SimilarHITs: Revealing the Role of Task Similarity in Microtask Crowdsourcing

Alan Aipe
IIT Patna
Patna, Bihar, India
alan.me14@iitp.ac.in

Ujwal Gadiraju
L3S Research Center,
Leibniz Universität Hannover
Hannover, Germany
gadiraju@L3S.de

ABSTRACT

Workers in microtask crowdsourcing systems typically consume different types of tasks. Task consumption is driven by the self-selection of workers in the most popular platforms such as Amazon Mechanical Turk and CrowdFlower. Workers typically complete tasks one after another in a chain. Prior works have revealed the impact of ordering tasks while considering aspects such as task complexity. However, little is understood about the benefits of considering task similarity in microtask chains.

In this paper, we investigate the role of task similarity in microtask crowdsourcing and how it affects market dynamics. We identified different dimensions that affect the perception of task similarity among workers, and propose a supervised machine learning model to predict the overall task similarity of a task pair. Leveraging task similarity, we studied the effects of similarity on worker retention, satisfaction, boredom and fatigue. We reveal the impact of chaining tasks according to their similarity on worker accuracy and their task completion time. Our findings enrich the current understanding of crowd work and bear important implications on structuring workflow.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; • **Human-centered computing**; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Crowdsourcing; Microtasks; Task Similarity; Workers; Performance

ACM Reference Format:
Alan Aipe and Ujwal Gadiraju. 2018. SimilarHITs: Revealing the Role of Task Similarity in Microtask Crowdsourcing. In *HT '18: 29th ACM Conference on Hypertext and Social Media, July 9–12, 2018, Baltimore, MD, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

HT '18, July 9–12, 2018, Baltimore, MD, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Crowdsourcing has evolved rapidly over the last decade and presently accounts for a multi-million dollar marketplace. There is a wide variety of Human Intelligence Tasks (HITs) which are typically crowdsourced. Thousands of diverse workers with different motivations complete HITs on crowdsourcing platforms based on availability, their eligibility to participate, and a number of other variables that drive the market dynamics [9]. In such a diverse environment, the role of parameters affecting efficiency and accuracy of workers gain significance and need to be studied in order to improve the overall effectiveness of the paradigm [15]. Task consumption in microtask crowdsourcing platforms is largely driven by a self-selection process, where workers meeting the required eligibility criteria select the tasks that they prefer to work on. Moreover, prior works have established that, during the course of completion of a variety of tasks, workers tend to perform better with experience due to learning effects [8, 11]. Thus, understanding the factors that drive dynamics of learning and devising intelligent task chaining techniques can augment the quality of contribution of workers in microtask crowdsourcing platforms.

In this paper, we focus on understanding the similarity between HITs and the role of task similarity in shaping market dynamics and outcomes. The study commences by identifying different similarity dimensions and their influence on overall similarity of a task pair. Using the above dimensions as key ingredients, we propose a supervised machine learning model to predict similarity of any given task pair. After this, we move forward and investigate the effects of task similarity on significant parameters like worker satisfaction, boredom, fatigue, and task completion time.

Definition : The *similarity* of a task pair can be defined as the degree of resemblance between two tasks, i.e., the extent to which they are identical in nature.

Research Questions and Original Contributions. This paper aims at filling this knowledge gap by contributing novel insights on the nature and importance of task similarity in microtask crowdsourcing. By combining qualitative and quantitative analysis, we seek to answer the following research questions.

> **RQ1**: What are the different dimensions of similarity between two HITs? How much influence does each dimension have on the perceived similarity of a HIT pair?

Through a study of 100 workers on CrowdFlower (Study I), we identified key dimensions of similarity between microtasks and

their corresponding influence in the overall task similarity perception of workers.

RQ2: How can we model the overall similarity between two HITs?

We used a supervised machine learning model and proposed features to predict the workflow and topic similarity of any given task pair. Building on our findings from Study I, we propose a weighted sum of task similarity over individual dimensions as the overall task similarity of a pair of tasks from a holistic standpoint. Our proposed stochastic gradient descent regressor model results in predicting task similarity accurately, with a mean average error of 0.68.

RQ3: What is the effect of task similarity on worker retention, worker satisfaction, boredom and fatigue?

Our studies reveal that ordering tasks in a chain according to overall task similarity results in improved accuracy, but at the cost of inducing boredom.

2 RELATED LITERATURE

2.1 Task Chaining and Complexity

Crowdsourcing microtasks are small units of work designed to be completed one bit at a time, eventually contributing to a much larger goal. Although they can be performed in isolation, in practice people often complete them one after another, in a chain [4]. Prior research has shown that task ordering or chaining has a profound impact on worker performance [4, 21]. This is because transitions between consecutive cognitive tasks have measurable effects on ongoing mental processes. Related literature in psychology suggests that the contribution ability of workers tends to wane and become more error-prone while switching between tasks, when compared to a single task scenario. This can be accounted by re-configuration of physical and psychological parameters to adapt to new task at hand [19, 25].

