various game elements, these mechanisms are driven by statistical estimation of workers' expertise/accuracy. The estimation is conducted over worker responses (1) using QCC aggregation models (e.g. majority vote) and (2) with control questions. The estimates can be directly fed back to the crowd-workers to motivate them to perform better. They also determine how much rewards (e.g. scores, badges, ranks) should be given to them accordingly.

7.1. 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 [141–143]. 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 [135,143]. 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 responses that they elicited with effort (e.g. by flipping their responses) if they believe that doing so could result in higher payoffs [134].

To address the above issues, 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) (see Fig. 8). The aim of the adaptiveness is to ensure that the expected payoffs of the workers are maximized only when they exert sufficient effort to produce high-quality responses and report these responses truthfully.

7.2. 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 (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. The accuracy of the responses to those control questions serve as the inputs to some carefully designed *payment*

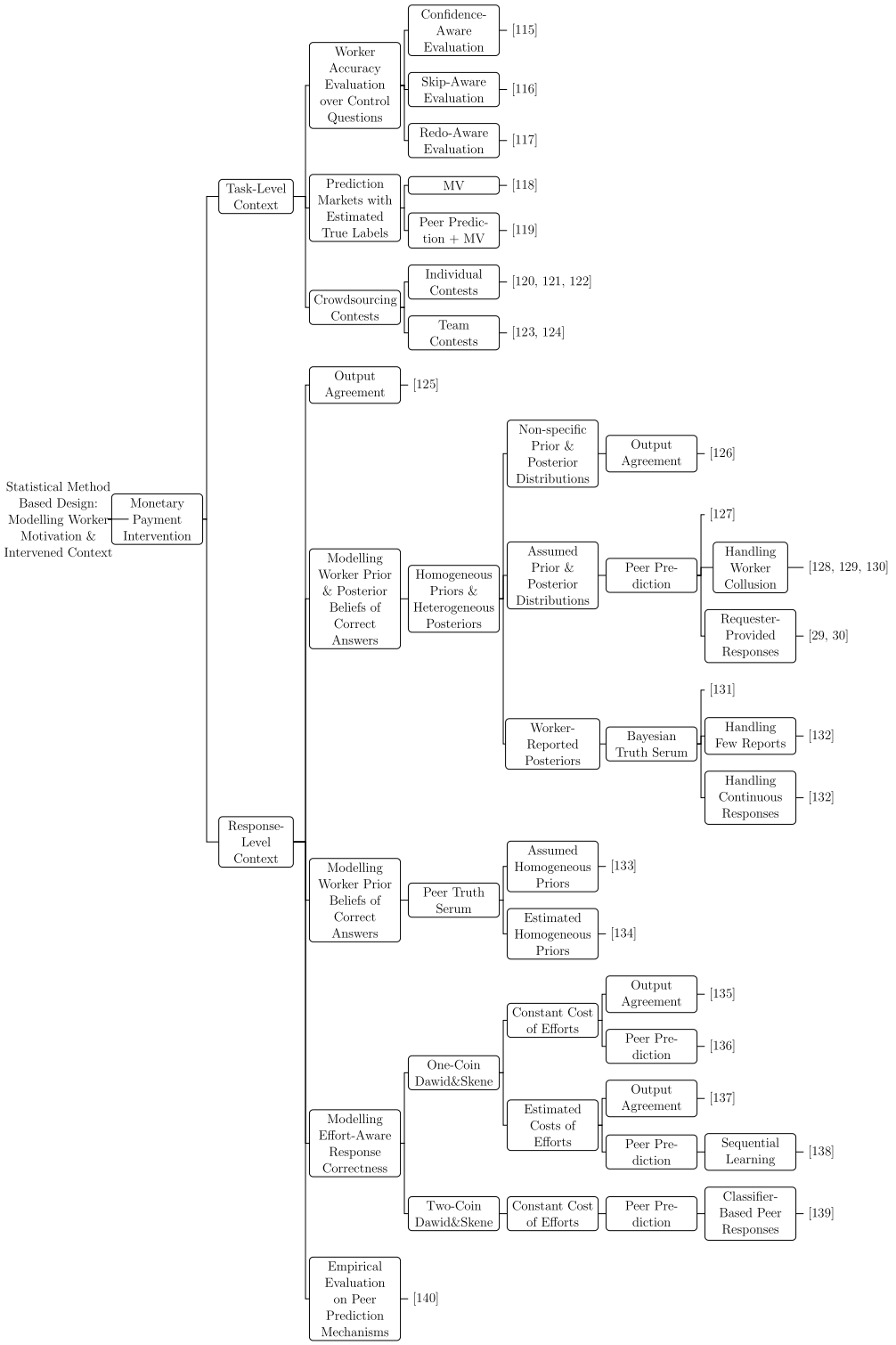


Fig. 8. A taxonomy of QCC papers that considered the interaction between worker context and (extrinsic) motivation. They focused on designing monetary payment mechanisms which rely on statistical modelling (and possibly estimation) of worker attributes.

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 [115], the 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. The accuracy estimate is then used to calculate the payment to the worker. In [116], 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 [117], the authors proposed a two-stage design 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. At both stages, the payment function evaluated over the worker's responses to the control questions ensures her truthful reporting.

Other than using control questions, some have approached the payment problem from different angles. In economics, prediction markets [144] use ground-truths, revealed as outcomes of targeted events, to compute rewards for participants in the markets. The participants can buy and sell contracts which will pay off certain amounts dependent on the ground-truths later revealed. Each contract corresponds to a possible outcome of an event. Market prices of the contracts reflect participants' consensus on the ground-truth probability distribution. Purchasing a contract raises its market price and hence increases the consensus probability that its corresponding outcome is the ground-truth. Every purchase also brings a cost which is modelled by various payment functions in the prediction-market literature [145,146]. Under certain conditions, these functions can achieve truth-telling equilibria in which participants maximize their expected payoffs by truthfully reporting what they believe the ground-truths are.

In the setting of crowdsourcing labels for machine learning systems, prediction markets are not directly applicable as true labels are hardly available in this case. Some recent prediction-market papers proposed to use inferred true labels as the proxy [118,119]. In these papers, once the prediction market was closed, workers who may or may not have participated in the market later observed the ground-truths and reported them with certain intentions. Aggregating these reports (e.g. using majority vote [118]) yields ground-truth estimates which can then be used to reward those who traded in the markets. These papers focus on designing payment functions for either the prediction markets [118] or the followed ground-truth reporting [119]. The former functions mitigate the *outside-market* manipulation from those who both participated in the markets and observed the ground-truths, and wanted to pump up their rewards from the markets. The latter functions incentivize all who observed the ground-truths to report truthfully with tailored peer prediction mechanisms.[6] We also suggest that prediction markets can be combined with QCC models which infer each true label by aggregating the probability estimates[7] from crowd-workers in the markets. In this case, it is necessary to design new payment functions or modify functions used by the markets to alleviate *inside-market* manipulation from those who understand how QCC models work in general. For example, some might deliberately cater to the majority's responses rather than express their own opinion as they know that most QCC models assume the majority of workers are reliable.

Another line of research that involves payment mechanism design focuses on *contest-based* crowdsourcing. Such type of crowdsourcing features prominently in online platforms for product design and development, such as TopCoder[8] for software development and 99designs[9] for graphic design. On these platforms, each task published by the product requesters looks for one or more solutions submitted by the crowd-workers. The workers compete with one another, in terms of the solution quality, for an often smaller number of rewards. Another key feature of these platforms leveraged by the research is that the requesters always assess the quality of the submissions and determine the winners so that they can be paid accordingly. This means the payment mechanisms in this case were designed based on some observed ground-truths.

Theoretically, the research for contest-based crowdsourcing is well founded on the *contest theory* [147] in economics. The theory models the quality of a solution submitted by a worker as either the probability that he/she wins, or a continuous latent variable for its ranking. The research has further refined the winning probability into either winning the whole contest/task if a *winner-takes-all* strategy is applied, or winning a certain prize if multiple prizes are awarded. In all cases, the quality is computed by some function (e.g. softmax [120] or linear [121] functions) that takes in variables corresponding to several attributes of crowd-workers on contest platforms. The two predominantly modelled attributes are *expertise* and *effort*, and their interaction is often modelled by a product or addition in those functions. This means workers with higher expertise and effort (level) are more likely to win. Based on the theoretical models, each contest among crowd-workers is portrayed as a *game-theoretic auction* [122] affected by a variety of factors such as number and sizes of prizes, number of workers involved, entrance fee, etc. From this auction model, a *game-theoretic equilibrium* is formed among the workers which maximizes (1) the expected quality of the top-N solutions for the requester (buyer) and (2) the expected payoff for each worker (sellers) with respect to the payment (function) parameters. Furthermore, the research also studied how sensitive the equilibrium is to changing the payment parameters (and other parameters) [120–122].

Empirically, there have been studies on how different (combinations of) factors affect the performance of crowd-workers in terms of solution quality [148,149] and the extent of participation [150,151], etc. In terms of the payment parameters, the size of the winner's prize has been primarily studied. When increasing the prize size, both positive [150,148] and negative [149] effects, especially on the solution quality, have been observed. However, empirical investigations are still

[6] We will review the peer prediction mechanisms in Section 7.3.

[7] Each worker's probabilistic belief on the true labels can be estimated from their contract buying and selling histories in the markets.

[8] https://www.topcoder.com.

