Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms

Mark E. Whiting, Dilrukshi Gamage, Snehalkumar (Neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan, Teogenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A. Mullings, Yoni Dayan, Kristy Milland, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, Michael S. Bernstein
Stanford Crowd Research Collective
daemo@cs.stanford.edu

ABSTRACT

Crowd workers are distributed and decentralized. While decentralization is designed to utilize independent judgment to promote high-quality results, it paradoxically undercuts behaviors and institutions that are critical to high-quality work. Reputation is one central example: crowdsourcing systems depend on reputation scores from decentralized workers and requesters, but these scores are notoriously inflated and uninformative. In this paper, we draw inspiration from historical worker guilds (e.g., in the silk trade) to design and implement *crowd guilds*: centralized groups of crowd workers who collectively certify each other's quality through double-blind peer assessment. A two-week field experiment compared crowd guilds to a traditional decentralized crowd work model. Crowd guilds produced reputation signals more strongly correlated with ground-truth worker quality than signals available on current crowd working platforms, and more accurate than in the traditional model.

Author Keywords
crowdsourcing platforms; human computation

ACM Classification Keywords
H.5.3. Group and Organization Interfaces

INTRODUCTION

Crowdsourcing platforms such as Amazon Mechanical Turk decentralize their workforce, designing for distributed, independent work [17, 44]. Decentralization aims to encourage accuracy through independent judgement [61]. However, by making communication and coordination more difficult, decentralization disempowers workers and forces worker collectives off-platform [43, 66, 17]. The result is disenfranchisement [24, 57] and an unfavorable workplace environment [43, 44]. Worse, while decentralization is motivated by a desire for high-quality work, it paradoxically undercuts behaviors and institutions that are critical to high-quality work. In many traditional organizations, for example, centralized worker coordination is a keystone to behaviors that improve work quality, including skill development [3], knowledge management [37], and performance ratings [60].

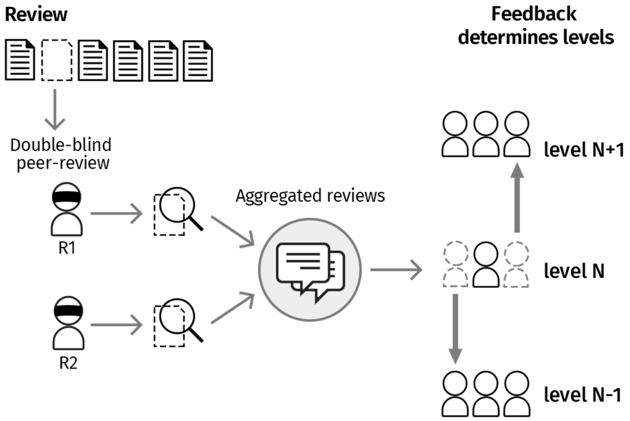


Figure 1. Crowd guilds provide reputation signals through double blind peer-review. The reviews determine workers' levels.

In this paper, we focus on *reputation* as an exemplar challenge that arises from worker decentralization: effective reputation signals are traditionally reliant on centralized mechanisms such as performance reviews [60, 25]. Crowdsourcing platforms rely heavily on their reputation systems, such as task acceptance rates, to help requesters identify high-quality workers [24, 45]. On Mechanical Turk, as on other on-demand platforms such as Upwork and Uber, these reputation scores are derived from decentralized feedback from independent requesters. However, the resulting reputation scores are no-

CSCW '17, February 25-March 01, 2017, Portland, OR, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4335-0/17/03 $15.00
DOI: http://dx.doi.org/10.1145/2998181.2998234

toriously inflated and noisy, making it difficult for requesters to find high-quality workers and difficult for workers to be compensated for their quality [45, 22].

To address this reputation challenge, and with an eye toward other challenges that arise from decentralization, we draw inspiration from a historical labor strategy for coordinating a decentralized workforce: *guilds*. Worker guilds arose in the early Middle Ages, when workers in a trade such as silk were distributed across a large region, as bounded sets of laborers who shared an affiliation. These guilds played many roles, including training apprentices [19, 46], setting prices [47], and providing mechanisms for collective action [54, 51]. Especially relevant to the current challenge, guilds measured and certified their own members' quality [19]. While guilds eventually lost influence due to exerting overly tight controls on trade [47] and exogenous technical innovations in production, their intellectual successors persist today as professional organizations such as in engineering, acting and medicine [48, 35]. Malone first promoted a vision of online "e-lancer" guilds twenty years ago [42], but to date no concrete instantiations exist for a modern, online crowd work economy.

We present *crowd guilds*: crowd worker collectives that coordinate to certify their own members and perform internal feedback to train members (Figure 1). Our infrastructure for crowd guilds enables workers to engage in continuous double-blind peer assessment [32] of a random sample of members' task submissions on the crowdsourcing platform, rating the quality of the submission and providing critiques for further improvement. These peer assessments are used to derive guild levels (e.g., Level 1, Level 2) to serve as reputation (qualification) signals on the crowdsourcing platform. As workers gather positive assessments from more senior guild members, they rise in levels within the guild. Guilds translate these levels into higher wages by recommending pay rates for each level when tasks are posted to the platform. While crowd guilds focus here on worker reputation, our experiment implementation also explores how crowd guilds could address other challenges such as collective action (e.g., collectively rejecting tasks that pay too little), formal mentorship (e.g., repeated feedback and training), and social support (e.g., on the forums). Because existing platforms cannot be used to support these affordances, we have implemented crowd guilds on the open-source Daemo crowdsourcing platform [15].

As with historical guilds, several crowd guilds can operate in parallel across different areas of expertise. To focus our exploration, our prototype in this paper implements a single platform-wide guild for microtask workers. While in theory anybody can perform microtask work, prior work has demonstrated that it requires a wide range of both visible and invisible expertise to perform supposedly 'simple' work effectively [43]. This need for expert, high-quality microtask work has previously driven both platforms and requesters to curate private groups of expert workers (e.g., Mechanical Turk Masters, requesters' private qualification groups) [45]. In our case, this expertise makes microtask work an appropriate candidate for a guild.

We performed a two-week field experiment to evaluate whether crowd guilds establish accurate worker reputation signals. We recruited 300 workers from Amazon Mechanical Turk, randomizing them between control and crowd guild conditions. We then launched tasks daily to the workers in these two groups, providing both conditions with a forum and automatically-generated peer assessment tasks. In the crowd guild condition, the peer assessment tasks determined guild levels and workers received the assessment feedback. In the control condition, no guild levels were available, and the peer assessments were never returned to the worker being assessed. We calculated each worker's accuracy using gold-standard tasks launched on the platform.

Guilds' peer assessed ratings were a significantly better predictor of workers' actual accuracy than the workers' acceptance rates on Amazon Mechanical Turk. Furthermore, workers' peer assessment was significantly more accurate and less inflated in the guilds condition, when peers' reputations were attached to the assessment. Workers in the guilds condition also provided one another with more actionable feedback and advice than those in the control condition.

In sum, this paper contributes a design and infrastructure for crowd guilds as a re-centralizing force for crowdsourcing marketplaces. We target crowd guilds at addressing reputation challenges, because reliable reputation is a difficult and representative problem of the negative outcomes of worker decentralization. To follow, we review related work on crowd collectives and historical guilds, then introduce our design and describe our field experiment.

