Modeling Task Complexity in Crowdsourcing

Jie Yang[*], Judith Redi[*], Gianluca Demartini[†], Alessandro Bozzon[*]
[*]Delft University of Technology, [†]University of Sheffield
[*]{j.yang-3, j.a.redi, a.bozzon}@tudelft.nl, [†]g.demartini@sheffield.ac.uk

Abstract

Complexity is crucial to characterize tasks performed by humans through computer systems. Yet, the theory and practice of crowdsourcing currently lacks a clear understanding of task complexity, hindering the design of effective and efficient execution interfaces or fair monetary rewards. To understand how complexity is perceived and distributed over crowdsourcing tasks, we instrumented an experiment where we asked workers to evaluate the complexity of 61 real-world re-instantiated crowdsourcing tasks. We show that task complexity, while being subjective, is coherently perceived across workers; on the other hand, it is significantly influenced by task type. Next, we develop a high-dimensional regression model, to assess the influence of three classes of structural features (metadata, content, and visual) on task complexity, and ultimately use them to measure task complexity. Results show that both the appearance and the language used in task description can accurately predict task complexity. Finally, we apply the same feature set to predict task performance, based on a set of 5 years-worth tasks in Amazon MTurk. Results show that features related to task complexity can improve the quality of task performance prediction, thus demonstrating the utility of complexity as a task modeling property.

Introduction

The concept of *task* is central to human computation and crowdsourcing in general. It refers to a set of activities to be performed by an individual, mapping to an operation, control or synthesis process in a human computation algorithm (Law and von Ahn 2011). Tasks play a significant role for all the actors involved in crowdsourcing campaigns: their designers – i.e. the *requesters*; their executors – i.e. the *workers*; and the *platform* (e.g., Amazon MTurk), which hosts them and enables their discovery, assignment, and reward.

Studying task properties is key for addressing core crowdsourcing problems such as task assignment and optimization, and worker retention (Yang and Bozzon 2016). To this end, previous research extensively covered the study of task meta-properties, e.g., setting appropriate rewards (Cheng, Teevan, and Bernstein 2015a), with the goal of improving the effectiveness of human computation algorithms, and the quality of crowdsourcing output in general. *Time* is the dimension along which the human cost (e.g., fatigue, stress) associated with the execution of crowdsourcing tasks has been typically quantified. Incentives such as monetary rewards strongly depend on the estimation of task completion time. *Complexity*, instead, reflects not only temporal demands of task execution, but more holistically the real cognitive effort that performers need to put into the completion of tasks. Interestingly, this task property has so far been mostly neglected by crowdsourcing research.

Objective *Complexity* has been identified as one of the most important task properties in a variety of fields studying the relationship between humans and computer artifacts. It is a construct that is widely used in behavioural sciences, towards investigating task cognitive load; nevertheless, it is surprisingly difficult to describe. To advance the theory and practice of crowdsourcing, we advocate the need for a systematic examination of how task complexity is perceived by workers, what contributes to task complexity, and of how it affects execution performance and output quality.

Being able to measure and predict task complexity can be highly beneficial for both workers and requesters. Example applications include more accurate reward estimation for a given task, better task routing approaches that take into account worker skills and expertise, and simpler worker reputation management by letting workers decide at which level of complexity to work without risks of compromising their reputation (e.g., a worker may want to start working on tasks with low complexity and increase it as she becomes more confident). Crowd answer aggregation (Venanzi et al. 2014; Sheshadri and Lease 2013) and evaluation of worker performance (Jung and Lease 2015a; 2015b), where the complexity of the completed task can be taken into account when assessing work quality, may also benefit from it. Finally, a better understanding of task complexity dimensions may also inform better governance within enterprise crowdsourcing (i.e., tasks crowdsourcing within an enterprise to the crowd consisting of company employees) where different sociological issues exist (e.g., workload balance, different reward schemes, etc.) (Vukovic, Laredo, and Rajagopal 2010).

Original Contributions This paper makes a concrete step towards advancing our understanding of the complexity of crowdsourcing tasks and its quantification. We borrow from one of the most widely-used task complexity taxonomies (Campbell 1988), focusing on task complexity as a property of tasks perceived by workers during the interactions

between them, i.e., subjective task complexity. Specifically, we seek answers to the following research questions:

- **RQ1**: How is the complexity of crowdsourcing tasks perceived by workers, and distributed over tasks?
- **RQ2**: Which task features can characterize crowdsourcing task complexity, and based on them how to predict task complexity in an automatic way?
- **RQ3**: How do task complexity features affect task performance?

By re-instantiating 61 tasks published in Amazon MTurk, we collect subjective complexity evaluations from workers by using the NASA Task Load Index. We define a quantifiable measure of subjective complexity (i.e., Mean Complexity Score), which shows that task complexity is coherently perceived by different workers, and which we show to be significantly influenced by task type. To understand the influence of task design on complexity perception and to ultimately develop an automatic way for measuring task complexity, we propose a high-dimensional regression method, named `MFLR`, to model the key dimensions of complexity from three classes of *structural* features of a crowdsourcing task, including metadata, semantic content, and visual appearance. We show that both the visual appearance and language used in task description can largely influence the perception of task complexity, and thus can be used to measure it. Then, we demonstrate that the same feature set can be used to improve task throughput prediction. We show that complexity is a relevant task modeling property, thus contributing a better understanding of how task formulation can affect task execution performance in crowdsourcing markets.

Related Work

Task Complexity Complexity is a multifaceted property of human-executed tasks, for which it is difficult to establish a holistic definition. It is generally agreed that *task complexity* (Campbell 1988; Wood 1986) depends on both objective task properties and individual characteristics of the task doer. As such, task complexity can be operationalized in terms of: 1) intrinsic complexity, which does not account for subjective perception and individual differences of task doers, but focuses exclusively on the features of the task; 2) subjective complexity, which measures the task as a function of the perception and handling of its performers; and 3) relative complexity, which considers the relative relationship between the difficulty of the task and the capabilities of the task doers. Subjective and intrinsic task complexity are related, as the former can be used to explain how the latter is handled by workers. In a comprehensive review of literature in information systems and organizational behavior, (Campbell 1988) concluded that "any objective task characteristic that implies an increase in information load, information diversity, or rate of information change can be considered a contributor to complexity". In this paper we focus on *subjective* complexity of crowdsourcing tasks, aiming at its quantification and the identification of task features that contribute to it.