[9] https://99designs.com.au.

scarce on the effects of the number of prizes and their respective sizes on the performance of workers. Recently, the idea of *team competition* was leveraged to incentivize low-performing workers through collaborative teamwork [123,124]. Two major types of team formation strategies were investigated: balanced-team (workers automatically assigned to one of the smallest teams uniformly at random), and self-organizing-team (smaller teams merged into larger teams by the agreement between team managers). Winning a prize as a team, each worker receives a reward proportional to his/her contribution (e.g. the proportion of correctly labelled items by the worker) to the teamwork. The empirical results show that both team formation strategies have boosted the quality of workers' labels compared to individual contests.

A promising but challenging research direction for the future is to introduce worker contests with monetary rewards into *non-contest* crowdsourcing platforms such as AMT and CrowdFlower. On these platforms, it is impossible for data requesters to verify ground-truths for most tasks. Therefore, it is questionable that the game-theoretic auction models developed by the previous research is still applicable in this case. Also, no empirical study has been dedicated to understanding the effects of different payment parameters on the performance of competing crowd-workers when ground-truths are absent. Similar to the treatment for prediction markets, we suggest to combine contest-based crowdsourcing with QCC aggregation models which infer the ground-truths. It will be interesting to investigate how much discrepancy there is in the performance of crowd-workers rewarded based on the inferred and the actual ground-truths.

7.3. 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 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 comparison with the target response. If these two responses are the same, then the target worker will be rewarded according to a payment function that is carefully designed to induce a *game-theoretic equilibrium* among all the workers [135,126,133,134]. 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 data requester. Such an equilibrium is thus also referred to as a *truthful* one.

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.

Output agreement [125] 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 [126] 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. In it, workers report the correct answer according to the common part of their understanding.

Peer prediction [127] 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 together with a random peer's response to the same question to calculate the reward for the worker. 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 workers. The collusion 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 [128,129] or designing payment functions that penalize the "collusion equilibria" to make them have smaller payoffs than the truthful one [130]. However, the techniques still assume that true answers and responses are binary.

Peer truth serum (PTS) [133] is an alternative when data 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. It also uses 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 to induce at least one "non-truthful" equilibrium where all workers collude with one another. They always give the least likely responses to each question. Instead of using a predefined prior distribution, subsequent work [134] focused on dynamically estimating the prior. It was done by using frequencies of responses from other workers to the same question to which the response of the target worker was given.

Bayesian truth serum (BTS) [131] 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 [132] 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 [152] 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 QCC models such as DS and GLAD which focus 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 are 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. 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 [135], referred to as the *DG* mechanism. They dealt with binary response options based on the one-coin DS model. 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 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. [136] 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). This also means that those who participate will be qualified and invest efforts in the equilibrium.

Both Dasgupta [135] and Witkowski et al. [136] have assumed the cost induced by non-zero efforts is a constant. Within the same modelling framework, Liu and Chen [137] 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 in [138]. 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 [139] 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 [133,140]. 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 any of the above incentive mechanisms is applied to the monetary payments for workers' responses, quality control methods can be further applied to the crowdsourced responses to produce more reliable estimates of the true answers. There are also unified frameworks proposed by Frongillo et al. [29] and Ho et al. [30] that integrate the above process. More specifically, Frongillo et al. [29] combined a payment function that ensures a truth-telling equilibrium with the Bayesian statistics to allow for both truthful response elicitation and Bayesian aggregation for inferring the true answers. In [30], a

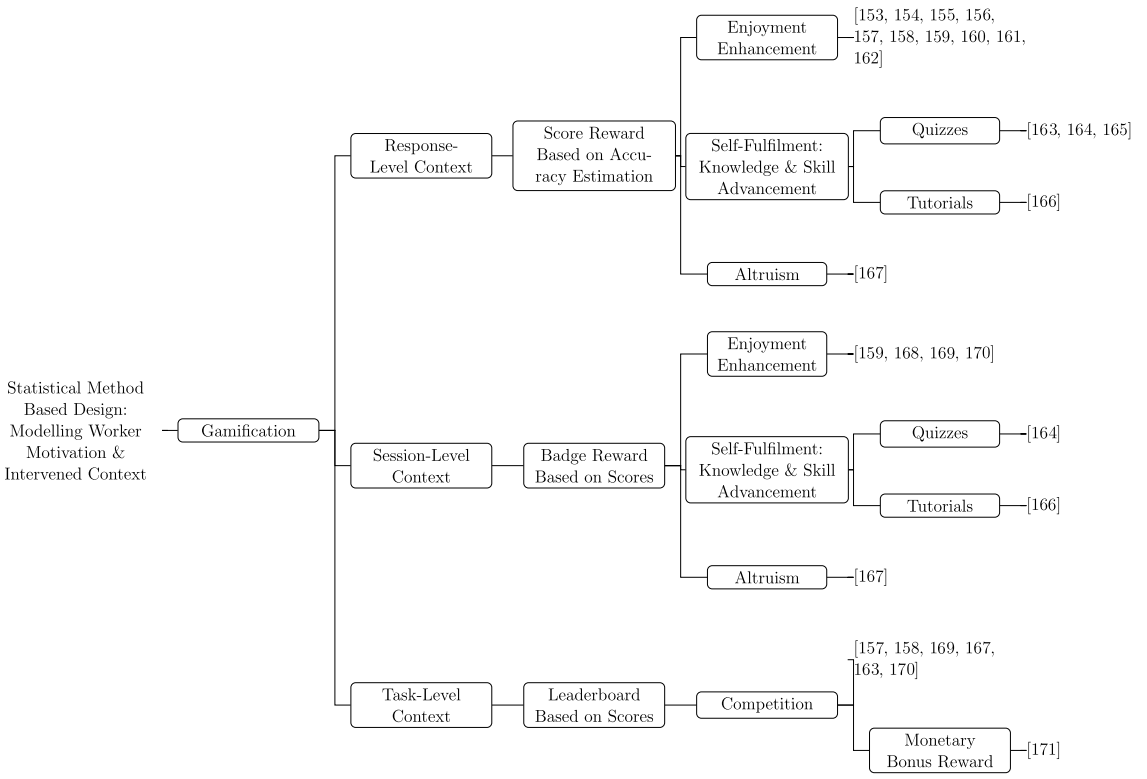


Fig. 9. A taxonomy of QCC papers that considered the interaction between worker context and (intrinsic) motivation. They focused on designing gamification mechanisms which, in most cases, rely on statistical estimation of worker ability.

further step was taken to optimize the multiple-choice interface with confidence shown to each worker (the interface being similar to the one employed in [115]) along with the optimization of the Bayesian aggregation.

7.4. 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 workers to make them more engaged in answering questions. This eventually leads to improved quality of their responses. The intrinsic motivations 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' responses. Fig. 9 shows the taxonomy that summarizes the gamified mechanisms.

7.5. 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 depending on the quality of her response. Higher-quality responses should be rewarded with more points. 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 as the building block of other more complicated game elements such as leaderboards and level systems.

In [153], 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 [154], relevance judgment in Information Retrieval [155–157], image classification [157], video captioning [158], language translation [159], and communication to the general public about culture [160] and science [161].

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 [163], 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 target questions 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 by AMT.

The main idea of the above work to use gamified quizzes to attract participation of (and contributions from) workers or volunteers has also been adopted in [164] and [165]. 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 [172] for engaging online players to provide answers to questions. The answers were used to train classifiers that can perform better automatic 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 [166], 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 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 compared to those produced by the originally novice workers.

Scoring mechanism can also help incentivize workers to make altruistic contributions to their communities. In [167], the focus is on motivating workers to contribute to the construction of a new online 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.

However, we have also found an exception in which the scoring mechanism failed to motivate participants to perform better. In [162], the authors developed a two-player object recognition online game in which one player (master) marks an object in the image as the target while the other player (seeker) tries to find the target object based on the hints provided by the master. The scoring mechanism, in this case, sets up a positive score and reduces it in front of both players every time the seeker clicks on the wrong object. The game is over when either the seeker clicks on the target object or the score becomes zero. It turned out that this scoring mechanism demotivated the participants as their performance (measured by the average number of wrong clicks) was significantly better when the scoring was removed. We conjecture that this negative result is due to the score decreasing policy of the scoring mechanism which has very much pressured the participants.

7.6. 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 tasks they have successfully completed. In a gamified task, badge awarding usually happens in session-level contexts. More specifically, when a 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 for various motivational purposes. They range from making laborious and tedious work (such as image annotation [168], proofreading [169], language translation [159] and mobile application testing [170]) more enjoyable, confirming one's learning progress, (e.g. on image editing [166] and enterprise knowledge [164]), to encouraging workers' commitment to building online communities [167].

7.7. 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 workers at any time 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 [157], steady improvements were observed in the quality of workers' relevance judgments for documents to search queries. Quality was measured in terms of the level of agreement between workers on the same document-query pairs. In [169,158], 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 [167], 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 [163], 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 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 in [170,171]. In [170], 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. In [171], workers were exposed to a task leaderboard when performing relevance judgment [173]. The authors found that providing such a leaderboard could significantly improve the quality of responses (compared to no leaderboard) only when the workers were informed of a top-10 bonus reward with one dollar per person. They also observed that considering the number of responses made by each worker thus far in the rank calculation resulted in lower accuracy across workers (than considering the percentage of correct responses). They argued that this indicated a discouraging factor of the task leaderboard.

To deal with the above issue, Ipeirotis and Gabrilovich [163] 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. As a result, workers who arrive late at a task page are less likely to be intimidated by the (page) leaderboard rankings. In [167], 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.