RELATED WORK

In this section, we first review literature on how crowdsourcing marketplaces can improve work quality, focusing on peer assessment methods. We then draw upon literature on the historical formation of guilds and their legacy to the modern world with a focus on structures for crowd workers to enhance their work reputation and improve communication in the online crowd work environment. Finally we discuss literature on crowdsourced worker communities and their collective behavior in online labor markets.

Improving work quality

Ensuring high-quality crowd work is crucial for the sustainability of a microtask platform. Mechanisms such as voting by peer workers [6], establishing agreement between workers [59], and machine learning models trained on low-level behaviors [56], have been used to gauge and enhance the quality of crowd work. In addition to these techniques, task-specific feedback help crowd workers augment their behavior and improve their performance [11].

Crowdsourcing platforms have collected work feedback through requesters, workers, using self-assessment rubrics, and with the help of expert evaluators. The Shepherd system [11] allows workers to exchange synchronous feedback with requesters. Crowd guilds scale up this notion of distributed feedback [7] to make peer assessment and collective reputation management a core feature of a crowdsourcing platform.

Self-assessment is another route to help workers reflect, learn skills and more clearly draw connections between learning goals and evaluation criteria [5]. However, workers in self-assessment systems become dependant on rubrics or use special domain language, which tends to be difficult for novices to understand [4]. Automated feedback [20] also enhances workers' performance. However, such systems are generally used to enhance the capabilities of specialized platforms; for example, Duolingo and LevelUp integrate interactive tutorials to enhance the quality of work [10], which requires significant customization for a given domain, and has not been demonstrated in general work platforms.

Peer assessment driving quality

If crowd workers can effectively assess each other, they could bootstrap their own work reputation signals [68, 32]. Worker peer review can scale more readily than external assessment, and leverages a community of practice built up amongst workers with the same expertise [11]. It can also be comparably accurate: peer assessment matches expert assessment in massive open online classes [32].

For effectively assessing each other's contributions, it may be prudent to recruit assessors based on prior task performance. Algorithms can facilitate this by adaptively selecting particular raters based on estimated quality, focusing high quality work where it's most needed [50]. The Argonaut system [20] and MobileWorks [31] demonstrate reviews through hierarchies of trusted workers. However, promotions between hierarchies in the Argonaut system require human managers and MobileWorks needs managers to play an essential procedural role in the quality assurance process by determining the final answer for tasks that are reported by workers as difficult, which restricts these systems' ability to scale. In contrast, crowd guilds provide automatic reputation levels based on peer assessment results to algorithmically establish who is qualified to evaluate which work.

The advice and feedback provided in peer assessment can facilitate distributed mentorship [7]. Self and peer assessment can train reviewers to become better at the craft themselves [32, 1, 11], and feedback from those with more expertise improves result quality [11, 41]. Returning peer assessment feedback rapidly can increase iteration and aid learning toward mastery [33]. In crowd guilds, we introduce a review framework that draws on these insights, using workers' assessments of their peers' work to offer quality based reputation categories and constructive feedback.

Building on the above approaches, crowd guilds utilize the observation that crowd members can evaluate each others' work accuracy [11, 20]. Additionally, we demonstrate that crowd guilds can create not just a rating for individual pieces of work, but a stable and informational reputation system.

Guilds

Existing crowdsourcing markets deploy ad hoc methods for improving and sustaining the community of workers. To design a holistic community, we take inspiration from *guilds*. Historically, guilds represented groups of workers with shared interests and goals, as well as enabling large-scale collective behaviors such as reputation management.

Guilds originally evolved as associations of artisans or merchants who controlled the practice of their craft in medieval towns [51]. These craftspeople's guilds, formed by experienced and confirmed experts in their respective fields of handicraft, behaved as professional associations, trade unions, cartels, and even secret societies [55]. They used internal quality evaluation processes to progress members through a system of titles, often starting with *apprentices*, who would go through some schooling with the guild, to then become *journeymen* [19], and eventually develop to *master craftsmen* [46]. Guilds prided themselves on their collective reputation, which was an aggregate of their members' progress through the quality system, and for high quality work, which enabled them to demand premium prices. Some guilds even fined members who deviated from the guild's quality standards [47].

Today, the intellectual inheritance of guilds persists via professional organizations, which replicate some of their benefits [29]. Professions such as architecture, engineering, geology, and land surveying require varying lengths of apprenticeships before one can gain professional certification, which holds great legal and reputational weight [48, 35]. However, these professions fall into traditional work models and do not cater for global and location-independent flexible employment [36]. Non-traditional work arrangements such as freelancing do not provide common work benefits such as economic security, career development, training, mentoring and social interaction with peers, which are legally essential for classification as full-time work in many cases [27]. Although the support of professional organizations exists for freelance work, much of the underlying value of guilds does not.

Guilds re-emerged digitally in Massively Multiplayer Online games (MMOs) and behave somewhat like their brick-and-mortar equivalents [52]. Guilds help their players level their characters, coordinate strategies, and develop a collectively recognized reputation [12, 26]. This paper draws on the strengths of guilds in assessing participants' reputation, and adapts them to the non-traditional employment model of crowd work. Crowd guilds formalize the feedback and advancement system so that it can operate at scale with a distributed membership.

Guilds offer an attractive model for crowd work because they provide reputation information for distributed professionals and carry a collective professional identity. They could also be used to train their own members [62, 53, 28], and may eventually give workers the opportunity to take collective action in response to issues with requesters or the platform, all of which are characteristics notably missing from today's crowdsourcing ecosystem.

Collective action

Despite being spread across the globe, crowd workers collaborate, share information, and engage in collective action [17, 66, 18]. The Turkopticon activist system, for instance, provides workers with a collective voice to publically rate requesters [24], though it remains a worker tool, external to the

Mechanical Turk platform. Along with Turkopticon, workers frequent various forums to share information identifying high-quality work and reliable requesters [43]. However, these forums are hosted outside the marketplace. This de-centralization from the main platform makes it harder to locate and obtain reputational information [25] and, when needed, bring about large scale collective action. Therefore, Dynamo identified strategies for small groups of workers to incite change [57].

In existing marketplaces, workers are frustrated by capricious requester behavior related to work acceptance and having limited say in cultivating platform policies [9, 30]. Issues related to unfair rejections, unclear tasks, and platform policies have been publicly discussed [43], but workers have limited opportunities to impact platform operations leaving less room to accommodate emerging needs. Therefore, crowd guilds focus on providing peer assessed reputation signals. To do this, they internalize affordances of previous tools for peer-review and gathering feedback, and give the guild power in the marketplace to manage some of these issues.

CROWD GUILDS

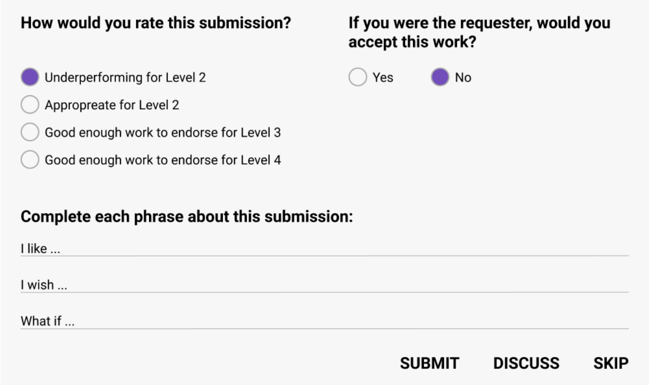
In this section, we describe technical infrastructure for *crowd guilds* in a paid crowdsourcing marketplace, enabling collective evaluation of members' reputations via peer feedback.