Measuring Complexity of Crowdsourcing Tasks Previous work investigated the relationship between task length and the trade-off between execution speed and result quality, showing how decomposition into simpler micro-tasks might lead to slower completion time but increased accuracy (Cheng et al. 2015b). This recent result indicates even further the need for understanding the impact of task complexity on the dynamics of a crowdsourcing marketplace and how this affects all actors involved. While task complexity has been used in other work on crowdsourcing – e.g., by means of a post-task questionnaire (Cheng et al. 2015b) or for the purpose of measuring worker effort and motivation (Rogstadius et al. 2011), our work is the first one looking at the different factors affecting perceived task complexity and to propose and evaluate methods to predict it given task properties.

(Cheng, Teevan, and Bernstein 2015a) propose a method to quantify the *effort* involved in task execution, meant to be used to estimate the appropriate monetary rewards for a given task. The approach is based on a pilot launch of the task, within which the time available for task completion is selectively manipulated. The approach is able to estimate the time needed to obtain completion while achieving a given output accuracy level. W.r.t. (Cheng, Teevan, and Bernstein 2015a), our work aims at measuring task complexity beyond completion time, and in a fully automated way, i.e., avoiding the need for a task to be "piloted" in order to estimate its required effort. We use a standard method to measure the ground truth perceived complexity of a task – that is the NASA Task Load Index (`NASA-TLX`) – and instrument a set of features (i.e., metadata, content, and visual, fully computable from the task data and html code) to predict task complexity.

`NASA-TLX` finds application in a variety of experimental tasks (Hart 2006), including the domains of information retrieval (Halvey and Villa 2014), Human Computer Interaction (Lischke et al. 2015). Recently, it has been also used in crowdsourcing, as a tool to assess subjective judgment of task difficulty (Mitra, Hutto, and Gilbert 2015), or to support the development of semi-automatic reward strategies (Cheng, Teevan, and Bernstein 2015a). The focus of our work is on reaching a better understanding of the different dimensions involved in measuring the subjective complexity of a crowdsourcing task. This will allow the creation of tools and quantitative measures to support worker and requester interaction, and it paves the way for novel research focusing on crowdsourcing of complex tasks (Kittur et al. 2013).

Micro-task Crowdsourcing Markets Micro-task crowdsourcing platforms are becoming more and more popular for both academic and commercial use. A variety of tasks is being crowdsourced over these platforms, with workers executing them in exchange of a small monetary reward. A taxonomy of popular micro-task type has been proposed by (Gadiraju, Kawase, and Dietze 2014), with audio transcription and survey being identified as two of the most popular task type (Difallah et al. 2015). In our work we use the same taxonomy to characterize tasks first and verify the impact of task type on complexity later.

Measuring and Modeling Task Complexity

To enable more efficient and fair mechanisms for microtask execution, we need a better understanding of which task features influence the success and effectiveness of online work. To this end, it is crucial to quantify the complexity of the tasks that humans (workers) have to carry out through (or in collaboration with) computer systems.

Measuring Complexity with NASA TLX

In literature, hardly any characterization of crowdsourcing task complexity has been explored. We introduce here methods for characterizing and quantifying subjective complexity towards answering **RQ1**.

Subjective complexity is defined in terms of workers' experience with a task (Campbell 1988). It relates to the notion of workload, since it is defined as the perception of the level of complexity associated with the performance of a task. Subjective complexity is related to intrinsic complexity, but they are not necessarily identical. Intuitively, tasks of a given intrinsic complexity might be more demanding for a worker than another. Also, tasks can require high effort (high subjective complexity) without necessarily being structurally more complex (arguably, a text summarization task is more demanding than an image annotation one), but other tasks might be demanding due to their structural complexity (e.g., a long survey with many radio buttons).

In the ergonomics literature, many instruments exist to measure workload, mostly based on self-assessment. Among them, the NASA Task Load Index (`NASA-TLX`) (Hart and Staveland 1988) is the most widely adopted. `NASA-TLX` is "a multi-dimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six sub-scales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort and Frustration". The test is straightforward: after completing the task to be evaluated, subjects are asked to rate workload along each of the subscales (typically, through 20-point likert scales ranging from 0 to 100, with endpoints anchored "low" and "high"). Then, subjects perform a pairwise comparison of the subscales, chosing among a pair the subscale which contributes the most to the overall workload. The pairwise comparison provides the relative weighting for each subscale (Hart and Staveland 1988)in the final overall `NASA-TLX` score, which is then determined as the weighted sum of each subscale score multiplied by the corresponding weight. In this work we adopt the resulting `TLX` score, ranging between 0 and 100, to measure subjective complexity of crowdsourcing tasks.

Modeling Complexity with Task Features

Our second research question (**RQ2**) concerns determining which *quantifiable* and *immutable* properties of a task can be used to estimate subjective complexity. The focus is on properties that relate to the task structure (e.g., instantiation properties, graphical layout, instructions) but not to its matter. The difference is subtle but significant. A task can be instantiated having the same structure (e.g. GUI and instructions), but different content. For example, the same task structure could be instantiated to ask workers to search for a bird in a picture. In this case, the specific picture proposed to the worker for the search would be the matter. Presenting the worker with a realistic picture of a bird rather than Miro's "The Migratory Bird" may significantly alter the complexity of the task. Investigating the influence of task matter on crowdsourcing task complexity poses issues of (1) data retrievability and (2) (multimedia) content interpretation, which is beyond the scope of this paper. In the following we explore three classes of task structural properties, and specifically: *metadata* features, *content features*, and *visual features*. We provide a brief introduction to each category, and refer the reader to the companion page for a full description of the feature set[1].

Metadata Features include task attributes defined at task creation time. They are used by a requester to provide an overview of the activities to be performed (i.e. `Title` and `Description` length), the required qualifications (e.g. worker `Location` and minimum `Approval Rate`), the estimated `Allotted Time` time and `reward`, and the number of `Initial Hits`. Our hypothesis is that metadata features reflect complexity as seen from the requester's point of view: intuitively, the task requester would reward higher compensation for more complex tasks, or provide a longer description to attract suitable workers.