Based on our review, we found that the majority of papers regarding gamification for crowdsourcing applications have shown positive results on boosting the quality of worker responses. However, we cannot rule out the possibility of potential publication biases which have prompted authors to only show positive results. Furthermore, not only the papers we reviewed in crowdsourcing have reported negative effects of scoring mechanisms [162] and leaderboards [163,170,171], but also those in other domains such as education [174,175] and e-commerce [176] have found mixed, if not all negative, effects of either game elements. We found that workers are more likely to be demotivated by competition, especially when competition results can be accumulated, and in general, bad design choices which can pressure or bore the workers early on. We suggest that game designers should first survey a sample of the targeted worker population (using their prototype games) on how they think particular design choices would incentivize their motivation and affect their performance. We also urge the gamification design in practice to systematically adopt more contributions from the theoretical work (e.g. crowdsourcing contests, peer consistency mechanisms).

8. 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 QCC literature, we have found the following two types of mechanisms in which the interaction takes place:

- Worker expertise is improved by *training mechanisms* deployed at different levels of contexts. These mechanisms are based on statistical models that capture workers' (domain) expertise. They include the one-coin DS model and the (multi-dimensional) GLAD model with the question difficulty variables replaced by observed question domains. These mechanisms are able to track, learn and control the learning curves of crowd-workers in such a way that their performance can be improved or maintained at some level.
- Worker expertise is leveraged by *question assignment mechanisms* to control the allocation of questions into different levels of worker contexts. These mechanisms are based on either statistics of worker expertise/accuracy or statistical models that account for it. The models are the same as those employed by the training mechanisms. These question assignment mechanisms are able to cost-effectively route unsolved questions in certain domains to workers with sufficient expertise in those domains.

Fig. 10 illustrates the taxonomy which summarizes the above two types of mechanisms that model the interaction between worker expertise and contexts.

8.1. Improving worker expertise using training mechanisms at different levels of contexts

Training mechanisms intervene in the context to directly affect the *expertise* of workers (see Fig. 11). They aim at improving workers' expertise regardless of their motivation. 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 [178] that the default training can significantly improve the quality of worker responses on a variety of crowdsourcing tasks. In [179], the authors proposed a novel teacher-learner framework which performs iterative teaching on crowd-workers who are modelled as logit learner models. The coefficient vector for a learner model characterizes the corresponding worker's expertise/skills over some observed domains of questions. Meanwhile, a teacher (agent) has access to a target coefficient vector which represents the skills it wants the workers to grasp after the teaching process.

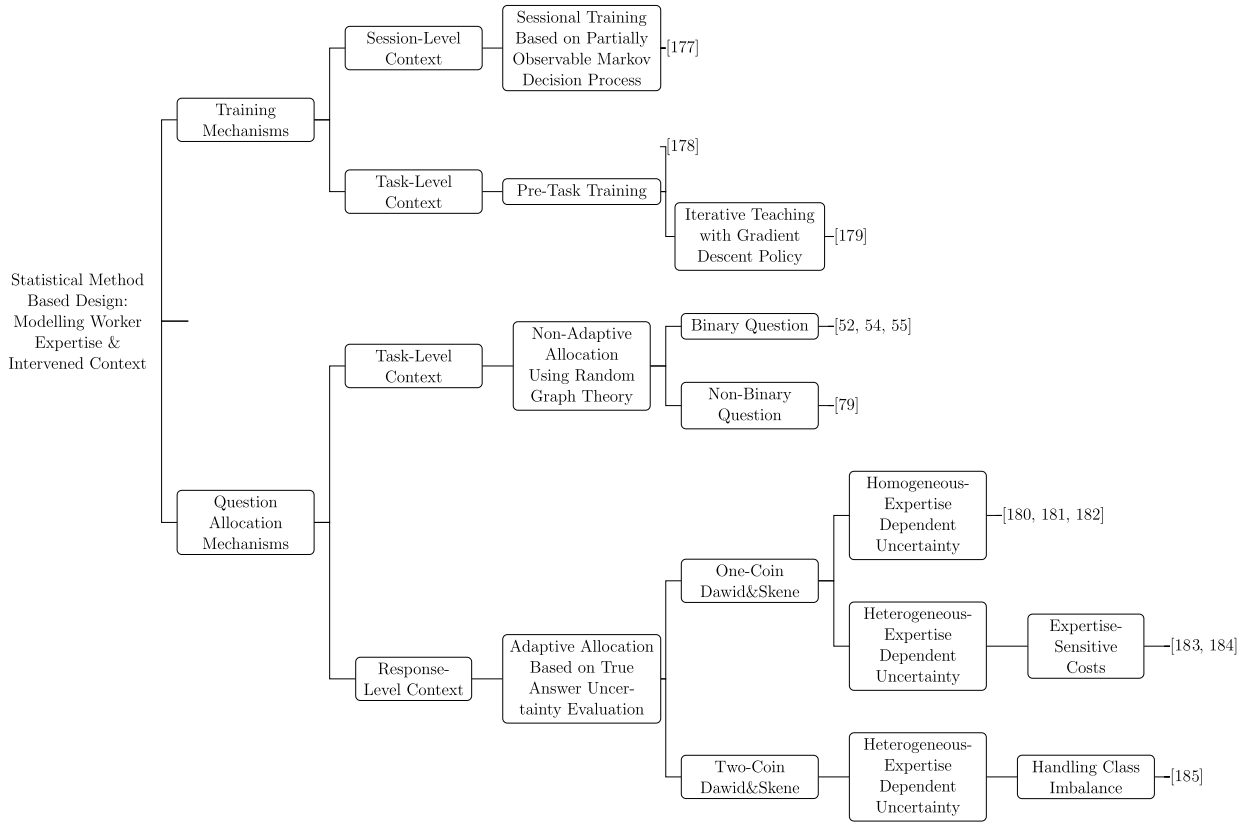


Fig. 10. A taxonomy of QCC papers that considered the interaction between worker expertise and context. These papers focus on designing either training mechanisms that alter contexts to improve worker expertise or question allocation mechanisms that use worker expertise to determine questions to be answered in the contexts.

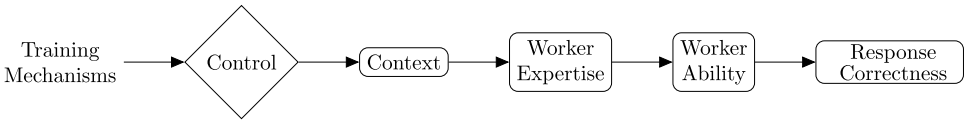


Fig. 11. A diagram shows training mechanisms control worker contexts to improve worker expertise, which further improves the correctness of their responses.

The learner models employ a gradient descent strategy to update their coefficients. The key aspect of this strategy is that the gradient descent update to a model's coefficients should depend on not only the worker's current response but also the model's update memory in the past. Moreover, by incorporating a decay factor, earlier updates will have less effect on the current update. This shows the fact that workers' knowledge is accumulated and earlier knowledge can be forgotten. At each teaching iteration, the teacher selects a (control) question which is going to give the largest descent in the gradient of a worker's coefficients towards the target coefficients. It is not until the gradient descent converges for a worker that can he or she enter the task.

The mechanisms reviewed above are restricted 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 workers is undesirable, they are not given a chance to be retrained to perform better. To solve this issue, Bragg et al. [177] proposed a training mechanism functioning in the session-level contexts. The mechanism models the decision-making process of whether to train a worker or collect responses from her in each of her working sessions as a partially observable Markov decision process (POMDP) [186]. It 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.

8.2. Question allocation in crowdsourcing

Reducing the number of responses collected using crowdsourcing to lower costs while maintaining the prediction accuracy on true answers has been an important QCC research topic over the years. A common crowdsourcing process involves

assigning each worker a set of questions which is selected uniformly at random according to some value. This value is 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.

8.3. 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. 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 worker. In [52,54,55], the authors proposed to draw a regular random bipartite graph based on the *configuration model* from the random graph theory [187]. 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[10]). 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 to achieve a target error is the error value scaled by the inverse of the expectation of each worker's expertise. In their following work [79], the authors investigated the same subjects but with non-binary questions. They derived similar results in terms of the near-optimality of the regular random graph in achieving any target error rate and the scaling effect of expertise on the cost per question.

8.4. 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. Sheng et al. [180] and Ipeirotis et al. [181] proposed to model one key aspect in the adaptive allocation, that is the *uncertainty* of each question's true answer. In their 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. In their following work [183,184], the authors addressed heterogeneous worker expertise. In this case, the entropy of the probability estimates from Eq. (1) embodies the uncertainty about the true answer. 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 [182], the authors modelled the uncertainty of the correct response l_j for question j 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 Eq. (6)). At each time 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 Eq. (6)). The selected question is then assigned to worker i who is able to maximize the probability of seeing the response r_{ij}:

$$P(r_{ij}|\boldsymbol{w}_i, \boldsymbol{x}_j) = \delta(\boldsymbol{w}_i, \boldsymbol{x}_j)^{\mathbb{1}\{r_{ij}=l_j\}}\big(1-\delta(\boldsymbol{w}_i, \boldsymbol{x}_j)\big)^{\mathbb{1}\{r_{ij}\neq l_j\}} \tag{9}$$

where $\delta(\boldsymbol{w}_i, \boldsymbol{x}_j) = \delta_{ij}$ defined by Eq. (6).