A crowd guild is a group of crowd workers who coordinate to manage their own reputation. As a research prototype, we have implemented the guild structure with the Daemo open-source crowdsourcing platform [15]. Crowd guilds could be built to focus on specific types of tasks and cultivate expertise for that labor. With a single platform-wide guild, however, we focus on the visible and invisible aspects of microwork, such as the expertise necessary to perform Mechanical Turk tasks effectively [43]. We thus designed the crowd guilds in active collaboration with workers on Amazon Mechanical Turk forums. Our discussions with workers on forums and over video chat led to several design decisions in crowd guilds (e.g., payment for peer assessment tasks).

Crowd guilds introduce a peer assessment infrastructure allowing workers to review each other's work, provide feedback via work critiques, and establish publicly-visible guild worker levels (e.g., Level 2, Level 3). To complement crowd guilds and explore the design space, we have also created prototype implementations of other behaviors that guilds can support: collective social spaces for informal engagement [67], group determination of appropriate wage levels, and collective rejection of inappropriate work.

Reputation via peer assessment and guild leveling

Because each worker on a crowdsourcing platform is essentially a freelancer, reputation information (e.g., acceptance rates and five-star ratings) is generally uncoordinated and often highly inflated [30, 22]. Historically, in similar situations, guilds stepped in to guarantee the quality of their own (similarly independent) members, just as professional societies do today. Thus, crowd guilds form around a community of practice [65] that can effectively assess its own members' skills. Crowd guilds manage their members' reputation by assigning public guild levels to each member. To do so, our infrastructure randomly samples guild members' work, anonymizes it, and routes it to more senior members for evaluation to determine each worker's appropriate level.



How would you rate this submission?

● Underperforming for Level 2
○ Appropreate for Level 2
○ Good enough work to endorse for Level 3
○ Good enough work to endorse for Level 4

If you were the requester, would you accept this work?

○ Yes ● No

Complete each phrase about this submission:

I like ...

I wish ...

What if ...

Figure 2. Crowd guilds peer assess a random sample of their own members' work in order to determine promotion between levels. The assessment form asks the double-blind reviewer to rate the work relative to the worker's current level, and give open-ended feedback.

Peer review

Our system automatically triggers double-blind peer reviews of tasks completed by guild workers on the crowdsourcing platform. Online peer assessment [32] is in regular use for filtering out low-quality work in crowdsourcing [40, 39]—crowd guilds adapt this process for use in promotion and reputation. Peers and professionals are accurate at assessing each other [13, 63], but people trust information more when it comes from those with more experience and authority [49]. Thus, a worker's peer reviews in a crowd guild can only be completed by guild members ranked one level above: for example, Level 2 workers can only be assessed by Level 3 workers.

To generate reviews, the crowd guilds infrastructure randomly samples a percentage of each worker's task submissions. Each sampled submission is wrapped inside an evaluation task and posted back onto the platform as a paid assessment task available to qualified members of the guild.

The reviews are double-blind—the workers do not know who reviewed their work, or which of their tasks will be reviewed, and the reviewers do not know which worker they are reviewing. As a result, workers cannot strategically increase their work effort to get more positive reviews, and reviewers have little opportunity to be selective about their reviews to benefit particular parties. Because reviews are conducted by guild members one level above the worker being reviewed, the quality expectations of the reviewer are not too distant from those of the worker, and when a guild member's level increases, the quality standards to proceed to the next level are proportionately higher. Quality standards are not explicitly set for any level, but are interpreted by those in the level as their reviews decide who will join them.

Review tasks consist of a copy of the original task, the worker's response and three review questions (Figure 2):

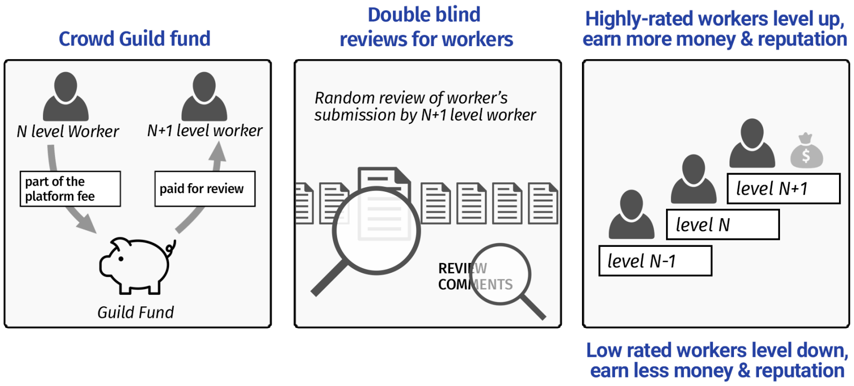

Crowd Guild fund
N level Worker
N+1 level worker
part of the platform fee
paid for review
Guild Fund
Double blind reviews for workers
Random review of worker's submission by N+1 level worker
REVIEW COMMENTS
Highly-rated workers level up, earn more money & reputation
level N+1
level N
level N-1
Low rated workers level down, earn less money & reputation

Figure 3. Review and leveling combine to form reputation ranks. Reviews are funded through the platform and are conducted, double-blind, on random tasks. The level of the worker is adjusted as the moving average of recent reviews meets thresholds. The moving average of recent reviews is used to adjust the level of the worker if they meet a threshold.

1. A four-point ordinal scale for evaluating the quality of the work. The scale points are calibrated to the worker's current level n, and ask the reviewer to rate the results as: a) Subpar: appropriate for level $n-1$, b) At par: appropriate for level n, c) Above par: appropriate for level $n+1$, or d) Far above par: appropriate for level $n+2$. Workers collectively develop criteria for each level as the guild matures. Over time, such ratings will determine whether the worker stays, moves up or moves down in the guild levels. The inclusion of a $n+2$ assessment option was based on crowd worker feedback that it should be possible to rapidly level up after a skilled worker joins a guild.

2. A question asking if the worker thinks the requester is likely to accept the work. Framing focusing on the requester's perspective ("If you were the requester...") results in lower variance responses compared with asking for egocentric reviews ("Would you accept this?") [16]. This question is not expected to be used to adjust actual acceptance of work, but gives useful insight into workers' perceptions of requesters' interests.

3. An open-ended text field asking how the worker might improve their work in the future. This field is designed to enable critique, utilizing a "I Like, I Wish, What If" model drawn from design thinking (e.g., [64]) because of its efficacy in motivating high quality responses that are actionable and prosocial.

To make it less likely that senior guild members exercise unfair power, review tasks are also included in the tasks sampled for review. Reviews themselves get reviewed—a form of meta-reviewing. These meta-reviews are completed by all guild members, not just higher-ranked members. If a guild member is reviewing unfairly, others in the guild can recognize and punish the behavior. Meta-reviews also ensure that the quality standards for a level are reasonable.