Content Features aim at capturing the semantic richness of a task. Research in related domains – e.g. community question-answering systems (Yang et al. 2014), website complexity (Ivory, Sinha, and Hearst 2001) – suggests that content features could be indicative of the requirements and quality of task specification, which could in turn influence the perceived complexity of a task. Content features include numerical properties like the `Amount` of `Words`, `Links`, and `Images` in the task body, but also the `Keywords` used to concisely describe the task; the actual words (`unigrams`) contained in its title, description, and body; and the high-level `Topics`, extracted from task title and content using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). We hypothesize that the content of a task captures its requirements and hence its complexity. Also, the use of language features like unigrams can help highlight poorly formulated task instructions, which may increase workers' cognitive load at execution time.

Visual Features relate to visual properties of a task, including its layout and color palette. As the visual complexity of a Web page is known to influence the users' cognitive effort during interaction (Harper, Michailidou, and Stevens 2009), as well as users' aesthetics impression (Reinecke et al. 2013), we hypothesize that the look and feel of a crowdsourcing task can also have influence on its complexity. We study visual features that insist on (1) the general Web page structure (Ivory, Sinha, and Hearst 2001) such as `Body Text Percentage`, amount `stylesheet` and `script` files, and (2) the `visual areas` (Reinecke et al. 2013) such as number of `Text groups` and `Image areas`[2]. Intuitively, one could expect features like the amount of text groups to be positively correlated with

[1] https://sites.google.com/site/hcomp2016complexity/

[2] The code used to extract visual areas is available at http://iis.seas.harvard.edu/resources/aesthetics-chi13/

Task Type	Count	Percentage
Survey (SU)	4	6.6%
Content Creation (CC)	19	31.15%
Content Access (CA)	4	6.6%
Interpretation and Analysis (IA)	17	27.87%
Verification and Validation (VV)	2	3.3%
Information Finding (IF)	14	22.95%
Other	1	1.6%

Table 1: Distribution of task type in dataset.

task complexity, while some others such as the amount of `Emphasized Body Text Percentage` to be negatively correlated with task complexity – a higher emphasized body text percentage may suggest that the task designer has put effort in managing the user experience. We also study color-related features, previously investigated on images (Hasler and Suesstrunk 2003) and websites (Reinecke et al. 2013). `Colorfulness` is considered as one of the most notable features that influence emotional reaction (Coursaris, Swierenga, and Watrall 2008). We therefore hypothesize that it could also influence worker's mood during task execution, and thus influence their perception of task complexity. The color features we will study include the `histogram` of colors, and `HSV values`.

Is Complexity Coherently Perceived by Workers?

Our study started with an experiment where we collected task complexity evaluation from workers. This section introduces our experiment and the analysis of how task complexity is perceived by workers and distributed over task type.

Experiment

Setup To collect subjective complexity data, we first created a dataset of real MTurk tasks, to be then re-instantiated, executed, and evaluated by workers. As a first step, we extended *mturk-tracker*[3] to enable the retrieval of the complete set of resources associated with a task posted on MTurk, including meta-data, formatting (e.g. Javascript, CSS) and, when publicly available, content (e.g. images, audios). We performed the crawling during the first week of August 2015, retrieving 5487 tasks. We composed our final dataset of 61 tasks by picking from each requester active in the observation week one HIT per task type. The considered tasks have been labeled by the task type classifier developed in (Difallah et al. 2015) for the task type taxonomy proposed in (Gadiraju, Kawase, and Dietze 2014). This was necessary because HITs from the same requester are often similar to each other, and the majority of HITs are posted by a small number of requesters. Selecting one HIT per task type and requester guaranteed diversity in the task set in terms of both HIT content and visual presentation, as well as in task type; an overview of the composition of the dataset is given in Table 1. We then instantiated the 61 tasks based on the crawled task data. To minimize the chance of learning bias, we use another platform, i.e. Crowdflower, to collect the complexity evaluation for these tasks. We turned to the CrowdFlower crowd also to include in the evaluation the judgment of workers from a broader set of countries, thus reducing the risk of country-specific biases. Then, we appended at the end of each task the `NASA-TLX` questionnaire of task complexity assessment, and we asked workers to fill it in with respect to the task they just executed. Workers were recruited among Crowdflower Level 3 contributors, and tasks and `TLX` completion were performed in an externally hosted server.

To obtain reliable complexity assessment, we aimed at having 15 workers executing each task and completing the related `TLX`. In addition, to control for the quality of the complexity evaluations, we implemented a post-hoc task execution filtering mechanism. Due to lack of supervision during task execution in crowdsourcing, misunderstandings of the task instructions, as well as low engagement or commitment to the task, unreliable task outputs are to be expected (Hoßfeld et al. 2014; Eickhoff and de Vries 2013). As `TLX` entails subjective assessment of complexity, it was not possible to adopt a quality control based on golden answers (Alonso, Marshall, and Najork 2014). Therefore, we implemented a mechanism of control for self-consistency. When presenting the workers with the 15 pairs of subscales at the end of `TLX` questionnaire, we repeated three pairs twice (non consecutively, so the workers would be less aware of the repetition). We would expect unreliable workers to skip through these pairs, randomly selecting an answer to finish the task fast, and possibly giving inconsistent answers on the repeated pairs.

Results The 61 crowdsourcing tasks were executed and evaluated by 13 to 16 workers each ($M = 14.8 \pm 0.572$). In total, we obtained 903 evaluations, including: (1) a set of 6 judgments of complexity, each expressed along a different subscale, representing a different *complexity factor* (Mental, Physical, Temporal, Performance, Effort and Frustration); (2) a set of six *weights*, each representing the relevance of each factor in computing the final task complexity and (3) an overall *task complexity score*. For each completed task, we also recorded both the time taken by each worker to complete it and the time taken to fill in the follow-up `TLX` questionnaire.

The filtering of task executions and evaluations was enacted when: 1) more than two mistakes were made in the control questions (i.e., repeated pairs of complexity factors in the questionnaire); and 2) the time taken to complete the `TLX` questionnaire was outside the range $MEAN\_CT(i) \pm 2 * std\_CT(i)$, where $MEAN\_CT(i)$ is the average across workers of the time taken to fill in the questionnaire for task i and $std\_CT(i)$ is the related standard deviation.[4] For 34 of the 903 `TLX` completions (3.8%), workers made mistakes

[3] http://www.mturk-tracker.com/

[4] $MEAN\_CT$ was computed per each task separately, rather than across tasks. This is due to the presence of a significant effect of the task i on the `TLX` completion time (Kruskal-Wallis Test, $H = 142.89, p < 0.001$). This suggests that the characteristics of the task may have influenced the engagement of the worker with the task execution (and thereby questionnaire completion); workers poorly engaged with the task may have completed the `TLX` skip-

in all control questions; in 107 cases (11.8%), 2 out of the three control questions were answered in a wrong way. As a result, 15.6% of the evaluations were discarded. Of the remaining 762, 52 took either too short or too long to be completed (5.8%); hence, the following analysis is based on the remaining 710 evaluations (11.64 ± 1.693 per task).