In [185], the question assignment strategies were further extended to be based on two-coin models which encode biases of individual crowd-workers towards positive and negative responses. They also devised an adaptive decision boundary for

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

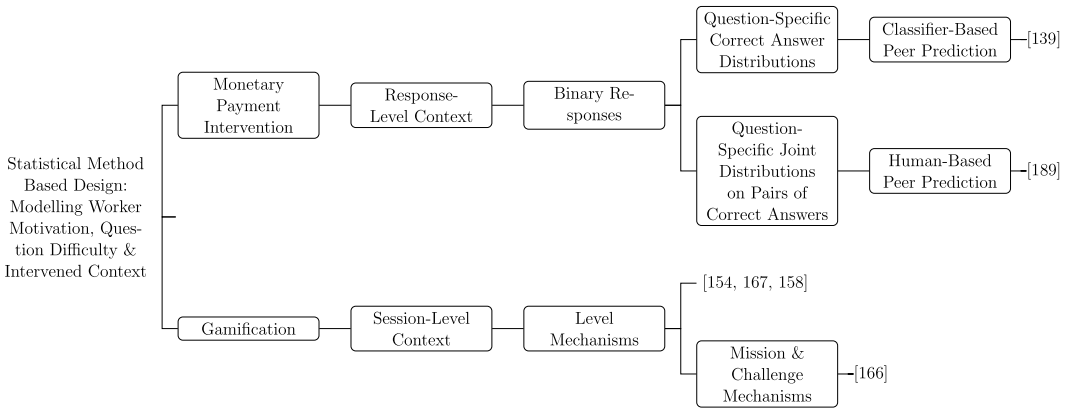


Fig. 12. A taxonomy of QCC papers that considered worker motivation, context and question difficulty. These papers focused on designing either monetary payment mechanisms which additionally modelled question difficulty or gamification mechanisms which increase question difficulty to challenge workers.

determining the true answer of each question and further, degrees of the uncertainty when class imbalance exists in the true answers. In [188], a hidden Markov model was proposed to capture the correlation in the time-varying ability of each crowd-worker. At each time step, the workers were ranked based on estimates of their current abilities and only the top worker was assigned the question to answer. Compared to the above work, this work fails to utilize all the crowdsourcing power available.

9. Modelling worker motivation, question difficulty and contexts

In Section 7, the response quality is modelled to be dependent on the worker motivation under different worker contexts. The *heterogeneity* in question difficulty was ignored in order to simplify both 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 and the gamification incentive mechanisms have successfully controlled the response quality by considering the question difficulty. In this case, the payment mechanisms are typically applied to the response-level context while the gamification mechanisms are applied to the session-level context. Both lines of research work have been summarized by the taxonomy in Fig. 12.

The payment mechanisms here are mostly extended from the work reviewed in Section 7.3. Their statistical models now consider question-specific true answer probabilities instead of a shared probability distribution. In this way, the conditions for these mechanisms to guarantee a truth-telling equilibrium become weaker. Therefore, the mechanisms are more realistic (as the questions are now allowed to be heterogeneous).

The gamification mechanisms here are extended from the work reviewed in Section 7.4. They now use the estimates of question difficulty, calculated using some agreement metrics (e.g. kappa statistics), to drive the level systems in games. In this case, the crowd-workers are more likely to feel being challenged and excited when facing questions whose difficulty levels match up with their expertise levels.

9.1. Monetary payment in response-level contexts

Only recently have payment-based incentive mechanisms started to consider the heterogeneity in question difficulty 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 questions is high rather than a lack of effort from the workers.

In the machine-aided peer prediction mechanism proposed in [139], 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 so difficult 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 [189], the proposed peer prediction mechanism was also applied to binary-response questions except that it relies solely on *human* peer consistency assessment. 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 or disagree with one another on the correct answer for the question. The larger the summation of the diagonal entries is, the easier the question and it is the opposite for the off-diagonal entries. 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

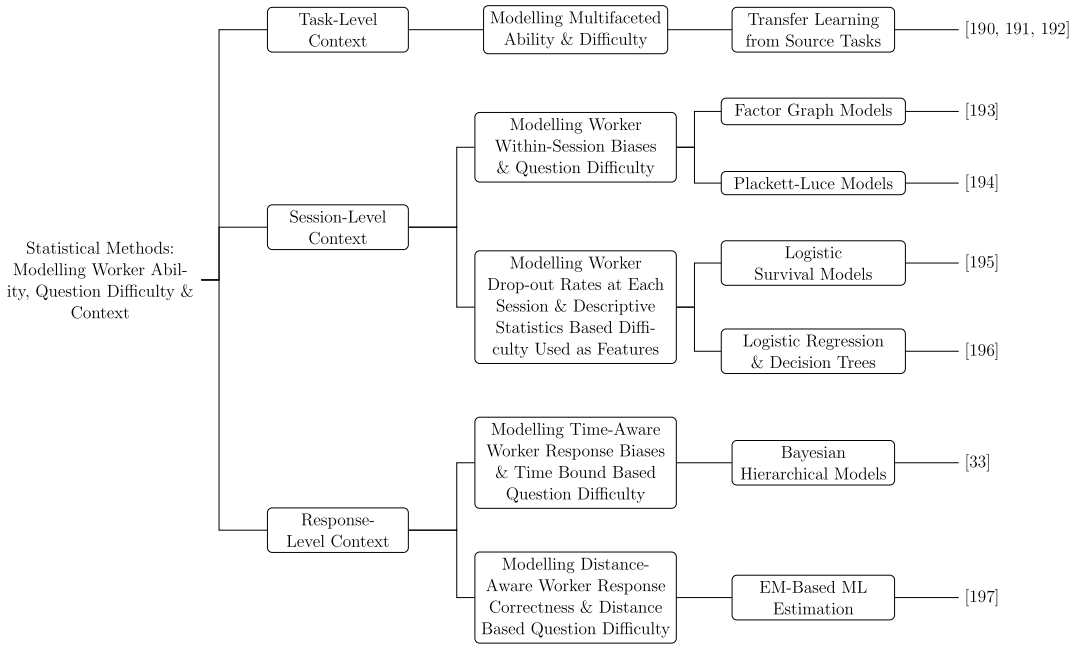


Fig. 13. A taxonomy of QCC papers that considered worker ability, context and question difficulty. They focused on statistical modelling of how worker ability varies with contexts at specific levels, and the interaction between the context-aware ability and the difficulty.

improved incentives for workers to be truthful and was less sensitive to their collusions compared to the previous mechanisms.

9.2. 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. The mechanism sets up different difficulty levels for the questions so that the workers progress from the easiest level to the hardest level to finish a 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 [154], 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. The mechanism was appreciated by the workers with 50% of them praising the level progression.

The level mechanism has played a similar role in [167,158] where higher worker scores trigger higher difficulty levels in the game for the workers to play. In [166], 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 training sessions for workers 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.

10. Modelling worker ability, question difficulty and contexts

A crowd-worker's ability can vary across different contexts in which this worker has been situated. For example, the ability can be dynamic across different tasks the worker has participated, or across different pages of the same task. There are statistical models that have considered the interactions among the worker ability, question difficulty and worker context. These models incorporate latent factors that represent different levels of contexts along with the worker ability and question difficulty factors (see Fig. 13). Most of these models are based on the GLAD framework with some using its multi-dimensional extension [93]. So far, question allocation mechanisms that change the working contexts to make workers perform more effectively have not yet considered modelling the question difficulty. This is mostly because the previous work reviewed in Sections 8.2 and 8.3 has used uncertainty measures on the true answer probabilities as surrogates for a question difficulty variable.

10.1. Modelling the interactions in task-level contexts

In this case, the context in which the interaction between the worker ability and the question difficulty takes place is the whole crowdsourcing platform. The proposed models utilize the response information from the same workers across various source tasks in which they have participated to improve the parameter estimation for a target task. This is referred as the *transfer learning* [198] in the Machine Learning literature.

In [190], the authors proposed a model which encoded multi-dimensional ability and difficulty. Since this multifaceted model has a larger number of parameters, it is intrinsically more vulnerable to response sparsity. The authors thus decided to transfer worker expertise estimates from source tasks to target tasks based on estimated similarity between the tasks. This cross-task transfer was able to smooth out the unreliable expertise estimation in the target tasks. However, no transfer learning was conducted for calibrating the estimation of question difficulty in the target tasks. In contrast, in [191,192], the transfer learning helps to learn a better latent feature representation for each question in the target task from the observed features of the questions in the source tasks. The latent feature representation for each target question can be interpreted as their multi-dimensional difficulty.

10.2. Modelling the 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 models learn latent (bias) variables specific to particular (types of) sessions. Zhuang and Young [193] did this by learning a factor graph model which encoded biases in workers' responses to questions within each session. The factor function is defined as the exponent of a linear regression over counts of different response options (within a particular session). The regression coefficients are global, meaning that they encode a latent bias structure shared across the sessions. This latent structure maps the response count distribution from a session to a bias value which *offsets* each response accuracy (determined by the expertise-difficulty interaction) within that session. In another work [194], the authors 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 [199]. 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 the ML estimation.

In [195], a worker drop-out modelling framework was proposed which consists of a sequence of logistic survival models each corresponding to a particular session/page. Each model determines the probabilities of workers surviving a particular session and moving to the next. The coefficients of 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.) within the session. Similarly, Mao et al. [196] 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.

10.3. Modelling the 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 [33], 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 [197], the quality of a worker's response to a spatial question (e.g. labelling point of interests) is jointly determined by three factors: the worker's intrinsic ability (i.e. the probability of her being reliable[11]), her (response-specific) location-aware ability, and the difficulty of the question. The location-aware ability 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.

11. Modelling worker expertise, question difficulty and response relationships

When semantic relationships between response options are present in crowdsourcing, QCC methods measure the relationships using some *distance* metrics such that options with closer relationships have smaller distances. A common distance

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

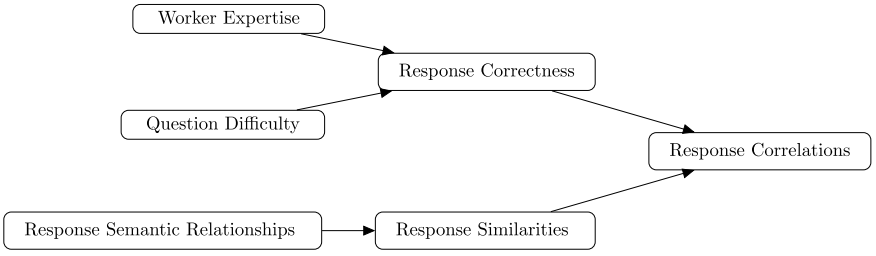


Fig. 14. A diagram shows how worker expertise, question difficulty and response semantic relationships contribute to the correlations within responses.