Task complexity has undoubtedly turned out to be an integral parameter driving the performance of workers on crowdsourcing platforms. Research work in the field of task chaining with respect to complexity revealed that lead-up microtasks have a significant impact on momentum and performance of workers performing a final task. Authors also proved that while the same operation chains aid in the efficiency of simple microtasks, same content chains might help alleviate mental burden on more complex microtasks [4, 26]. Therefore, task chaining has to be operationalized with a broader set of dimensions under consideration. In this paper, we extend such prior work by investigating the role of task similarity in microtask chains. In contrast to prior works, we propose a holistic definition of task similarity between a pair of tasks.

2.2 Workflows and Task Clarity

Task Clarity is also perceived as a worthwhile parameter affecting worker performance. Several authors have stressed about the positive impact of task design, clear instructions and descriptions on the quality of crowdsourced work. As per studies carried out in this domain [14], clarity was found to have direct relationship with cognitive load experienced by workers pursuing a continuous sequence of tasks, thus forming a quantifiable influence on performance.

Workflow of tasks can potentially affect worker performance in task chains. Task workflows can be defined as a sequence of steps that are required to be performed in order to complete a given task. Frequently, a task requester experiments with several alternative workflows to accomplish the task, but choose a single one for the production runs that workers can follow. Several works in the past have addressed the importance of good workflows. Workflows that create short-term goals have been shown to help workers by increasing the perceived likelihood of success [1, 10]. Lin, Mausam and Weld showed that selecting a single best workflow is suboptimal, because alternative workflows can compose synergistically to attain higher quality results [17]. Moreover, collaborative workflows have also been explored to bring about better overall performance [3].

Recent work has sought to improve workers' experiences with microtasks by inserting micro-diversions to provide timely relief during long chains [6]. Large organizations have explored redesigning assembly lines to build task specialization while still enabling task switching and creativity [2]. On crowd platforms, priming effects [20] and monetary interventions [27] can also improve performance.

In summary, prior works have touched upon some of the similarity dimensions individually and its effects on engagement of workers. Moreover it has also portrayed the importance of task chaining in a macro-scale crowdsourcing environment. In this paper, we build on the current understanding of crowd work and conceptualize task similarity as a worthwhile parameter in microtask crowdsourcing.

3 STUDY I : DIMENSIONS OF SIMILARITY

In this pilot study, we aimed to identify different dimensions of task similarity and their influence on the overall perceived similarity of a task pair (**RQ1**).

3.1 Methodology

To address **RQ1**, we manually identified a set of different dimensions based on which two tasks can be considered to be similar. These dimensions included the *workflow* of tasks (i.e., the steps required to complete the tasks successfully), *topic* of tasks (i.e., the topic(s) related to the content of the task), *time required* to complete the tasks, *time available* to complete the tasks, *batch size* of tasks, the associated monetary *reward*, *type of data* in the tasks (i.e., text, images, audio/video), task *metadata* such as title, description and keywords, and the *country* of origin of the task requesters. In light of prior works, we also considered similarity based on task types (proposed via a goal-oriented taxonomy of microtasks [12]), similarity based on task complexity [26], and on task clarity [14]. The different dimensions we considered are presented in Table 1.

We then designed and deployed a survey on the CrowdFlower[1] platform. In this survey, workers were asked to rate the influence of

[1] https://crowdflower.com/

different dimensions on the overall perceived similarity of a generic task pair, on a 7-point Likert scale (from *1: No Influence* to *7: High Influence*). In an open-ended question, workers were also asked to suggest any missing dimensions and the extent to which such dimensions influenced the overall similarity of a task pair. In order to detect untrustworthy workers and ensure reliability of responses, the survey was designed by following the guidelines proposed by Gadiraju et al. for running crowdsourced surveys [13]. To further ensure reliability, we restricted the participation of workers to *Level-3 workers*[2] on CrowdFlower.

On average, workers took 3.68 minutes to complete the survey and were compensated at an hourly rate of 7 USD. Responses from 100 different high quality workers were thus collected and aggregated, obtaining the following results.

3.2 Results

Workers in general did not mention any influential dimension of similarity that they felt was not reflected in the set of 12 similarity dimensions presented to them. Therefore, we limit our analysis to these. To compare the effect of the similarity dimensions on the overall perceived similarity of a pair of tasks, we computed a one-way between workers ANOVA across the 12 similarity dimensions considered. We found significant differences in the influence of the dimensions on the overall perceived similarity across the 12 conditions at the $p < .001$ level; $F(11, 1199) = 10.479$. Post-hoc comparisons using the Tukey-HSD test revealed that workers believe the *task type* of tasks and *workflow* influence their perception of similarity to a significantly greater extent[3] than *task complexity***, *reward***, *batch size***, *time required***, *time available***, and *country of requester***. Workers believe that the *type of data* in tasks influences their perception of similarity to a significantly greater extent than *time required**, *time available**, and *country of requester***. We found that the *topic* and *metadata* of the tasks were significantly more influential dimensions of similarity in comparison to the *country of requesters***. Finally, *task complexity* and the associated monetary *reward* were deemed to be significantly more influential in determining the overall task similarity than the *type of data* in tasks*.