Perception and Distribution of Task Complexity

Inter-evaluator Agreement The purpose of the subjective study was to establish a *Mean Complexity Score* (MCS) for each task, expressing the complexity of a given task as perceived, on average, by a crowd of workers executing it. This was functional to our follow-up experiment of complexity prediction. To establish MCS as ground truth, we verify the presence of a sufficient agreement among workers (evaluators) in expressing complexity scores: a too large disagreement would give large confidence intervals of the MCS, making them unreliable as ground truth.

Traditional inter-evaluator agreement measures are not applicable in our case, as they often assume repeated measures (e.g., Cronbach's alpha). Krippendorff's alpha does not have this limitation, but it has been shown to be of limited reliability for subjective evaluation tasks (Alonso, Marshall, and Najork 2014). A possible different way to look at the problem is to check the extent to which individual evaluations are spread around the mean of the complexity (or complexity factor) value per task. This type of analysis, often used in visual quality assessment tasks, can be applied to any type of subjective evaluation involving a pool of participants scoring the same item (in our case, a task i), for which a mean opinion score along some dimensions (complexity and its factors, in our case) is sought. The SOS hypothesis (Hoßfeld, Schatz, and Egger 2011) is a useful tool to perform such type of analysis. It stems from the observation that Mean Opinion Scores (MOS; in our case, mean complexity - or complexity factor - scores for the tasks) and the spread of the individual scores around the MOS (as measured by their Standard Deviation, or SOS) are linked by a squared relationship. Specifically, if the dependent variable (i.e., complexity) is measured on a K-point scale, with v_1 being the lowest value of the scale and v_K being the highest, for each task i we can define the relationship $SOS(i)^2 = \alpha(-MOS(i)^2 + (v_1 + v_K)MOS(i) - (v_1 * v_K))$, with $MOS(i)$ the Mean Opinion Score for task i and $SOS(i)$ the related standard deviation. The parameter α that regulates this relationship can be found by fitting the MOS to the SOS data. Its value can be compared with that of other subjective evaluation tasks, which are deemed to be more (high α) or less (low α) prone to high variability in evaluations.

Table 2 shows α values computed for complexity evaluations as well as evaluations of the individual complexity factors (mental, physical, temporal, performance, effort and frustration). α values range between 0.25 for the effort factors and performance and 0.29 for the frustration factors. Complexity evaluations have an α value of 0.28. This value is similar to what could be obtained in other subjective evaluations tasks (e.g. smoothness of web surfing, VoIP quality, cloud gaming quality, or image aesthetic appeal (Hoßfeld, Schatz, and Egger 2011; Redi and Povoa 2014), and we consider it acceptable. We thus conclude that subjective task complexity is coherently perceived by workers. The scores per task could be therefore aggregated into Mean Complexity Scores (MCS), and Mean Factor Scores (MFS) for each individual complexity factor. Figure 1 shows the MCS and MFS per task plotted against their related SOS, along with the curve fitting the data (p-value < 0.001 for each complexity factor) according to the SOS relationship described before.

Subjective Task Complexity Scores We now analyse the output of the NASA−TLX questionnaire. We are interested in observing whether some complexity factors would get higher scores than others, and how users deemed each factor relevant in the final complexity score. In this respect, we computed, in addition to MCS and MFS, also Mean Factor Weights (MFW), as the average weight value given by individual evaluators (i.e., workers) to the same task.

Kolmogorov-Smirnoff and Levene tests revealed MCS, MFS and MFW to be normally and homogeneously distributed, thus allowing the application of parametric analysis. The 61 tasks obtained an average complexity value of 63.78 ± 11.56. Figure 2 shows the factor values scored by the 61 tasks on average. A one-way ANOVA setup with complexity factor as independent value and mean score per task as dependent variable, revealed that tasks were scored significantly different along the six factors ($F = 22.414, df = 6, p < 0.001$). Specifically, tasks scored lower (according to post-hoc tests with Bonferroni correction) in physical complexity, and highest in mental complexity and effort. Considering that online crowdsourcing tasks usually involve very little physical labour, and more mental effort, these results are in line with the expectations. Perceived performance was also scored significantly lower than mental complexity and effort ($p < 0.001$ in both cases); this indicates that workers were, on average, relatively happy with their performance with the task (the lower the score on the performance sub-scale, the better the perceived performance). Finally, frustration scored significantly lower than all other factors (except for physical complexity), which suggests workers to be rather satisfied with the tasks we instantiated.

To understand the relative importance of each factor in the overall complexity perception, we also looked at the distribution of the MFW across tasks. An ANOVA with Bonferroni correction ($F = 170.821, df = 5, p < 0.001$), established that all factors had significantly different weights for our tasks (see Figure 3), except for effort and mental (p = 0.767). The latter two obtained, across tasks, relatively high weights indicating that mental complexity and effort to complete the task are very relevant in judging the complexity of crowdsourcing tasks. Interestingly, the highest weight was assigned to performance. Whereas mental complexity was expected to have high relevance in the overall complexity, the fact that workers seem to care about the degree to which they perform the task properly, suggests that intrinsic motivation (as in willingness to perform the assigned job properly) may play more than a marginal role in workers sat-

ping through it (too fast) or doing other things while at it (too slow), either ways producing poorly reliable task evaluations.

Factor	Complexity	Mental	Physical	Temporal	Performance	Effort	Frustation
α	0.2785	0.2627	0.2745	0.2897	0.2507	0.2503	0.2937

Table 2: SOS Hypothesis α values for Mean Complexity Scores and Mean Factor Scores.