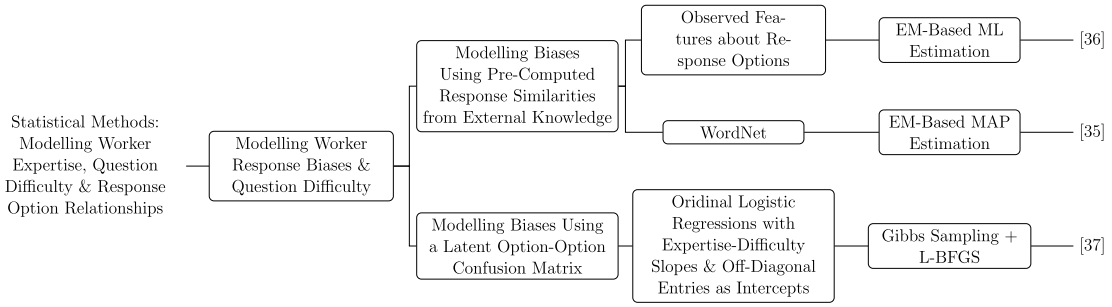


Fig. 15. A taxonomy of QCC papers that considered worker expertise, question difficulty and semantic relationships between responses. These papers focused on statistical modelling which leveraged pre-computed response similarities from external knowledge to account for response biases.

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 options in a semantic structure (e.g. WordNet[12]).

Three papers have considered leveraging semantic relationships between response options [35–37] for improving the quality control of crowdsourced responses (see Figs. 14 and 15). In [35], 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. 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 [36], the authors proposed a model which shares the same idea as [35] 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 semantic relationships to pre-compute the semantic distances.

Recently, Yuan et al. [37] proposed the first QCC model that directly infers the semantic relationships from responses. Their model captures the relationships using a *symmetric latent* relatedness matrix. In this matrix, each off-diagonal entry is a real-valued score representing how related the response options are to one another. The authors also observed two phenomena: (i) capable workers are more likely to select options that are more (semantically) related to the correct answer and (ii) difficult questions are more likely to receive responses less related to their correct answers. The proposed model captures the above phenomena using ordinal logistic regressions. Each regression has a *response-specific slope* being the expertise-difficulty difference, and the off-diagonal relatedness scores as the *ordinal intercepts* specific to each option (other than the correct answer). The authors also showed that their model can elegantly incorporate the external knowledge, via a linear regression, into the prior mean of the Normal distribution which generates the off-diagonal scores. They finally showed that their model outperformed the previous two models [35,36] in true answer prediction.

12. Summary and discussion

Crowdsourcing has become a principle tool that allows research communities and companies to collect data for (machine learning) 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 recorded responses (and their costs). The correspond-

[12] https://wordnet.princeton.edu/

ing solution is the quality control methods which aim to remove or suppress low-quality responses (and possibly amplify the high-quality ones).

The two major categories of quality control methods are quality control mechanism designs and statistical inference models. Surveys exist regarding each category but their reviews were isolated without recognizing the strong connection between the categories. We view the quality control for crowdsourcing as a unified cyclic process that integrates both categories of methods (as shown in Fig. 1). This viewpoint allows us to conduct a more comprehensive survey that bridges the two categories in terms of how their methods work together for crowdsourcing. To link these two categories, we proposed in the survey a framework that systematically unifies all the crowdsourcing aspects (and their important attributes) modelled by both categories to determine the response quality. Many of these aspects and attributes are, for the first time, identified from the quality control literature. Based on this framework, we proposed another graph framework that unifies the past quality control research in terms of the aspects and attributes they have exploited.

Our survey flows by following the (quality control research) graph framework and provides systematic technical insights to a wide variety of quality control methods. It also contributes, for the first time, organized taxonomies of quality control papers. Each taxonomy is characterized by the considered aspects and attributes, modelling assumptions, parameter estimation techniques, and design features of quality control methods of the same type (deemed by the graph).

According to our survey, there are several limitations existing in the current quality control research. First, even though the GLAD and DS models have been widely applied and extended in the research, these applications and extensions rarely aim at *large-scale sparse* and *diverse* crowdsourcing. This type of crowdsourcing is featured by large numbers of questions and crowd-workers with sparse responses across them, and the characteristics of questions vary greatly (e.g. by having diverse topics). It is prevalent in user content generation Websites where the quality of user generated content poses an issue and needs to be controlled (by estimation from responses of other users to the same content).

In this case, workers can react very differently due to the question diversity, and their response correlations will be much weaker than small-scale (non-sparse and non-diverse) crowdsourcing which is the focus of current quality control statistical models. As for the future research directions, Bayesian hierarchical models have great opportunities to properly handle large-scale sparse and diverse crowdsourcing due to its hierarchical nature that smooths the parameter estimation. Among them, hierarchical topical models that have been successfully applied to mining large corpora with diverse topics and sparse word counts [200] can be adapted. Deep generative models [201] are another alternative to this purpose. They can be adapted to capture complex question and worker behavioural diversity directly from responses.

The second limitation is the lack of advanced statistical models for handling the subjectivity that might exist in crowdsourcing. Human judgment and generated content is intrinsically subjective due to personal preferences and opinions. This leads to the variation in responses (to the generated content) and current quality control methods seldom separate it from the quality of responses in their models. Although recent work [112,31] has made some progress in this regard, more need to be done in the future. Especially, how to design measures of subjectivity in crowdsourcing remains an open question. So far, only Yuan et al. [31] directly defined a question-specific subjectivity measure, which was the expected number of correct answers (for a question) with respect to underlying groups of workers. Each group embodies a particular school of thought. Other subjectivity measures can be defined specific to not only questions but also the entire tasks or individual workers. The corresponding statistical models can encode (and quantify) these measures as variables different from the response quality variables.

Another interesting direction is to investigate the dependency between the difficulty and subjectivity of questions. According to the literature, the difficulty determines the response quality while the subjectivity results in response correlations. Yuan et al. [31] made an *independence* assumption about the two attributes. Future study on this subject will allow us to derive more reasonable models if the dependency does exist. More importantly, these subjectivity-aware models can be further fused with those aiming for large-scale sparse and diverse crowdsourcing as user generated contents can exhibit subjectivity. Controlling the quality of these contents requires models that separate out the subjectivity.

The third limitation is the incapability of current quality control methods in handling either *highly multi-class* or *highly multi-label crowdsourcing*. The former concerns crowdsourcing with many response options but there is only a single correct answer for each question. Examples include image categorization based on large numbers of ImageNet categories [41], Web-page categorization based on large numbers of Wikipedia (topical) categories, and etc. The latter concerns each question having many different correct answers (from many response options). A typical example is using crowdsourcing to tag user generated contents [202]. Both types of crowdsourcing could exhibit response correlations resulted from *semantic relationships between the response options*. Highly multi-label crowdsourcing is also a special case of partially subjective crowdsourcing which means that the subjectivity could also have contributed to the response correlation. However, the dependency between the subjectivity and the semantic relationships remains unclear and needs to investigated in the future.

Currently, only three papers have dealt with the quality control for highly multi-class crowdsourcing [35–37] and only one of them shows the potential of reconstructing the semantic relationships directly from responses [37]. As for the highly multi-label crowdsourcing, an effective QCC model is still missing. According to the literature, semantic relationships between response options can be represented by an option-option matrix. This matrix is quadratic in the number of options, and therefore will become very large when the number of options is large. Thus, when estimating this matrix (to reconstruct the semantic relationships), future models need to make use of matrix factorization techniques. Such techniques are

useful for lowering the complexity of the models to prevent over-fitting. The resulting factors for individual options can be modelled to interact with other factors (e.g. the expertise and difficulty factors) to determine the response quality.

The final limitation we have found regards the quality control mechanism designs, particularly for the *worker payment* and the *task gamification*. For the payment design, most of the current techniques rely on game-theoretic approaches. These approaches are theoretically elegant and sound but very few empirical studies have been conducted to systematically verify and compare their quality control performance in practice. Thus, studies in this regard are needed in the near future. As for the gamification design, its methodologies are much less developed compared with the methodologies for either worker payment or question allocation design. 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 responses in general crowdsourcing. These insights will be critical for constructing a common gamification methodology which provides empirically justified guidelines on what combinations of game elements should be used (possibly together with other mechanisms) to improve the quality of responses.

Declaration of competing interest

To the best of our knowledge, there is no conflict of interests.

References

[1] A. Marcus, A. Parameswaran, Crowdsourced data management: industry and academic perspectives, Found. Trends® Databases 6 (1–2) (2015) 1–161, https://doi.org/10.1561/1900000044.

[2] O.F. Zaidan, 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, R. Kuipers, Crowdsourcing for usability testing, in: Proceedings of the Association for Information Science and Technology, vol. 49, Wiley Online Library, 2012, pp. 1–10.

[4] K.-J. Stol, 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] P.G. Ipeirotis, Analyzing the Amazon mechanical turk marketplace, XRDS: crossroads, ACM Mag. Stud. 17 (2) (2010) 16–21.

[6] A. Kumar, 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.