Review tasks are paid tasks on the platform. Funds for review could come from the requesters, workers or platform, and could operate like a subscription, tax, or donation. Requesters are familiar with paying platform fees such as on Amazon Mechanical Turk, while platforms already invest in attracting workers and requesters; each stand to benefit from crowd guilds. However, the workers benefit most directly through increased wages as a result from leveling. Donation models can lack stability, and subscriptions can suffer from a lack of granularity, while a tax on individual tasks avoids these problems. In our design, we chose for crowd guilds to exact a marginal cost per task from worker earnings. In practice, this means that as workers complete tasks, funds will accumulate to pay for a review of one of those tasks, selected at random.

Based on initial pilot analysis to establish the average time per review for several different task types, charging 10% per task and conducting a review after 10 tasks have been completed is a default that serves both reviewers and workers well. Reviews for many task types are significantly faster than the original task. Because review tasks are fast, a 10% overhead is enough for the more expensive higher-leveled workers to perform the review. However, an algorithm could be designed to tune these defaults dynamically, if task duration and review duration were measured by the platform. For example, in the long term, 10% may in fact be oversampling the reviews, as workers perform hundreds of tasks per day.

In practice, this system must be adjusted to deal with cold-start problems and privacy issues. If there are no workers of higher reputation (e.g., if a guild has just started or if a top-level worker is being reviewed), peers at the same level can review the task. In addition, some requesters do not want other workers seeing their tasks in order to protect private information. For this reason, Daemo requesters can opt out of their tasks being used for review purposes, which will avoid tasks being seen by anyone other than the initial worker, but this will also mean that additional information gained by the review process will not be returned to them.

Levels

Levels (Figure 3) provide trusted reputation signals to requesters on the platform. All workers begin at Level 1.

Promotion is determined using the peer review feedback. The four-point ordinal scale from peer review is converted into a numeric scale 1–4, and a moving numeric average is calculated across a window of 10 reviews. When a worker's moving

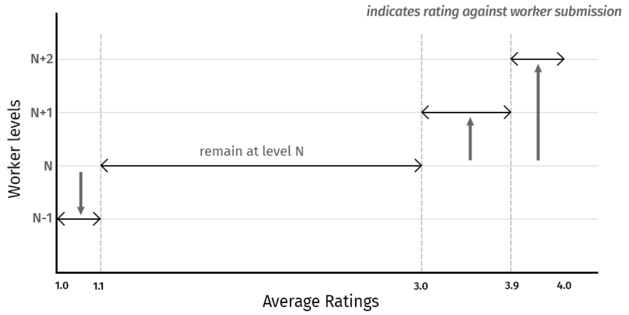


Figure 4. Example of leveling thresholds. The average review on a 1-4 scale, is used to choose one of 4 level shifts relative to the worker's current level: move down one level, remain in the same level, move up one level, or move up two levels.

average reaches a threshold such as those shown in Figure 4, their level will be updated. After every level change, the moving average is reset. Unlike Mechanical Turk's acceptance rate, in which a blunder can mean a permanent mark on a worker's reputation, crowd guilds consider a moving average so as not to hold workers back by their past mistakes. On the other hand, a worker can also be leveled down if they consistently produce low quality work. In this way, workers have an incentive to continue doing good work.

The thresholds in Figure 4 only level down if work quality is particularly low, rated 1 out of 4 for 90% of reviews. This set of thresholds has been used for our current deployment and has worked effectively as we are testing with a small number of workers, but the thresholds should be tuned to deal with large numbers of workers and large numbers of tasks to perform at scale.

Other approaches to leveling we considered include a periodic review schedule or a manual review board. A periodic review cycle ties workers to a timeline that is not related to actual changes in their quality and can move too slowly for active and skilled workers. A manual board voting on promotions, as in Wikipedia [38], would incur a larger cost overhead and increase the risk of an oligarchic regime managing the guild. Thus, we chose a continuous random sampling of work.

Exploring social effects

Historically, guilds did not restrict their attention to reputation management; they also engaged in a broader set of collective behaviors to support their trade and their members' welfare. While our attention in this paper is focused on reputation, crowd guilds can also explore other collective behaviors. We present several such prototypes: informal social engagement through collective spaces, group determination of wage levels, and collective rejection of inappropriate work.

Level-based pricing

If requesters can trust workers to be higher quality, then they can rely less on redundancy and pay individual workers more. When a requester posts a new task on Daemo and selects a guild level to target, higher levels show higher wage recommendations, based on the average hourly wage of workers at that level. The average hourly wage is calculated in aggregate over all members of a level, only counting work they perform that was posted to that level. This offers a useful insight into how much workers expect to make from a task—wage is often hard for requesters to make a judgement about without a trial and error process [8]. Leaving this as a recommendation, and not a requirement, allows for some flexibility and for the pricing scheme to organically evolve.

Collective rejection

Because some requesters routinely disregard ethical wage standards, crowd guilds also provide workers with a task rejection mechanism. Guild workers are able to collectively reject tasks if they feel the price is unfairly low for the requested level. When a worker rejects the task from their own feed, the task is no longer available for that worker, and the requester is sent a notice. However, when more than a small percentage of workers from a level decide to reject a task, then the task is removed from everybody's task feed in the guild. With this design, requesters are incentivized to price their tasks favorably enough to match the expectations of the worker community and not underprice their tasks. In our current implementation, 3% of workers rejecting a task removes it from the platform. 3% was chosen arbitrarily and we hope to consider ways to optimize this value in the future.

Open forum

Forums are widely used by the crowdsourcing community. However, many of the most popular ones are external to the platforms they serve, such as Turker Nation and the /r/mturk subreddits [43]. The crowd guild forum is linked from within Daemo platform: "Discuss" links are available within each task to enable workers to quickly begin discussions focused on specific issues. Workers' reputation levels are publicized with badges, and separate badges are provided identifying requesters and administrators, affording direct communication and helping to humanize the platform.

EVALUATION

Crowd guilds introduce a system for crowd workers to communicate unified reputation signals and share feedback. To provide initial insight into guilds' ability to achieve this goal, we performed a two-week field experiment in which 300 crowd workers were randomized into either a crowd guild or a traditional decentralized design within the Daemo crowdsourcing marketplace. This evaluation strategy seeks evidence that crowd guilds can form meaningful reputation signals, focused on a medium-length field deployment. We frame our evaluation around the following hypothesis:

> *Peer assessment in crowd guilds produces more accurate reputation information than peer assessment when a guild is not present.*

Participants

We conducted a field experiment over a two-week period. We recruited 300 workers from Amazon Mechanical Turk to perform as many tasks as they wanted to on Daemo. They were paid for all tasks completed on the platform. Our recruitment task required that participants have more than 1000 accepted tasks on Mechanical Turk. N = 196 completed the two-week study. A new set of tasks was released on the platform each morning.

The 196 participants had an average of 78,965 tasks accepted on Amazon Mechanical Turk before starting our experiment, were 56% female, had an average of 14.3 years of education, had been working on Amazon Mechanical Turk for an average of 2 years and had an average acceptance rate of 99.7%. Participants reported using 3 external platforms or tools to facilitate their crowd work on average. Most popular were Mechanical Turk, TurkOpticon (85%), Reddit (40%); the least popular were CrowdFactory (0%), LeadGenius (0%), and CrowdFlower (5%). 15% reported using Facebook for crowd work related activities. The participants were paid for all their work submissions in line with the current ethical standards for research on Mechanical Turk [57]. Participants earned an average of $10 per hour for tasks they performed as part of this study.