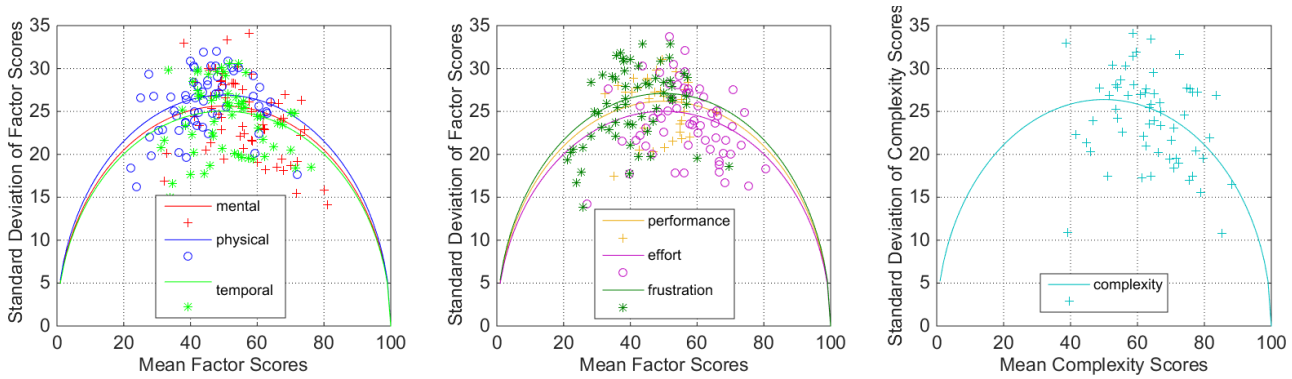

Figure 1: SOS Hypothesis plots for complexity (rightmost) and its factors (left and center plots). The continuous lines depict the square fitting found applying the SOS hypothesis.

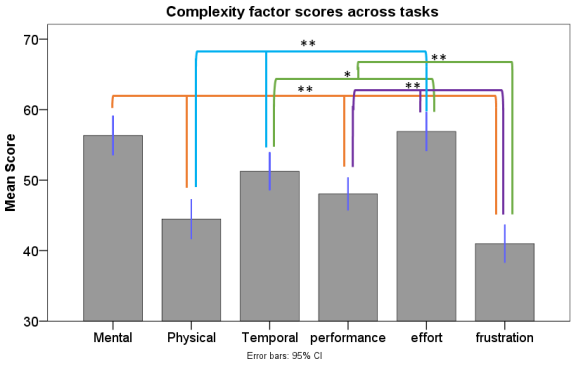

Figure 2: Mean Scores per complexity factors across the 61 instantiated tasks. "*" indicates mean differences with significance to the .05 level, "**" indicates mean differences with significance to the .01 level.

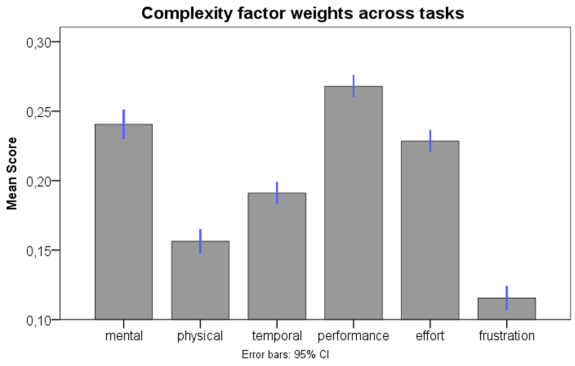

Figure 3: Mean values of the weights assigned to the six complexity factors throughout the 61 instantiated tasks. All means are significantly different except those of mental complexity and effort.

isfaction and overall motivation to complete complex crowdsourcing tasks. Frustration, which obtained on average the lowest weight, seems instead not impact majorly the overall perceived complexity of the task.

Finally, we verified whether task type (see Table 1) had an influence on perceived complexity, as well as task performance time, and TLX completion time. As complexity scores were normally and homogeneously distributed when clustered according to task, we used again parametric testing. Note that we excluded from the analysis the task of type "other" as we only had one datapoint for that category. Task type was found to have a significant effect on complexity ($F = 3.729$, $df = 5$, $p = 0.06$). This effect could be entirely explained by the significant difference in complexity between *Content Creation* and *Interpretation and Analysis* tasks, the latter found to be less complex than the former by an average 11.61 points ($p = 0.017$ after Bonferroni correction). Task performance time and TLX completion time were found to be non-normally distributed. A Kruskal Wallis test revealed that for both dependent variables, task type did not have a significant effect (for task execution time: $H = 9.067$, $p = 0.170$; for TLX completion time: $H = 2.159$, p = 0.866).

Task Complexity Prediction

Having collected subjective complexity scores for our 61-task dataset, we then checked whether our proposed features were useful to predict it.

Feature sets. As input for our model, we experiment separately with each of the feature categories described before (metadata, content, and visual features). We also test a fourth prediction configuration, in which a LDA is applied to content features to extract semantic topics, thereby reducing the dimensionality of the feature set for model training and testing. The dimensionality of metadata, visual and semantic features was 9, 14, and 1447, respectively.

Feature Set	Regression Model			
	Linear	Lasso	MFLR	Random Forest
Metadata	13.37±4.18	13.16±4.24	–	**9.94±1.68**
Visual	14.86±4.01	12.50±2.07	**9.97±1.28**	10.21±1.15
Content	12.87±1.64	9.97±1.27	**9.18±1.83**	10.00±1.47
Content LDA	10.34±1.84	**9.23±1.44**	–	11.80±1.18

Table 3: Subjective complexity estimation, measured by mean absolute error (MAE). The best predicting model for each feature class is highlighted in bold; the best performance across all features with different regression models is underlined.

Regression models. Due to the high dimensionality of content features, we propose a novel regression model, i.e. matrix factorization boosted linear regression (MFLR), that performs at the same time 1) dimension reduction, via non-negative matrix factorization (Lee and Seung 1999), and 2) regression, via Lasso (Tibshirani 1996). The key idea of MFLR is to learn latent complexity dimensions (LCD's) that are predictive for task complexity. To do so, the model jointly optimizes matrix factorization and regression functions as follows

$$\min_{\mathbf{U},\mathbf{V}\geq\mathbf{0},\mathbf{W}} \|\mathbf{X}-\mathbf{U}\mathbf{V}^T\|_F^2+\lambda\|\mathbf{U}\mathbf{W}^T-\mathbf{Y}\|_F^2 + \lambda_X(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)+\lambda_Y\|\mathbf{W}\|_1 \quad (1)$$

where $\|\mathbf{X}-\mathbf{U}\mathbf{V}^T\|_F^2$ factorizes the task-feature matrix $\mathbf{X}$ to two smaller matrices, i.e. $\mathbf{U}$ as the task-LCD matrix and $\mathbf{V}$ as the feature-LCD matrix; $\|\mathbf{U}\mathbf{W}^T-\mathbf{Y}\|_F^2$ takes task LCD's ($\mathbf{U}$) for predicting the complexity ($\mathbf{Y}$). The regularization terms $\mathcal{L}_X(\mathbf{U},\mathbf{V})$ and $\mathcal{L}_Y(\mathbf{W})$ are implemented with Frobenius and L_1 norm, and regulated by λ_X and λ_Y respectively [5].