[7] V. Ambati, Active learning and crowdsourcing for machine translation in low resource scenarios, Ph.D. thesis, 2012.

[8] A. Brew, D. Greene, P. Cunningham, The interaction between supervised learning and crowdsourcing, in: NIPS Workshop on Computational Social Science and the Wisdom of Crowds, 2010.

[9] J. Deng, J. Krause, 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.

[10] J. Cheng, 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.

[11] C. Eickhoff, A.P. de Vries, Increasing cheat robustness of crowdsourcing tasks, Inf. Retr. 16 (2) (2013) 121–137.

[12] J. Surowiecki, The Wisdom of Crowds, Anchor, 2005.

[13] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H.R. Motahari-Nezhad, E. Bertino, S. Dustdar, Quality control in crowdsourcing systems: issues and directions, IEEE Internet Comput. 17 (2) (2013) 76–81.

[14] A.I. Chittilappilly, L. Chen, S. Amer-Yahia, A survey of general-purpose crowdsourcing techniques, IEEE Trans. Knowl. Data Eng. 28 (9) (2016) 2246–2266.

[15] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, M. Allahbakhsh, Quality control in crowdsourcing: a survey of quality attributes, assessment techniques, and assurance actions, ACM Comput. Surv. 51 (1) (2018) 7.

[16] J. Muhammadi, H.R. Rabiee, A. Hosseini, A unified statistical framework for crowd labeling, Knowl. Inf. Syst. 45 (2) (2015) 271–294.

[17] J. Zhang, X. Wu, V.S. Sheng, Learning from crowdsourced labeled data: a survey, Artif. Intell. Rev. 46 (4) (2016) 543–576.

[18] Y. Zheng, G. Li, Y. Li, C. Shan, R. Cheng, Truth inference in crowdsourcing: is the problem solved?, Proc. VLDB Endow. 10 (5) (2017) 541–552.

[19] A. Kittur, J.V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, J. Horton, The future of crowd work, in: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, ACM, 2013, pp. 1301–1318.

[20] X. Yin, W. Liu, Y. Wang, C. Yang, 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.

[21] D. Geiger, M. Schader, Personalized task recommendation in crowdsourcing information systems – current state of the art, Decis. Support Syst. 65 (2014) 3–16.

[22] N. Luz, N. Silva, P. Novais, A survey of task-oriented crowdsourcing, Artif. Intell. Rev. 44 (2) (2015) 187–213.

[23] M. Lease, E. Yilmaz, Crowdsourcing for information retrieval: introduction to the special issue, Inf. Retr. 16 (2) (2013) 91–100.

[24] K. Mao, L. Capra, M. Harman, Y. Jia, A survey of the use of crowdsourcing in software engineering, J. Syst. Softw. 126 (2017) 57–84.

[25] G. Xintong, W. Hongzhi, Y. Song, G. Hong, Brief survey of crowdsourcing for data mining, Expert Syst. Appl. 41 (17) (2014) 7987–7994.

[26] B.L. Ranard, Y.P. Ha, Z.F. Meisel, D.A. Asch, S.S. Hill, L.B. Becker, A.K. Seymour, R.M. Merchant, Crowdsourcing–harnessing the masses to advance health and medicine, a systematic review, J. Gen. Intern. Med. 29 (1) (2014) 187–203.

[27] C. Gomes, D. Schneider, K. Moraes, J.d. Souza, Crowdsourcing for music: survey and taxonomy, in: 2012 IEEE International Conference on Systems, Man, and Cybernetics, 2012, pp. 832–839.

[28] C. Heipke, Crowdsourcing geospatial data, ISPRS J. Photogramm. Remote Sens. 65 (6) (2010) 550–557.

[29] R.M. Frongillo, Y. Chen, I.A. Kash, Elicitation for aggregation, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI Press, 2015, pp. 900–906.

[30] C.-J. Ho, R. Frongillo, Y. Chen, Eliciting categorical data for optimal aggregation, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2016, pp. 2450–2458.

[31] Y. Jin, M. Carman, Y. Zhu, W. Buntine, Distinguishing question subjectivity from difficulty for improved crowdsourcing, in: J. Zhu, I. Takeuchi (Eds.), Proceedings of the 10th Asian Conference on Machine Learning, in: Proceedings of Machine Learning Research, PMLR, vol. 95, 2018, pp. 192–207.
[32] H.J. Jung, Y. Park, M. Lease, Predicting next label quality: a time-series model of crowdwork, in: Second AAAI Conference on Human Computation and Crowdsourcing, 2014.
[33] M. Venanzi, J. Guiver, P. Kohli, N.R. Jennings, Time-sensitive Bayesian information aggregation for crowdsourcing systems, J. Artif. Intell. Res. 56 (2016) 517–545.
[34] H.J. Jung, M. Lease, Modeling temporal crowd work quality with limited supervision, in: Third AAAI Conference on Human Computation and Crowdsourcing, 2015.
[35] T. Han, H. Sun, Y. Song, Y. Fang, 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.
[36] Y.-L. Fang, H.-L. Sun, P.-P. Chen, T. Deng, Improving the quality of crowdsourced image labeling via label similarity, J. Comput. Sci. Technol. 32 (5) (2017) 877–889.
[37] Y. Jin, L. Du, Y. Zhu, M. Carman, Leveraging label category relationships in multi-class crowdsourcing, in: Proceedings of the 22nd Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Springer International Publishing, 2018, pp. 128–140.
[38] B. Loni, M. Menendez, M. Georgescu, L. Galli, C. Massari, I.S. Altingovde, D. Martinenghi, M. Melenhorst, R. Vliegendhart, M. Larson, Fashion-focused creative commons social dataset, in: Proceedings of the 4th ACM Multimedia Systems Conference, 2013, pp. 72–77.
[39] B. Loni, L.Y. Cheung, M. Riegler, A. Bozzon, L. Gottlieb, M. Larson, Fashion 10000: an enriched social image dataset for fashion and clothing, in: Proceedings of the 5th ACM Multimedia Systems Conference, 2014, pp. 41–46.
[40] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[41] A. Khosla, N. Jayadevaprakash, B. Yao, F.-F. Li, Novel dataset for fine-grained image categorization, in: First Workshop on Fine-Grained Visual Categorization, CVPR, 2011, Citeseer.
[42] A.P. Dawid, A.M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, Appl. Stat. (1979) 20–28.
[43] R. Snow, B. O'Connor, D. Jurafsky, 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.
[44] W. Tang, M. Lease, Semi-supervised consensus labeling for crowdsourcing, in: SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval, 2011, pp. 1–6.
[45] Y. Zhang, X. Chen, D. Zhou, M.I. Jordan, Spectral methods meet em: a provably optimal algorithm for crowdsourcing, in: Advances in Neural Information Processing Systems, 2014, pp. 1260–1268.
[46] Y. Zhang, X. Chen, D. Zhou, M. Jordan, Spectral methods meet em: a provably optimal algorithm for crowdsourcing, J. Mach. Learn. Res. 17 (1) (2016) 3537–3580.
[47] C. Gao, D. Zhou, Minimax optimal convergence rates for estimating ground truth from crowdsourced labels, arXiv e-prints arXiv:1310.5764, 2013.
[48] B. Carpenter, A hierarchical Bayesian model of crowdsourced relevance coding, in: Text REtrieval Conference, 2011.
[49] H.-C. Kim, Z. Ghahramani, Bayesian classifier combination, in: Artificial Intelligence and Statistics, 2012, pp. 619–627.
[50] E. Simpson, S. Roberts, I. Psorakis, A. Smith, Dynamic Bayesian combination of multiple imperfect classifiers, in: Decision Making and Imperfection, Springer, 2013, pp. 1–35.
[51] A. Ghosh, S. Kale, 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.
[52] D.R. Karger, S. Oh, 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.
[53] N. Dalvi, A. Dasgupta, R. Kumar, V. Rastogi, Aggregating crowdsourced binary ratings, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 285–294.
[54] D.R. Karger, S. Oh, D. Shah, Iterative learning for reliable crowdsourcing systems, in: Advances in Neural Information Processing Systems, 2011, pp. 1953–1961.
[55] D.R. Karger, S. Oh, D. Shah, Budget-optimal task allocation for reliable crowdsourcing systems, Oper. Res. 62 (1) (2014) 1–24.
[56] Q. Liu, J. Peng, A.T. Ihler, Variational inference for crowdsourcing, in: Advances in Neural Information Processing Systems, 2012, pp. 692–700.
[57] J. Ok, S. Oh, J. Shin, Y. Yi, Optimality of belief propagation for crowdsourced classification, in: International Conference on Machine Learning, 2016, pp. 535–544.
[58] T. Bonald, R. Combes, A minimax optimal algorithm for crowdsourcing, in: Advances in Neural Information Processing Systems, 2017.
[59] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, 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.
[60] P.G. Moreno, A. Artés-Rodríguez, Y.W. Teh, F. Perez-Cruz, Bayesian nonparametric crowdsourcing, J. Mach. Learn. Res. 16 (2015) 1607–1627.
[61] H. Imamura, I. Sato, M. Sugiyama, Analysis of minimax error rate for crowdsourcing and its application to worker clustering model, in: Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 2152–2161.
[62] C. Liu, Y.-M. Wang, Truelabel+ confusions: a spectrum of probabilistic models in analyzing multiple ratings, in: Proceedings of the 29th International Conference on International Conference on Machine Learning, Omnipress, 2012, pp. 17–24.
[63] E. Kamar, A. Kapoor, E. Horvitz, Identifying and accounting for task-dependent bias in crowdsourcing, in: Third AAAI Conference on Human Computation and Crowdsourcing, 2015.
[64] O. Isupova, Y. Li, D. Kuzin, S.J. Roberts, K. Willis, S. Reece, Bccnet: Bayesian classifier combination neural network, arXiv e-prints arXiv:1811.12258, 2018.
[65] J. Yang, T. Drake, A. Damianou, Y. Maarek, Leveraging crowdsourcing data for deep active learning an application: learning intents in alexa, in: Proceedings of the 2018 World Wide Web Conference, WWW'18, International World Wide Web Conferences Steering Committee, 2018, pp. 23–32.
[66] F. Rodrigues, F.C. Pereira, Deep learning from crowds, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018, pp. 1611–1618.
[67] P. Cao, Y. Xu, Y. Kong, Y. Wang, Max-mig: an information theoretic approach for joint learning from crowds, in: Proceedings of the 7th International Conference on Learning Representations (ICLR), 2019.
[68] K.G. Dizaji, H. Huang, Sentiment analysis via deep hybrid textual-crowd learning model, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018, pp. 1563–1570.
[69] M.Y. Guan, V. Gulshan, A.M. Dai, G.E. Hinton, Who said what: modeling individual labelers improves classification, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.
[70] B.I. Aydin, Y.S. Yilmaz, Y. Li, Q. Li, J. Gao, M. Demirbas, Crowdsourcing for multiple-choice question answering, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 2946–2953.
[71] R. Yan, Y. Song, C.-T. Li, M. Zhang, 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.
[72] S. Kajimura, Y. Baba, H. Kajino, H. Kashima, Quality control for crowdsourced poi collection, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2015, pp. 255–267.