Experimental conditions

Participants were randomized into treatment and control conditions. Both conditions were hosted on isolated instances of Daemo and had their own connected discussion forum. Both groups were given the same tasks with the same schedule. Each worker saw an average of 10 tasks per day that would take approximately 10 minutes of work in total. In order to simulate a diverse crowdsourcing platform, the tasks included a range of popular microtasks [23]: tagging image contents, analysing the sentiment of tweets, writing product reviews, and assessing positions of online articles. The tasks were posted from a variety of different requester accounts to make the experiment appear realistic. The peer-review interface remained the same for all types of tasks during the study, meaning, for example, that sometimes a peer could provide assessment of a hand written product review, while other times they would be assessing a single forced choice response.

The control condition replicated current best practices on microtask crowdsourcing platforms. Control workers were invited to the study with no mention of guilds or leveling. Some tasks presented to control workers involved the review of other workers' submissions, however, these appeared identical to a standard task on the system; similar types of verification tasks are common on microtasking platforms (e.g., [2]).

The treatment (crowd guilds) condition used a separate instance of Daemo that included all the features of the control condition with the addition of crowd guild leveling. Workers in this condition were informed of the leveling features through in-context tooltips and forum posts. All participants in the treatment condition started as Level 1 workers. To kickstart leveling, at the end of the second day of the study, we reviewed submissions and selected 10 workers to promote to Level 2 manually. All other level adjustments during the study occurred automatically based on the leveling thresholds demonstrated in Figure 4. To achieve a spread of levels more quickly during the study period, reviews were sampled for 20% of submissions.

Evaluating crowd guilds

To measure accuracy, we used mostly close-ended tasks with ground truth answers. For open-ended tasks, such as product reviews, we used rubrics developed in prior work [11], and then asked five other workers to evaluate the submission according to the rubric, taking the modal response as the accuracy score. We calculated the workers' accuracy as the percentage of their tasks answered correctly (0-100%).

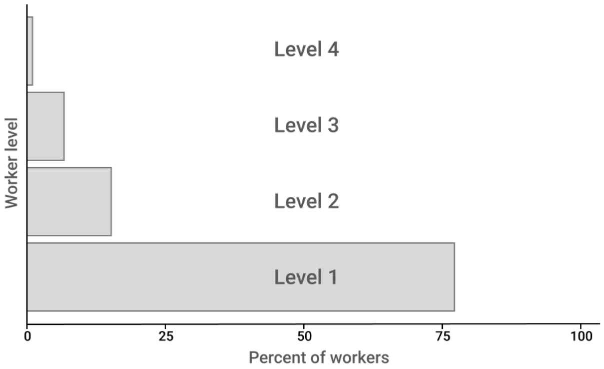


Figure 5. By the conclusion of the study, the crowd guild had evaluated its workers and distributed them across four levels.

At the conclusion of the study we measured how many posts workers made on the forums, and how useful they thought feedback from task reviews was (5-point Likert scale). These factors help us understand any auxiliary effects crowd guilds may have had. Additionally, we conducted exploratory content analysis on both the forum and review feedback to understand how these features were used differently when crowd guilds were present.

We then used regression analysis to investigate how well workers' accuracy could be predicted by their mean peer assessment scores obtained through the peer-review process in our study, their condition in the study (1 = crowd guild, 0 = control), the number of posts they shared on the forum (a proxy for their engagement in the guild social environment), and the self-reported usefulness of the guild's feedback (5-point Likert scale). We added controls for their approval rate on Amazon Mechanical Turk (0-100%), as well as their total number of accepted tasks on Mechanical Turk. A multiple regression allows us to understand which variables were significantly associated with ground-truth accuracy.

RESULTS

After two weeks in the study, workers in the crowd guild condition spread themselves into Levels 1 to 4, with 151 of workers at Level 1, 30 in Level 2, 13 in Level 3 and 2 in Level 4 (Figure 5). During this time, workers in the guild condition completed a total of 15176 tasks, with an average of 87 tasks per worker. Workers in the control condition completed a total of 13427 tasks, with an average of 113 tasks per worker.

Crowd guilds generate accurate reputation signals

We analyzed the relationship between workers' ground-truth accuracy and the reputation signals accessible to the system. Table 1 contains the estimate, std. error, and p-values for the associated variables (*adj.* $R^2 = 0.038$). Mean peer assessment rating was a significant predictor of ground truth accuracy ($\beta = 4.1$, $p < 0.005$), supporting the hypothesis that crowds can collectively author accurate reputational signals.

Coefficient	Value	Error	t-value	p-value
Approval rate	1.98	240	-0.59	0.55
Accepted tasks	0.00	0.00	0.21	0.83
Mean peer assessment	4.08	1.39	2.93	**0.004**
Feedback usefulness	-0.34	0.75	-0.45	0.65
Forum activity	0.09	0.09	1.01	0.31
Condition	3.00	1.66	1.81	0.07

Table 1. A regression predicting workers' ground truth accuracy uncovered significant effects of average peer review score (1–4), verifying that continuous peer review can be used to establish accurate reputation credentials for workers.

Traditional Mechanical Turk reputation signals were not significantly correlated with ground truth accuracy, nor were the self-reported guild feedback usefulness, nor the forum activity level (all p > 0.05). Study condition (coded as 1 for guilds) trended toward significance p = 0.07, suggesting that the workers in the guilds condition may have produced higher-quality results on average.

Was there a difference in the accuracy of peer feedback between conditions? The distribution of the overall peer assessment ratings of each worker at the conclusion of the study is shown in Figure 6. Testing correlations between workers' ground truth accuracies and their average peer assessment ratings, there was a significant correlation in the guilds condition (p < 0.001, r = .36) and no significant correlation in the control condition (p = 0.09, r = .18). This result suggests that feedback in the guild condition was more informative, again supporting the hypothesis that the guilds condition produced more accurate reputational information than the control condition.

To understand why crowd guilds produced more accurate reputation scores, we address two questions. First, what differed about the reputation scores between conditions? There is a statistically significant difference in the overall peer assessment ratings between conditions (t(130.12) = 6.33, p < 0.001), with reviews in the guild condition (µ = 2.5) significantly less inflated than those in the control condition (µ = 3.0). Reputation inflation is a challenge in crowdsourcing marketplaces [22, 14], so a lower average score is preferable, which indicates that reviewers in the guilds condition were more discerning and spare with high ratings. Why does this difference occur? We hypothesize that workers in each condition interpreted the meaning of each rating scale level differently based on the perceived impact of ratings. This rating deflation and discernment was the likely mechanism behind the existence of a correlation between peer assessment and accuracy in the guild condition but not the control condition: in the control condition, score inflation caused workers of different ground-truth accuracies to have similar average feedback scores.

The second question: what about the crowd guilds design was responsible for producing more accurate ratings than the control condition? Potential hypotheses include the content of the free-text feedback, engagement on the forums, or the ratings themselves. The self-reported feedback usefulness and forum activity variables in the regression provide some insight into this mechanism. Neither feedback usefulness nor forum activity were significant predictors of worker accuracy (both p > .05). This result suggests that the community significance of crowd guild ratings, the only other major feature in this system, was the main influencer leading to improved rating accuracy. In other words, current evidence suggests that crowd guilds produce more accurate reputation signals because workers treat the ratings differently when the ratings determine other members' levels. The guild's social structures and textual feedback may produce other benefits (e.g., a collective identity, ideas for improvement), but they do not appear to directly influence rating accuracy.