In MFLR, feature importance can be computed as $F_{importance} = \mathbf{V}\mathbf{W}^T$, where feature f_i has contribution $\mathbf{V}[i]\mathbf{W}^T$, with $\mathbf{V}[i]$ denoting the distribution of feature f_i over the learned latent complexity dimensions.

To assess the performance of MFLR, we experiment with 3 methods: 1) Linear regression, used as the baseline model; 2) Lasso, which is a linear model with feature selection; and 3) Random forest, which is an ensemble model known for its good generalization capabilities.

Parameter setting. For Lasso, we use the default setting in sklearn[6], in which the regularization parameter is set to 1, which implies that the error function and the regularization term have the same weight. To make the results comparable, we also set $\lambda = \lambda_Y$ in MFLR. Using a grid search on the training data, we set the parameters of MFLR as follows: $\lambda_X = 1, \lambda = \lambda_Y = 0.01$. To account for the Content features dimensionality variance, we set the number of topics to be extracted by LDA to 10, so as the number of latent complexity dimensions of MFLR.

Results Table 3 summarizes the resulting prediction performance. We ran a 5-fold cross-validation and report the averaged performance measured by mean absolute error (MAE). The ground truth task complexity is 63.78 ± 11.46 based on crowd worker annotations. The MAEs of different configurations fall into the range (9, 14), an acceptable error considering the ground truth figures. Overall, content features yield the best prediction performance with MFLR, although LDA-reduced content features with Lasso follow closely. Notably, prediction with visual features is also significantly improved with MFLR, a result that shows the good performance achievable by MFLR also with a lower number of dimensions. We account these results to the fact that MFLR re-combines features into latent complexity dimensions, projecting the input data into a (low-dimensional) transformed space better characterizing task complexity. This has obvious benefits with respect to Lasso, which uses the original feature space instead; with respect to LDA, the advantage of MFLR is in the fact that the latent complexity factors derivation is guided by the regression, thereby increasing the predictive power of LCDs'.

Visual Feature	Imp.	Semantic Feature	Imp.
visualAreaCount	3.25	linkCount	2.42
hueAvg	0.09	wordCount	1.37
		keyword: audio	0.09
		keyword: transcribe	0.07
		keyword: writing	0.06
imageAreaCount	-0.27	unigram: clear	-0.06
colorfulness1	-0.63	unigram: identify	-0.07
scriptCount	-1.52	uigram: date	-0.09
valAvg	-1.71	keyword: easy	-0.10
cssCount	-1.82	imageCount	-1.01

Table 4: Features more correlated (positively and negatively) with subjective complexity.

Important features. Table 4 summarizes the visual and content features most positively and negatively correlated with complexity. The most positively correlated visual feature is the number of visual areas, indicating that more visual items lead to higher task complexity perceived by workers. The presence of curated layouts and interactions (CSS and Javascript) are negatively correlated features with complexity, suggesting that a better design of the task presentation and more interactive components could decrease the complexity perceived by workers. Notably, the stronger visual factor is valAvg, i.e. the "value" (or "lightness") of the task design. In content features, complexity is reflected from the point of view of required actions to be performed by workers (e.g., *transcribe*), task type (e.g., *writing*), and content matter (e.g. *audio*). The amount of task elements (text, images, links) also plays a relevant role.

[5] We refer the reader to our companion page for the detailed optimization algorithm.

[6] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

Can Task Complexity Features Help Improve Task Performance Prediction?

As structural features of tasks have an influence on subjective complexity, we are interested in understanding if they are also helpful in predicting task performance. This section reports our efforts in applying the task complexity feature set on task performance prediction, where we approximate task performance by execution throughput. Our interest is in 1) evaluating the utility of using the complexity feature set for throughput prediction, 2) understanding the features that yield best prediction performance, and compare them with the ones best for complexity prediction.

Experimental Setup

Data Preparation We operate on a five year dataset (Difallah et al. 2015), collected from Amazon MTurk from 2009 to 2014. The dataset contains more than $2.56M$ distinct batches, and $130M$ HITs. All batches in the dataset are described by metadata, but only 55% of them are provided with task content. No task in the dataset contains all the information (e.g. CSS stylesheets, external resources) required to evaluate visual features. The throughput prediction problem has been previously studied in (Difallah et al. 2015). The goal is to predict the number of tasks executed in a predefined time span. We apply the same experimental methodology as in (Difallah et al. 2015). We randomly select 50 time points, uniformly sampled from 5 months (June to October 2014) of MTurk market data. For each time point, we predict the #tasks executed in the following hour, using as training data information about tasks published and executed in the previous four hours. To avoid biases due to the over-representation in the marketplace of a specific requester, we sample the task space by selecting, for each requester, a single batch per task type. This choice allows to keep an accurate distribution of task type, while avoiding compensation for the inevitable skewness introduced by big market players (Difallah et al. 2015). The result is an experimental dataset consisting of 8675 tasks. To better characterize prediction performance on batches with different throughput magnitude, we divide the dataset into three subgroups, according to the power-law distribution, with batch size varying in the range of $[1, 10)$, $[10, 100)$ and $[100, 1000)$, respectively. For evaluation, we again use the Mean Absolute Error (MAE) as a metric to measure prediction performance.

Feature Sets As input for throughput prediction, we experiment separately with each metadata and content features. Visual features could not be tested due to their absence in the majority of tasks. We also include task and market *dynamic* features, which are related to the execution context of the task in the marketplace. Inspired by (Difallah et al. 2015), we consider properties of the marketplace in the observation interval, including: the total number of `Available HITs` in the market (to account for market size); the total number and relative size of `Published` and `Completed` HITs (to account for market speed); and the total amount of `Available` and `Obtained Rewards` (to account for market value). The resulting feature dimensionality is 9 and 11, respectively; content features dimensionality varies according the number of unigrams in the subgroup, leading to $26.4k$ ($[1, 10)$), $15.2k$ ($[10, 100)$) and $1.3k$ ($[100, 1000)$) content features, respectively.