[73] T. Sunahase, Y. Baba, 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.
[74] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, ACM SIGKDD Explor. Newsl. 17 (2) (2016) 1–16.
[75] R.W. Ouyang, L. Kaplan, P. Martin, A. Toniolo, M. Srivastava, 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.
[76] R.W. Ouyang, L.M. Kaplan, A. Toniolo, M. Srivastava, T.J. Norman, Aggregating crowdsourced quantitative claims: additive and multiplicative models, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1621–1634.
[77] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, J. Mach. Learn. Res. 11 (April 2010) 1297–1322.
[78] C. Gao, Y. Lu, D. Zhou, Exact exponent in optimal rates for crowdsourcing, in: International Conference on Machine Learning, 2016, pp. 603–611.
[79] D.R. Karger, S. Oh, D. Shah, Efficient crowdsourcing for multi-class labeling, ACM SIGMETRICS Perform. Eval. Rev. 41 (1) (2013) 81–92.
[80] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[81] G. Hinton, N. Srivastava, K. Swersky, Overview of mini-batch gradient descent, Lecture Notes Distributed in CSC321 of University of Toronto 2014, Lecture 6a.
[82] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (December 2010) 3371–3408.
[83] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632.
[84] J. Whitehill, T.-f. Wu, J. Bergsma, J.R. Movellan, 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.
[85] Y. Bachrach, T. Graepel, T. Minka, J. Guiver, How to grade a test without knowing the answers—a Bayesian graphical model for adaptive crowdsourcing and aptitude testing, arXiv e-prints arXiv:1206.6386, 2012.
[86] X. Zhang, H. Shi, Y. Li, W. Liang, Spglad: a self-paced learning-based crowdsourcing classification model, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2017, pp. 189–201.
[87] D. Zhou, S. Basu, Y. Mao, J.C. Platt, Learning from the wisdom of crowds by minimax entropy, in: Advances in Neural Information Processing Systems, 2012, pp. 2195–2203.
[88] D. Zhou, Q. Liu, J.C. Platt, C. Meek, N.B. Shah, Regularized minimax conditional entropy for crowdsourcing, arXiv e-prints arXiv:1503.07240, 2015.
[89] P. Ruvolo, J. Whitehill, J.R. Movellan, Exploiting structure in crowdsourcing tasks via latent factor models, Tech. Rep., 2010.
[90] P. Ruvolo, J. Whitehill, 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.
[91] H. Kajino, Y. Tsuboi, H. Kashima, A convex formulation for learning from crowds, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012.
[92] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, J. Han, Faitcrowd: fine grained truth discovery for crowdsourced data aggregation, in: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 745–754.
[93] P. Welinder, S. Branson, P. Perona, S.J. Belongie, The multidimensional wisdom of crowds, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2010, pp. 2424–2432.
[94] F.L. Wauthier, M.I. Jordan, Bayesian bias mitigation for crowdsourcing, in: Advances in Neural Information Processing Systems, 2011, pp. 1800–1808.
[95] Y. Jin, M. Carman, D. Kim, L. Xie, Leveraging side information to improve label quality control in crowd-sourcing, in: Proceedings of the 5th AAAI Conference on Human Computation and Crowdsourcing (HCOMP), 2017, pp. 79–88.
[96] L. Yin, J. Han, W. Zhang, 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, 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, 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, Y.W. Teh, Inferring ground truth from multi-annotator ordinal data: a probabilistic approach, arXiv e-prints arXiv:1305.0015, 2013.
[100] H.J. Jung, 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.
[101] H.J. Jung, M. Lease, Improving quality of crowdsourced labels via probabilistic matrix factorization, in: Proceedings of the 4th Human Computation Workshop at AAAI, 2012, pp. 101–106.
[102] H.J. Jung, M. Lease, Crowdsourced task routing via matrix factorization, arXiv e-prints arXiv:1310.5142, 2013.
[103] 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.
[104] D. Zhou, Q. Liu, J. Platt, 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.
[105] G. Chen, S. Zhang, D. Lin, H. Huang, P.A. Heng, Learning to aggregate ordinal labels by maximizing separating width, in: Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, PMLR, vol. 70, 2017, pp. 787–796.
[106] M.P. Kumar, B. Packer, D. Koller, Self-Paced Learning for Latent Variable Models, Advances in Neural Information Processing Systems, vol. 23, Curran Associates, Inc., 2010, pp. 1189–1197.
[107] T.L. Griffiths, Z. Ghahramani, The Indian buffet process: an introduction and review, J. Mach. Learn. Res. 12 (April 2011) 1185–1224.
[108] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (January 2003) 993–1022.
[109] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, in: Proceedings of the 2nd International Conference on Learning Representations, 2013.
[110] P. Metrikov, V. Pavlu, 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, 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] A.T. Nguyen, M. Halpern, B.C. Wallace, M. Lease, Probabilistic modeling for crowdsourcing partially-subjective ratings, in: Fourth AAAI Conference on Human Computation and Crowdsourcing, 2016.
[113] C.E. Antoniak, Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, Ann. Stat. (1974) 1152–1174.
[114] R.M. Ryan, E.L. Deci, Intrinsic and extrinsic motivations: classic definitions and new directions, Contemp. Educ. Psychol. 25 (1) (2000) 54–67.
[115] N. Shah, D. Zhou, Y. Peres, Approval voting and incentives in crowdsourcing, in: Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 10–19.
[116] N.B. Shah, D. Zhou, Double or nothing: multiplicative incentive mechanisms for crowdsourcing, in: Advances in Neural Information Processing Systems, 2015, pp. 1–9.
[117] N. Shah, 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.