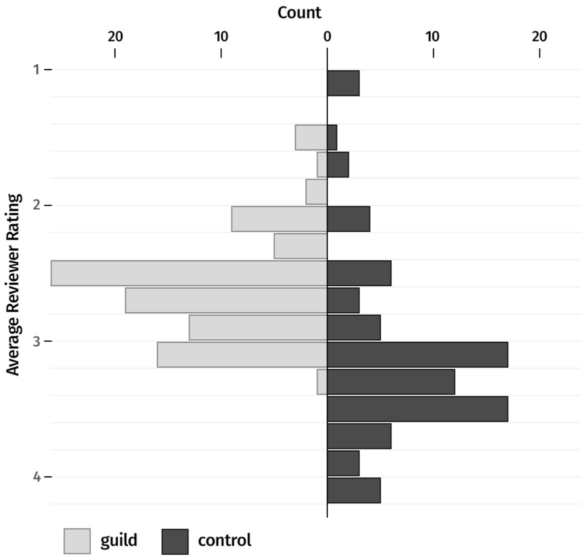


Figure 6. Workers' peer assessment ratings in the guild condition were less inflated than those in the control condition at the end of the study.

Cumulatively, these results support the hypothesis that crowd guilds produce more accurate reputation information than no guilds, and furthermore that they produce more accurate reputation information than current signals on Mechanical Turk.

Assessing crowd guilds' qualitative impacts

The results so far have investigated whether crowd guilds produce more accurate reputational information. However, our study also included other prototype community and feedback mechanisms, and it is important to understand the guilds' effect on the *community* as well as on the reputation system. In this section, we report a mixed-methods analysis of the feedback, forums, and survey results to understand the differences in workers' community behavior between conditions. We observed that incentives pragmatize workers' advice-giving behaviour, that community integration can have strategic values for workers and requesters, and that at least some workers argue that decentralization is a benefit of crowdsourcing and should not be meddled with. Here we focus on qualitative observations made during the study on each condition's community forum and peer assessment feedback.

Crowd guilds pragmatize feedback

Workers in both the treatment and control communities were mutually supportive. While conversations involving moderators tended to relate to bug reports or bringing up specific issues in a task, nearly all other threads were focused specifically on sharing know-how or awareness of platform features.

In the guilds condition, workers often focused on more pragmatic feedback on strategies for instrumental and informational support, while control workers were more likely to offer emotional support rather than information. For example, when one worker in the control condition raised issues they were facing, another control worker empathized and replied:

> *I'm sorry. hugs I hope your day gets better. Maybe there is something the people running the study can do to fix the problem.*

On the other hand, guild workers responded to similar situations with pragmatic advice that focused on the challenge being faced. When a worker expressed concern that review tasks were not available to everyone, the response provided direct informational support, for example:

> *I have been able to do review tasks since being upped to level 2 /shrug*

This difference was also seen in the peer review feedback as part of task reviews. Reviews in the guild condition were fewer characters on average ($\mu = 58.71$, $\sigma = 76.94$ vs. $\mu = 68.66$, $\sigma = 69.60$ in control, Kruskal-Wallis $p < 0.001$) and generally more focused and pragmatic. In the guild condition, responses used to let a worker know that they had done part of a task incorrectly were very direct:

> *Did not appropriately answer the question by specifying how often they used each tool.*

In the control condition, review responses exhibited doubt and consideration about the worker's thinking:

> *The page doesn't say or give any clue other than people growing marijuana as to whether they are pro-legalization or not. Because of this, I personally would have marked neither. But, then again, I could be wrong. Marking neither is what I would have done to improve it.*

This difference, when considered in connection to the improved accuracy of reviews in the treatment condition mentioned above, suggests that crowd guilds improve the effectiveness of peer review by professionalizing it. In contrast, we had expected that guilds would result in a more supportive environment in their peer review. When the community was filled with workers who have control over each other's reputation, the community behaviors became more pragmatic.

Love it or hate it: workers' preferences about centralization

Crowd guilds offer improved worker credentialing and feedback through re-centralization of reputation. In general, workers on the forum appreciated and favored these goals. A few workers in the guild condition, however, indicated a preference for an entirely decentralized platform. These "lone wolf" workers valued the independence of crowd work and being masters of their own fate. Lone wolf workers would often criticize guild features such as review and leveling, on the basis that they would remove independence from the worker. In our formative design feedback from crowd workers on Mechanical Turk forums, this sentiment also arose amongst a few members of the community. However, the significant majority of workers in the community expressed high levels of interest in this platform becoming a reality. We thus hypothesize that workers who have found success on Mechanical Turk as it is today see increased interdependence as a threat to their livelihood, stability and freedom.

The most salient complaints from workers in the guild condition were that they did not necessarily trust the reviewers and the leveling systems to serve the interests of the community:

> *It seems to me that people are overly judgmental in order to show that they deserve to be in Level 2 rather than in worrying about whether the rest of us should be.*

However, our quantitative analysis suggests that reviews on the treatment condition outperform the control condition in predicting actual quality of work.

Crowd guilds replace an *algorithm* (acceptance rate) with a *social process* (peer assessment), and that social processes can legitimately trigger disagreements and concerns between community members. It replaces an "us vs. them" dynamic (workers vs. requesters) with an "us vs. us" dynamic (Level 1 workers vs. Level 2 reviewers). These interdependencies between workers can be good—they enable high-quality reputation signals and community feedback, for instance—but they can also cause strife and disagreements. However, if the guild develops clear metrics, norms and accountability processes over time, it can overcome many of these issues, and based on our qualitative analysis, the benefit far outweighs the cost.

DISCUSSION

In this section, we reflect on the methodological limitations of our experiment, the major design challenges in organizing a crowd guild for a paid crowdsourcing platform, and the future directions of guilds in the crowdsourcing design space.

Limitations

The experiment suggests that crowd guilds enable workers to collectively manage their reputation metrics. Our evaluation was able to populate a guild with 104 workers, but large-scale crowdsourcing platforms have tens of thousands of active workers. We are unable to observe how crowd guild dynamics would play out at this scale, or over very long time periods, although crowd workers have been shown to maintain consistent performance over time [21]. These efforts remain future work. It is possible that the most effective route will be to allow workers to create their own guilds, and join as many as they wish, to avoid massive guilds that are not differentiable.

External validity is also limited by Daemo's existence as a small-scale crowdsourcing platform today. We paid workers to join the evaluation; we cannot know for certain that workers would actually prefer a crowd guild like on Daemo to their existing platforms. For crowd guilds to succeed, high-quality workers and requesters must remain in the social system for long periods of time and thus have a stake in making the social system useful. An ideal evaluation would involve workers and requesters who hold such a stake, either by changing Amazon Mechanical Turk, or recruiting long-term workers and requesters onto the Daemo platform. Since Daemo is not fully evolved to conduct the entire functional experiment with real requesters and does not already have a workforce, we recruited users from Mechanical Turk. This decision made study less realistic; however, because it minimized participants' long term motivations, it is also biased toward a conservative estimate of the strength of the effect. There are possible novelty effects which will need to be teased out in future work via longitudinal studies and qualitative analysis. Finally, although we tried to recruit a wide variety of real crowd workers with real tasks from crowd marketplaces, our results might not yet generalize to crowd work at large.