Models and Parameter Setting We experimented again with `Linear` regression, `Lasso` and `Random Forest` regression. In this case of task performance predict on, `MFLR` finds latent performance dimensions (`LPD`) drawn from the considered feature classes. We also applied `LDA` on content features for topic modeling. We set both the number of topics for `LDA` and the number of latent performance dimensions of `MFLR` to 100.

Results

Table 5 summarizes the prediction performance obtained by applying different feature classes to the considered regression models. The comparison of prediction performance on different feature classes highlights how dynamic features (i.e., the marketplace context) provide the least prediction support; this is an interesting result, that hints to the greater importance of objective task features for throughput prediction. We observe that metadata feature achieves good performance with `Random Forest` and content features achieves good performance with `Lasso`. This is not surprising, given that content features are of high dimensionality and `Lasso` properly select predictive features for task performance estimation. Interestingly, high-level semantic features, i.e. topics extract via LDA, achieves better performance than plain content feature with `Lasso`, which shows that dimension reduction improves the predictive power of content features. Most importantly, content features achieves the best performance with `MFLR` compared to all configurations when applied to (in batch groups $[1-10)$ and $[10-100)$), which is consistent with the complexity prediction experiment. It is worth noting that the configurations of $[1-10)$ and $[10-100)$ are the ones where the other models make the biggest errors, relatively to the range within throughput varies.

Content features are more informative than metadata features in low throughput groups ($[1, 10)$ and $[10, 100)$), with significant performance differences in both cases ($p < 0.01$, Wilcoxon signed-rank test); however, metadata features work better in the group $[100, 1000)$, although with non significant differences in performance ($p = 0.36$, Wilcoxon signed-rank test). These results complement previous work in throughput prediction (Difallah et al. 2015), where high prediction performance is mainly guided by the `Initial Hits` metadata feature. We show how the importance of such feature significantly decreases in low throughput groups, where content features have a prominent role. These results might indicate that workers execute large batches mainly due to the high volume of HITs available to be executed, while for batches of smaller sizes the content of the task have a bigger influence on workers' decision for task execution. Therefore requesters are recommended to better design their tasks of small batch sizes.

Important Features. Table 6 (line 2-6) shows the latent performance dimensions of content features that contribute the most in prediction: the most important $\texttt{LPD}_1$ contains terms related to the type of actions required from the worker

Feature Set	[1,10) (GT: 3.01±2.35)				[10,100) (GT: 29.92±21.63)				[100,1000) (GT: 265.45±185.97)			
	Linear	Lasso	MFLR	RForest	Linear	Lasso	MFLR	RForest	Linear	Lasso	MFLR	RForest
Metadata	3.88	3.78	–	**3.65**	18.35	18.11	–	**17.45**	126.81	126.45	–	**107.60**
Dynamic	**3.73**	3.61	–	3.93	20.31	**18.95**	–	20.47	**126.76**	131.30	–	138.55
Content	518.29	3.42	**2.72**	4.50	576.44	16.86	**15.65**	20.53	265.45	**110.14**	112.48	116.33
Content LDA	4.11	**3.42**	–	4.55	18.92	**17.88**	–	19.75	123.11	**118.15**	–	128.43

Table 5: Throughput prediction performance of different feature classes and regression models, measured by MAE. Results are reported for three batch groups: throughput within the range of $[1, 10)$, $[10, 100)$, and $[100, 1000)$. **GT** is the ground truth throughput. In every group, the best performance among different regression models for each class of features is highlighted in bold; the best performance among all feature classes for every group is underlined.

Impo. LPD's and Their main unigrams	$\mathbf{LPD}_1$	$\mathbf{LPD}_2$	$\mathbf{LPD}_3$	$\mathbf{LPD}_4$
	text	find	search	expressions
	clip	company	relevance	faces
	copy	online	google	emotions
	easy	fast	internet	mimicking
	quick	entry	information	camera
Impo. unigrams	bonus	copy	easy	categorization
	image	search	instances	indentification

Table 6: Main LPD's and unigrams that contribute the most to throughput prediction.

	- TP	**+ TP**
- CPX	know, words, exact guidelines, shown free, ask, based change, load, tell notes, number, need	photo, picture, text item, identify, type thank, best, easy address, job, pay accept, right
+ CPX	questions, answer file, provide create, options, listed save, issues, result personal, send, unable	transcribe, audio, images review, feedback, search read, article, try code, writing, sentence carefully, follow, instructions

Table 7: Top-15 unigrams associated with each positive **(+)** and negative **(-)** correlation with complexity (**CPX**) and throughput (**TP**) prediction.

(arguably, *copying* and *pasting* of *text* are a simple actions to perform); the second and third most important LPD's reflect performance from the point of view of task type (information finding – *find*, *online*, *google* – is known to be among the most demanding type of crowdsourcing tasks, in terms of worker labour). Finally, the fourth LPD hints to the type of annotation to be performed (subjective, looking for expressions and emotions), and to the type of content (images and videos). Table 6 (line 7-8) summarizes the most important unigrams across factors. Terms related to actions, task type, and task matter emerge as most representative.

Features for Complexity vs. Performance. Content features can be used to predict both task complexity and task performance. Are there content features able to predict both? We tested the presence of correlation between the *importance* of features in complexity prediction and in performance prediction. Results show a weak negative correlation of feature importance. Significance (Pearson correlation: -0.1308; $p < 0.05$) is present only for low throughput (**[1,10)**) tasks. This result supports the findings from the previous section, where low throughput tasks were the ones more likely to benefit for content features for prediction purposes. The negative correlation indicates the presence of features that, while hinting to higher task complexity, can signal low completion performance, and viceversa. We show these features in Table 7. The weak – yet significant – correlation is due to the existence of features having consistent correlation, i.e. both associated positively (respectively, negatively) with complexity and performance. Unigrams describing task actions and task type are again the ones more likely to be associated with consistent complexity and performance prediction. For instance *transcribe audio* unigrams are predictive of high complexity[7] and higher throughput in the low throughput (**[1,10)**) batches. This is an unexpected result, that we can explain only by looking at the MTurk market as a whole (Difallah et al. 2015), where the presence of large batches devoted to audio transcription might influence workers' task selection strategy. Clearly, further investigation focusing on the relationship between task complexity, market dynamics, and execution performance is needed, and it will be tackled in future work.