[118] M. Chakraborty, S. Das, Trading on a rigged game: outcome manipulation in prediction markets, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, AAAI Press, 2016, pp. 158–164.
[119] R. Freeman, S. Lahaie, D.M. Pennock, Crowdsourced outcome determination in prediction markets, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press, 2017, pp. 523–529.
[120] G. Azmat, M. Möller, Competition among contests, Rand J. Econ. 40 (4) (2009) 743–768.
[121] A. Ghosh, P. Hummel, Cardinal contests, ACM Trans. Econ. Comput. 6 (2) (2018) 7.
[122] D. DiPalantino, M. Vojnovic, Crowdsourcing and all-pay auctions, in: Proceedings of the 10th ACM Conference on Electronic Commerce, ACM, 2009, pp. 119–128.
[123] M. Rokicki, S. Zerr, S. Siersdorfer, Groupsourcing: team competition designs for crowdsourcing, in: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2015, pp. 906–915.
[124] M. Rokicki, S. Zerr, S. Siersdorfer, Just in time: controlling temporal performance in crowdsourcing competitions, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 817–827.
[125] L. Von Ahn, L. Dabbish, Designing games with a purpose, Commun. ACM 51 (8) (2008) 58–67.
[126] B. Waggoner, Y. Chen, Output agreement mechanisms and common knowledge, in: Second AAAI Conference on Human Computation and Crowdsourcing, 2014.
[127] N. Miller, P. Resnick, R. Zeckhauser, Eliciting informative feedback: the peer-prediction method, Manag. Sci. 51 (9) (2005) 1359–1373.
[128] R. Jurca, B. Faltings, Collusion-resistant, incentive-compatible feedback payments, in: Proceedings of the 8th ACM Conference on Electronic Commerce, ACM, 2007, pp. 200–209.
[129] R. Jurca, B. Faltings, et al., Mechanisms for making crowds truthful, J. Artif. Intell. Res. 34 (1) (2009) 209.
[130] Y. Kong, K. Ligett, 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.
[131] D. Prelec, A Bayesian truth serum for subjective data, Science 306 (5695) (2004) 462–466.
[132] D.C. Parkes, 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.
[133] B. Faltings, R. Jurca, P. Pu, B.D. Tran, Incentives to counter bias in human computation, in: Second AAAI Conference on Human Computation and Crowdsourcing, 2014.
[134] G. Radanovic, B. Faltings, R. Jurca, Incentives for effort in crowdsourcing using the peer truth serum, ACM Trans. Intell. Syst. Technol. 7 (4) (2016) 48.
[135] A. Dasgupta, A. Ghosh, Crowdsourced judgement elicitation with endogenous proficiency, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 319–330.
[136] J. Witkowski, Y. Bachrach, P. Key, D.C. Parkes, Dwelling on the negative: incentivizing effort in peer prediction, in: First AAAI Conference on Human Computation and Crowdsourcing, 2013.
[137] Y. Liu, Y. Chen, Learning to incentivize: eliciting effort via output agreement, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, 2016, pp. 3782–3788.
[138] Y. Liu, Y. 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.
[139] Y. Liu, Y. Chen, Machine-learning aided peer prediction, in: Proceedings of the 2017 ACM Conference on Economics and Computation, ACM, 2017, pp. 63–80.
[140] X.A. Gao, A. Mao, Y. Chen, 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.
[141] Y. Singer, M. Mittal, Pricing mechanisms for crowdsourcing markets, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 1157–1166.
[142] C.-J. Ho, A. Slivkins, S. Suri, J.W. Vaughan, Incentivizing high quality crowdwork, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 419–429.
[143] B. Faltings, G. Radanovic, Game theory for data science: eliciting truthful information, Synth. Lect. Artif. Intell. Mach. Learn. 11 (2) (2017) 1–151.
[144] Y. Chen, D.M. Pennock, Designing markets for prediction, AI Mag. 31 (4) (2010) 42–52.
[145] R. Hanson, Combinatorial information market design, Inf. Syst. Front. 5 (1) (2003) 107–119.
[146] S. Agrawal, E. Delage, M. Peters, Z. Wang, Y. Ye, A unified framework for dynamic pari-mutuel information market design, in: Proceedings of the 10th ACM Conference on Electronic Commerce, ACM, 2009, pp. 255–264.
[147] L.C. Corchón, The theory of contests: a survey, Rev. Econ. Des. 11 (2) (2007) 69–100, https://doi.org/10.1007/s10058-007-0032-5.
[148] T.X. Liu, J. Yang, L.A. Adamic, Y. Chen, Crowdsourcing with all-pay auctions: a field experiment on taskcn, Manag. Sci. 60 (8) (2014) 2020–2037.
[149] K.J. Boudreau, N. Lacetera, K.R. Lakhani, Incentives and problem uncertainty in innovation contests: an empirical analysis, Manag. Sci. 57 (5) (2011) 843–863.
[150] R.M. Araujo, 99designs: an analysis of creative competition in crowdsourced design, in: First AAAI Conference on Human Computation and Crowdsourcing, 2013.
[151] D.P. Gross, Performance feedback in competitive product development, Rand J. Econ. 48 (2) (2017) 438–466.
[152] G. Radanovic, 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.
[153] M. Brenner, N. Mirza, 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.
[154] A. Dumitrache, L. Aroyo, C. Welty, R.-J. Sips, A. Levas, Dr. detective: combining gamification 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.
[155] 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.
[156] J. He, M. Bron, L. Azzopardi, 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.
[157] C. Eickhoff, C.G. Harris, A.P. de Vries, 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.
[158] S. Saito, T. Watanabe, M. Kobayashi, H. Takagi, Skill development framework for micro-tasking, in: International Conference on Universal Access in Human-Computer Interaction, Springer, 2014, pp. 400–409.
[159] L. Guillot, Q. Bragard, R. Smith, A. Ventresque, Towards a gamified system to improve translation for online meetings, in: The 3rd International Workshop on Gamification for Information Retrieval, CEUR, 2016.
[160] J. Schlotterer, C. Seifert, L. Wagner, M. Granitzer, A game with a purpose to access Europe's cultural treasure, in: The 2nd International Workshop Gamification for Information Retrieval, 2015.
[161] W. Moazzam, M. Riegler, S. Sen, M. Nygaard, Scientific hangman: gamifying scientific evidence for general public, in: The 2nd International Workshop Gamification for Information Retrieval, 2015.

[162] A. Carlier, A. Salvador, F. Cabezas, X. Giro-i-Nieto, V. Charvillat, O. Marques, Assessment of crowdsourcing and gamification loss in user-assisted object segmentation, Multimed. Tools Appl. 75 (23) (2016) 15901–15928.
[163] P.G. Ipeirotis, 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.
[164] L.C. Stanculescu, A. Bozzon, R.-J. Sips, 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.
[165] J. Boyd-Graber, B. Satinoff, H. He, 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.
[166] M. Dontcheva, R.R. Morris, J.R. Brandt, 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.
[167] T.Y. Lee, C. Dugan, W. Geyer, T. Ratchford, J.C. Rasmussen, N.S. Shami, S. Lupushor, Experiments on motivational feedback for crowdsourced workers, in: The 7th International AAAI Conference on Web and Social Media, 2013, pp. 341–350.
[168] A.D. Mason, G. Michalakidis, 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.
[169] T. Itoko, S. Arita, M. Kobayashi, H. Takagi, Involving senior workers in crowdsourced proofreading, in: International Conference on Universal Access in Human-Computer Interaction, Springer, 2014, pp. 106–117.
[170] E. Massung, D. Coyle, K.F. Cater, M. Jay, 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.
[171] Y. Jin, M. Carman, X. Lexing, A Little Competition Never Hurt Anyone's Relevance Assessments, Proceedings of the Third International Workshop on Gamification for Information Retrieval (GamifIR), vol. 1642, CEUR Workshop Proceedings, 2016, pp. 29–36.
[172] K. Jennings, Brainiac: Adventures in the Curious, Competitive, Compulsive World of Trivia Buffs, Villard Books, 2007.
[173] M. Lease, G. Kazai, Overview of the TREC 2011 crowdsourcing track, in: Proceedings of the Text Retrieval Conference, 2011.
[174] K.R. Christy, J. Fox, Leaderboards in a virtual classroom: a test of stereotype threat and social comparison explanations for women's math performance, Comput. Educ. 78 (2014) 66–77.
[175] L. Hakulinen, T. Auvinen, A. Korhonen, The effect of achievement badges on students behavior: an empirical study in a university-level computer science course, Int. J.: Emerg. Technol. Learn. 10 (1) (2015) 18–29.
[176] J. Hamari, Transforming homo economicus into homo ludens: a field experiment on gamification in a utilitarian peer-to-peer trading service, Electron. Commer. Res. Appl. 12 (4) (2013) 236–245.
[177] J. Bragg, W. Edu, D.S. Weld, Learning on the job: optimal instruction for crowdsourcing, in: ICML Workshop on Crowdsourcing and Machine Learning, 2015.
[178] U. Gadiraju, B. Fetahu, 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.
[179] Y. Zhou, A.R. Nelakurthi, J. He, Unlearn what you have learned: adaptive crowd teaching with exponentially decayed memory learners, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'18, ACM, 2018, pp. 2817–2826.
[180] V.S. Sheng, F. Provost, 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.
[181] P.G. Ipeirotis, F. Provost, V.S. Sheng, J. Wang, Repeated labeling using multiple noisy labelers, Data Min. Knowl. Discov. 28 (2) (2014) 402–441.
[182] Y. Yan, R. Rosales, G. Fung, J.G. Dy, Active learning from crowds, in: Proceedings of the 28th International Conference on Machine Learning, Omnipress, 2011, pp. 1161–1168.
[183] J. Wang, P. Ipeirotis, A framework for quality assurance in crowdsourcing, No. CBA-13-06, NYU Faculty Digital Archive, 2013.
[184] J. Wang, P.G. Ipeirotis, F. Provost, Cost-effective quality assurance in crowd labeling, Inf. Syst. Res. 28 (1) (2017) 137–158.
[185] J. Zhang, X. Wu, V.S. Shengs, Active learning with imbalanced multiple noisy labeling, IEEE Trans. Cybern. 45 (5) (2015) 1095–1107.
[186] L.P. Kaelbling, M.L. Littman, A.R. Cassandra, Planning and acting in partially observable stochastic domains, Artif. Intell. 101 (1–2) (1998) 99–134.
[187] B. Bollobás, Random Graphs, No. 73, Cambridge University Press, 2001.
[188] P. Donmez, J. Carbonell, 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.
[189] D. Mandal, M. Leifer, D.C. Parkes, G. Pickard, V. Shnayder, Peer prediction with heterogeneous tasks, arXiv preprint arXiv:1612.00928.
[190] K. Mo, E. Zhong, 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.
[191] M. Fang, J. Yin, 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.
[192] M. Fang, J. Yin, D. Tao, Active learning for crowdsourcing using knowledge transfer, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 1809–1815.
[193] H. Zhuang, 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.
[194] H. Zhuang, A. Parameswaran, D. Roth, J. Han, Debiasing crowdsourced batches, in: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1593–1602.
[195] A. Kobren, C.H. Tan, P. Ipeirotis, 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.
[196] A. Mao, E. Kamar, E. Horvitz, Why stop now? Predicting worker engagement in online crowdsourcing, in: First AAAI Conference on Human Computation and Crowdsourcing, 2013.
[197] H. Hu, Y. Zheng, Z. Bao, G. Li, J. Feng, 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.
[198] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[199] R.L. Plackett, The Analysis of Permutations, Applied Statistics, 1975, pp. 193–202.
[200] K.W. Lim, W. Buntine, C. Chen, L. Du, Nonparametric Bayesian topic modelling with the hierarchical Pitman-Yor processes, Int. J. Approx. Reason. 78 (C) (2016) 172–191.
[201] A. Srivastava, C. Sutton, Autoencoding variational inference for topic models, arXiv e-prints arXiv:1703.01488, 2017.
[202] C. Jonathan, M.F. Mokbel, A demonstration of Stella: a crowdsourcing-based geotagging framework, Proc. VLDB Endow. 10 (12) (2017) 1969–1972.