The ramifications of crowd guilds

Introducing a sociotechnical system such as crowd guilds will inevitably shift the social and power dynamics of the crowdsourcing ecosystem. Some of these shifts are already visible. For example, the study made clear that guilds shifted the forum away from just being a water cooler and toward a work environment. In addition, the lone wolf workers were less enthused about the prospect of their reputation riding on other workers' behaviors.

We might extrapolate from these visible shifts to ones that may occur in the longer term. There will clearly be instances where workers' peer evaluations are inaccurate, and these unfair ratings may sow distrust within the guild. Over time, the guild levels may ossify and become oligarchical [58], further making workers feel distrustful of each other. It remains to be seen whether the meta-reviews, or the ability to split off and form a separate guild, might keep these pressures in balance.

Finally, while we focused our prototype on guilds as a vehicle for reputation, they could grow to encompass other functions as well. We prototyped several of these functions—collective rejection and wage setting, for example. We intend these prototypes to stand as example guild functions based on the roles that previous guilds played. However, crowd guilds may introduce entirely new roles not seen in historical guilds. Could they produce new forms of training? New social environments? Stronger relationships amongst crowd workers?

Future work

The future of crowd work depends on strong worker motivation, feedback, and pay [30]. There exist many other mechanisms for achieving these goals, especially increasing a sense of belonging instead of isolation. Crowd guilds begin with a narrow goal and pragmatic design, aimed to provide mechanisms for stronger reputation for workers and thus fairer pay on the platform. A second step is to translate this centralization into increased worker collective action in solving problems such as asymmetric access to information, limitations on open innovation and governance problems within the online labour marketplace [57]. More broadly, we believe that effective social computing designs can go further in enabling more prosocial and equitable environments for crowd workers. One worker mentioned:

> *"I was just thinking that it would be very helpful/useful to have some type of notification that signifies if a Forum Member is Currently Active/On or if they are Away etc.. If we knew that specific Members were Online, we would feel more connected and have a sense that we have someone to turn to in real time."*

Just as forum designs affect the trust in crowd worker forums [34], it is important to consider how guilds design will influence governance. How will crowd guilds govern themselves long term? The guild in our experiment had no leadership structure, but if it were to persist, it would need to develop one. What policies can they control, how do they make decisions, and how do they collect and redistribute their own income? These questions combine social computing design and political science. In the future, we hope to analyze variation in guild governance policies to better understand the forms of self-governance that predict long-term engagement and satisfaction.

Finally, in the current design, the guild represents all the workers on the platform: where every worker is a part of the guild, and they all collectively assess everyone's work. However, reviewers may not have the domain knowledge in every subject area to produce a quality review. We envision that different guilds will form around particular communities of practice.

CONCLUSION

Crowd workers in microtask platforms have been decentralized in order to reap efficiencies from independent work. However, in the long term, it is crucial for workers to have opportunities to connect with each other, learn from each other, and impact the platforms they use. In order to address this, we have drawn upon the historical example of guilds to bring workers into a loose affiliation that can certify each other's quality. Thus, this paper introduces crowd guilds, a system that empowers worker communities with peer assessed reviews with feedback leveling, and a connected community forum. Our evaluation of crowd guilds demonstrated that crowd guilds lead to improved reputation signals and community behaviour shifts toward efficient feedback. More generally, crowd guilds offer opportunities to co-design crowdsourcing platforms with worker platforms.

ACKNOWLEDGEMENT

We thank the following members of the Stanford Crowd Research Collective for their contributions: Alison Cossette, Trygve Cossette, Ryan Compton, Pierre Fiorini, Flavio S Nazareno, Karolina Ziukloski, Lucas Bamidele, Durim Morina, Sharon Zhou, Senadhipathige Suminda Niranga, Gagana B, Jasmine Lin, William Dai, Leonardy Kristianto, Samarth Sandeep, Justin Le, Manoj Pandey, Vinayak Mathur, Kamila Mananova, Ahmed Nasser, Preethi Srinivas, Sachin Sharma, Christopher Diemert, Lakmal Rupasinghe, Aditi Mithal, Divya Nekkanti, Mahesh Murag, Alan James, Vrinda Bhatia, Venkata Karthik Gullapalli, Yen Huang, Armine Rezaiean-Asel, Aditya V. Nadimpalli, Jaspreet Singh Riar, Rohheit Nistala, Mathias Burton, Anuradha Jayakody, Vinet Sethia, Yash Mittal, Victoria Purynova, Shrey Gupta, Alex Stolzoff, Paul Abhratanu, Carl Chen, Namit Juneja, Jorg Nathaniel Doku, Anmol Agarwal, Chen Xi, Zain Rehmani, Yashovardhan Sharma, Yash Sherry, Sree Nihit Munakala, Shivi Bajpai, Shivam Agarwal, Sanjiv Lobo, Raymond Su, Prithvi Raj, Nikita Dubnov, Kevin Viet Le, Klerisson Paixao, Haritha Thilakarathne, David Thompson, Varun Hasija, Lucas Qiu, Kushagro Bhattacharjee, Karthik Paga, Ishan Yelurwar, Archana

Dhankar, Aalok Thakkar. We also thank the Amazon Mechanical Turk workers who participated in this study. This work was supported by NSF award IIS-1351131, Office of Naval Research award N00014-16-1-2894, Toyota and the Hasso Plattner Design Thinking Research Program.

REFERENCES

1. Anderson, M. Crowdsourcing higher education: A design proposal for distributed learning. *MERLOT Journal of Online Learning and Teaching*, 7(4):576–590, 2011.
2. Bernstein, M., et al. Soylent. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology-UIST'10*, p. 313. 2010.
3. Billett, S. *Learning in the workplace: Strategies for effective practice.* ERIC, 2001.
4. Boud, D. Sustainable assessment: rethinking assessment for the learning society. *Studies in continuing education*, 22(2):151–167, 2000.
5. Boud, D. et al. *Enhancing learning through self-assessment*. Routledge, 2013.
6. Callison-Burch, C. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 286–295. Association for Computational Linguistics, 2009.
7. Campbell, J.A., et al. Thousands of positive reviews: Distributed mentoring in online fan communities. *arXiv preprint arXiv:1510.01425*, 2015.
8. Cheng, J., Teevan, J., and Bernstein, M.S. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1365–1374. ACM, 2015.
9. Deng, X.N. and Joshi, K. Is crowdsourcing a source of worker empowerment or exploitation? understanding crowd workers' perceptions of crowdsourcing career. 2013.
10. Dontcheva, M., Morris, R.R., Brandt, J.R., and Gerber, E.M. Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 3379–3388. ACM, 2014.
11. Dow, S., Kulkarni, A., Klemmer, S., and Hartmann, B. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 1013–1022. ACM, 2012.
12. Ducheneaut, N., Yee, N., Nickell, E., and Moore, R.J. The life and death of online gaming communities: a look at guilds in world of warcraft. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 839–848. ACM, 2007.
13. Falchikov, N. and Goldfinch, J. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research*, 70(3):287–322, 2000.
14. Gaikwad, S., et al. Boomerang: Rebounding the consequences of reputation feedback on crowdsourcing platforms. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 625–637. ACM, 2016.
15. Gaikwad, S.N., et al. Daemo: A self-governed crowdsourcing marketplace. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pp. 101–102. ACM, 2015.
16. Gilbert, E. What if we ask a different question?: social inferences create product ratings faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2759–2762. ACM, 2014.
17. Gray, M.L., Suri, S., Ali, S.S., and Kulkarni, D. The crowd is a collaborative network. *Proceedings of Computer-Supported Cooperative Work*, 2016.
18. Gupta, N., Martin, D., Hanrahan, B.V., and O'Neill, J. Turk-life in india. In *Proceedings of the 18th International Conference on Supporting Group Work*, pp. 1–11. ACM, 2014.
19. Guthrie, C. On learning the research craft: Memoirs of a journeyman researcher. *Journal of Research Practice*, 3(1):1, 2007.
20. Haas, D., Ansel, J., Gu, L., and Marcus, A. Argonaut: macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12):1642–1653, 2015.
21. Hata, K., Krishna, R., Fei-Fei, L., and Bernstein, M.S. A glimpse far into the future: Understanding long-term crowd worker accuracy. *arXiv preprint arXiv:1609.04855*, 2016.
22. Horton, J. and Golden, J. Reputation inflation: Evidence from an online labor market. *Work. Pap., NYU*, 2015.
23. Ipeirotis, P.G. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.
24. Irani, L.C. and Silberman, M. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 611–620. ACM, 2013.
25. Jøsang, A., Ismail, R., and Boyd, C. A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644, 2007.
26. Kang, J., Ko, I., and Ko, Y. The impact of social support of guild members and psychological factors on flow and game loyalty in mmorpg. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pp. 1–9. IEEE, 2009.
27. Karoly, L.A. and Panis, C.W. *The 21st century at work: Forces shaping the future workforce and workplace in the United States*, vol. 164. Rand Corporation, 2004.