Conclusions

This work studied the subjective complexity (i.e., as perceived by workers) of crowdsourcing tasks. We defined (1) an operational way to quantify subjective complexity; (2) a set of objective features (i.e., quantities computable from the task metadata and HTML code) from which to predict complexity automatically; and (3) a novel regression model, MFLR, that shows superior performance for complexity prediction. We were able to deliver a model predicting subjective task complexity in an accurate and fully automatic way; features describing the semantic content of the task are most predictive for complexity, followed closely by the features describing the visual appearance of the task. We show how this feature set is also useful to improve the prediction of task throughput when workers' task selection strategy is not influenced by batch size.

Our automatic complexity predictors have the potential to impact crowdsourcing research widely. We expect them to be useful in deploying new strategies for workers retention by, e.g., adjusting task complexity over a batch of tasks. We

[7] Arguably, audio transcription requires a lot of attention for proper execution.

also expect that measuring task complexity will create a better communication channel between requesters and workers thanks to the shared understanding of the required effort to complete the task, as well as the implementation of more fair compensation mechanisms. Our future work will focus on understanding the effect of varying task complexity on work quality, efficiency, and retention; efforts will also be devoted in extending the breadth and scope of our subjective complexity data, to yield more generalizable results.

Acknowledgement

This work was partially supported by the UK EPSRC grant number EP/N011589/1. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

References

Alonso, O.; Marshall, C.; and Najork, M. 2014. Crowdsourcing a subjective labeling task: a human-centered framework to ensure reliable results. Technical report, MSR-TR-2014-91.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of Machine Learning Research* 3:993–1022.

Campbell, D. J. 1988. Task complexity: A review and analysis. *Academy of Management Review* 13(1):40–52.

Cheng, J.; Teevan, J.; Iqbal, S. T.; and Bernstein, M. S. 2015b. Break it down: A comparison of macro- and microtasks. In *CHI*, 4061–4064. ACM.

Cheng, J.; Teevan, J.; and Bernstein, M. S. 2015a. Measuring crowdsourcing effort with error-time curves. In *CHI*, 1365–1374. ACM.

Coursaris, C. K.; Swierenga, S. J.; and Watrall, E. 2008. An empirical investigation of color temperature and gender effects on web aesthetics. *Journal of Usability Studies* 3(3):103–117.

Difallah, D. E.; Catasta, M.; Demartini, G.; Ipeirotis, P. G.; and Cudré-Mauroux, P. 2015. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *WWW*, 238–247. ACM.

Eickhoff, C., and de Vries, A. P. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16(2):121–137.

Gadiraju, U.; Kawase, R.; and Dietze, S. 2014. A taxonomy of microtasks on the web. In *HyperText*, 218–223. ACM.

Halvey, M., and Villa, R. 2014. Evaluating the effort involved in relevance assessments for images. In *SIGIR*, 887–890. ACM.

Harper, S.; Michailidou, E.; and Stevens, R. 2009. Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Transactions on Applied Perception (TAP)* 6(2):10.

Hart, S. G., and Staveland, L. E. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology* 52:139–183.

Hart, S. G. 2006. Nasa-task load index (nasa-tlx); 20 years later. In *HFES*, volume 50, 904–908. Sage Publications.

Hasler, D., and Suesstrunk, S. E. 2003. Measuring colorfulness in natural images. In *Electronic Imaging*, 87–95. International Society for Optics and Photonics.

Hoßfeld, T.; Keimel, C.; Hirth, M.; Gardlo, B.; Habigt, J.; Diepold, K.; and Tran-Gia, P. 2014. Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *IEEE Transactions on Multimedia* 16(2):541–558.

Hoßfeld, T.; Schatz, R.; and Egger, S. 2011. Sos: The mos is not enough! In *QoMEX*, 131–136. IEEE.

Ivory, M. Y.; Sinha, R. R.; and Hearst, M. A. 2001. Empirically validated web page design metrics. In *CHI*, 53–60. ACM.

Jung, H. J., and Lease, M. 2015a. A discriminative approach to predicting assessor accuracy. In *ECIR*, 159–171. Springer.

Jung, H. J., and Lease, M. 2015b. Modeling temporal crowd work quality with limited supervision. *HCOMP*.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In *CSCW*, 1301–1318. ACM.

Law, E., and von Ahn, L. 2011. *Human Computation*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Lischke, L.; Mayer, S.; Wolf, K.; Henze, N.; Schmidt, A.; Leifert, S.; and Reiterer, H. 2015. Using space: Effect of display size on users' search performance. In *CHI*, 1845–1850. ACM.

Mitra, T.; Hutto, C.; and Gilbert, E. 2015. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *CHI*, 1345–1354. ACM.

Redi, J., and Povoa, I. 2014. Crowdsourcing for rating image aesthetic appeal: Better a paid or a volunteer crowd? In *CrowdMM*, 25–30. ACM.

Reinecke, K.; Yeh, T.; Miratrix, L.; Mardiko, R.; Zhao, Y.; Liu, J.; and Gajos, K. Z. 2013. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *CHI*, 2049–2058. ACM.

Rogstadius, J.; Kostakos, V.; Kittur, A.; Smus, B.; Laredo, J.; and Vukovic, M. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *ICWSM*.

Sheshadri, A., and Lease, M. 2013. Square: A benchmark for research on computing crowd consensus. In *HCOMP*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Venanzi, M.; Guiver, J.; Kazai, G.; Kohli, P.; and Shokouhi, M. 2014. Community-based bayesian aggregation models for crowdsourcing. In *WWW*, 155–164. ACM.

Vukovic, M.; Laredo, J.; and Rajagopal, S. 2010. Challenges and experiences in deploying enterprise crowdsourcing service. In *ICWE*, 460–467. Springer.

Wood, R. E. 1986. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* 37(1):60 – 82.

Yang, J., and Bozzon, A. 2016. On the improvement of quality and reliability of trust cues in micro-task crowdsourcing. In *Weaving Relations of Trust in Crowd Work: Transparency and Reputation across Platforms. Workshop at WebScience 2016*.

Yang, J.; Hauff, C.; Bozzon, A.; and Houben, G.-J. 2014. Asking the right question in collaborative q&a systems. In *HyperText*, 179–189. ACM.