28. Kaye, L.K. and Bryce, J. Putting the fun factor into gaming: The influence of social contexts on the experiences of playing videogames. *International Journal of Internet Science*, 7(1):24–38, 2012.

29. Kieser, A. Organizational, institutional, and societal evolution: Medieval craft guilds and the genesis of formal organizations. *Administrative Science Quarterly*, pp. 540–564, 1989.

30. Kittur, A., et al. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 1301–1318. ACM, 2013.

31. Kulkarni, A., et al. Mobileworks: Designing for quality in a managed crowdsourcing architecture. *Internet Computing, IEEE*, 16(5):28–35, 2012.

32. Kulkarni, C., et al. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(6):33, 2013.

33. Kulkarni, C.E., Bernstein, M.S., and Klemmer, S.R. Peerstudio: Rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pp. 75–84. ACM, 2015.

34. LaPlante, R. and Silberman, M.S. Building trust in crowd worker forums: Worker ownership, governance, and work outcomes. In *Proceedings of WebSci16*. ACM, 2016.

35. Larson, M.S. and Larson, M.S. *The rise of professionalism: A sociological analysis*, vol. 233. Univ of California Press, 1979.

36. Laubacher, R.J., Malone, T.W., et al. Flexible work arrangements and 21st century worker's guilds. Tech. rep., MIT Center for Coordination Science, 1997.

37. Lee, H. and Choi, B. Knowledge management enablers, processes, and organizational performance: An integrative view and empirical examination. *Journal of management information systems*, 20(1):179–228, 2003.

38. Leskovec, J., Huttenlocher, D., and Kleinberg, J. Signed networks in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1361–1370. ACM, 2010.

39. Little, G., Chilton, L.B., Goldman, M., and Miller, R.C. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*, pp. 68–76. ACM, 2010.

40. Little, G., Chilton, L.B., Goldman, M., and Miller, R.C. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 57–66. ACM, 2010.

41. Luther, K., et al. Structuring, aggregating, and evaluating crowdsourced design critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 473–485. ACM, 2015.

42. Malone, T.W. and Laubacher, R.J. How will work change? elancers, empowerment, and guilds. *THE PROMISE OF GLOBAL NETWORKS*, p. 119, 1999.

43. Martin, D., Hanrahan, B.V., O'Neill, J., and Gupta, N. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 224–235. ACM, 2014.

44. McInnis, B., Cosley, D., Nam, C., and Leshed, G. Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in amazon mechanical turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2271–2282. ACM, 2016.

45. Mitra, T., Hutto, C.J., and Gilbert, E. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1345–1354. ACM, 2015.

46. Mocarelli, L. Guilds reappraised: Italy in the early modern period. *International review of social history*, 53(S16):159–178, 2008.

47. Ogilvie, S. Guilds, efficiency, and social capital: evidence from german proto-industry. *Economic history review*, pp. 286–333, 2004.

48. Ogilvie, S. The use and abuse of trust: the deployment of social capital by early modern guilds. *Jahrbuch für Wirtschaftsgeschichte*, 1:15–52, 2005.

49. Patel, N., et al. Power to the peers: authority of source effects for a voice-based agricultural information service in rural india. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, pp. 169–178. ACM, 2012.

50. Peng Dai, M.D. and Weld, S. Decision-theoretic control of crowd-sourced workflows. In *In the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*. Citeseer, 2010.

51. Pérez, L. Inventing in a world of guilds: silk fabrics in eighteenth-century lyon. *Guilds, innovation, and the European economy*, pp. 1400–1800, 2008.

52. Poor, N. What mmo communities don't do: A longitudinal study of guilds and character leveling, or not. In *Ninth International AAAI Conference on Web and Social Media*. 2015.

53. Rees Lewis, D., Harburg, E., Gerber, E., and Easterday, M. Building support tools to connect novice designers with professional coaches. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pp. 43–52. ACM, 2015.

54. Renard, G. Guilds in the middle ages. 1918.

55. Rosser, G. Crafts, guilds and the negotiation of work in the medieval town. *Past & Present*, (154):3–31, 1997.

56. Rzeszotarski, J.M. and Kittur, A. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 13–22. ACM, 2011.

57. Salehi, N., et al. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1621–1630. ACM, 2015.

58. Shaw, A. and Hill, B.M. Laboratories of oligarchy? how the iron law extends to peer production. *Journal of Communication*, 64(2):215–238, 2014.

59. Sheng, V.S., Provost, F., and Ipeirotis, P.G. 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*, pp. 614–622. ACM, 2008.

60. Sparrowe, R.T., Liden, R.C., Wayne, S.J., and Kraimer, M.L. Social networks and the performance of individuals and groups. *Academy of management journal*, 44(2):316–325, 2001.

61. Surowiecki, J. *The wisdom of crowds*. Anchor, 2005.

62. Suzuki, R., et al. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. *arXiv preprint arXiv:1602.06634*, 2016.

63. Topping, K. Peer assessment between students in colleges and universities. *Review of educational Research*, 68(3):249–276, 1998.

64. Ulibarri, N., et al. Research as design: Developing creative confidence in doctoral students through design thinking. *International Journal of Doctoral Studies*, 9:249–270, 2014.

65. Wenger, E. Communities of practice: A brief introduction. 2011.

66. Yin, M., Gray, M.L., Suri, S., and Vaughan, J.W. The communication network within the crowd. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 1293–1303. International World Wide Web Conferences Steering Committee, 2016.

67. Yu, L., André, P., Kittur, A., and Kraut, R. A comparison of social, learning, and financial strategies on crowd engagement and output quality. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 967–978. ACM, 2014.

68. Zhu, H., Dow, S.P., Kraut, R.E., and Kittur, A. Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 1445–1455. ACM, 2014.