Having identified a set of dimensions that influence the overall perception of task similarity, we aim to model task similarity between a pair of tasks next.

4 STUDY II : MODELING TASK SIMILARITY

In this study, we aim to propose a supervised machine learning model to predict overall similarity of a task pair (**RQ2**).

4.1 Methodology

To address **RQ2**, we adopted the following steps:

(1) Establishing ground truth for perceived task similarity.
(2) Modeling overall similarity of a task pair based on the different dimensions of similarity studied earlier.

[2] *Level-3 contributors* on CrowdFlower comprise workers who completed over 100 test questions across hundreds of different types of tasks, and have a near perfect overall accuracy.

[3] ** denotes statistical significance at the $p < .01$ level, and * denotes significance at the $p < .05$ level.

Table 1: Different similarity dimensions and their influence on the overall task similarity (aggregated from 100 distinct judgments on a 7-point Likert-scale).

Similarity Dimension	Influence (on 7-point scale)
Task Type (w.r.t. goal [12])	5.51 ± 1.42
Workflow	5.33 ± 1.32
Task Clarity	5.32 ± 1.42
Task Complexity	5.30 ± 1.36
Reward	5.30 ± 1.64
Topic	5.28 ± 1.56
Datatype	5.04 ± 1.37
Metadata	4.84 ± 1.40
Batch Size	4.53 ± 1.60
Time Required	4.29 ± 1.91
Time Available	4.25 ± 1.67
Country of Requester	3.83 ± 1.88

(3) Learning a supervised model to predict overall similarity of a task pair.

Establishing Ground Truth

To train a supervised model, we needed to establish ground truth for the overall task similarity between pairs of tasks. For this purpose, we considered the publicly available dataset of tasks sampled from Amazon Mechanical Turk used by Yang et al. to model task complexity in their work [26]. The dataset comprises of 61 distinct tasks of different types, as shown in Table 2.

Table 2: Tasks in the AMT dataset [26].

Task Type	#Tasks (in%)
Survey	6.60%
Content Creation	31.15%
Content Access	6.60%
Interpretation and Analysis	27.87%
Verification and Validation	3.30%
Information Finding	22.95%
Other	1.60%

We re-instantiated all 61 tasks and hosted them on an external server. Next, we designed and deployed a task on CrowdFlower to acquire similarity ratings from workers on different task pairs. We considered all possible pairs from the set of 61 tasks in the AMT dataset, resulting in 1,891 task pairs. In each case, workers were provided with links to the pair of tasks, asked to explore the tasks and then rate the (i) *workflow similarity*, and (ii) *topic similarity* of the given pair. To detect untrustworthy workers and ensure reliability of responses, we included test questions with priorly know answers [13, 22]. We also restricted the participation to *Level-3* workers on CrowdFlower. We gathered 3 responses for each task pair, and workers were compensated with monetary rewards at the hourly rate of 7.5 USD. Task pairs corresponding to at least one response from an unreliable worker (who failed to answer at least one test question correctly) were discarded. Thus, 1,877 task pairs were selected for learning the model.

Task Similarity Based on Individual Dimensions

Next, we computed the similarity of the task pairs with respect to each of the individual dimensions described earlier.

Task type similarity: The goal-oriented task types of each of the 61 tasks in the dataset were determined according to the taxonomy proposed by previous work [12]. Considering two generic tasks P and Q, with their corresponding task types, we compute the task type similarity $TTS(P,Q)$ as follows –

$$TTS(P,Q) = \begin{cases} 1 & \text{if the task types of } P \text{ and } Q \text{ are different} \\ 7 & \text{otherwise} \end{cases} \quad (1)$$

We use the extremities of the 7-point Likert scale used to build our ground truth to ensure adequate distinction between a pair of tasks, even if tasks are identical with respect to all other dimensions.

Workflow similarity: We propose a supervised machine learning model for computing workflow similarity between two tasks. To learn the model, we rely on the workflow similarity labels acquired from workers for each task pair as described earlier. Considering tasks P and Q, the similarity between the task pair based on each parameter (shown in Table 3) except the goal-oriented task type were calculated as follows –

$$S_i(P,Q) = 4 - 6f(\cosh(1 + |u^2 - v^2|)) \quad (2)$$

where u denotes the absolute value of parameter i of P, v denotes the absolute value of parameter i of Q and f denotes the mean normalization function. Our rationale behind choosing such a function was to ensure a considerable difference in similarity scores even after mean normalization of the difference. As *cosh* is a monotonously increasing function with steeper slopes at higher values, the mean normalized value of task P will not be in the close neighborhood of that of task Q. Task type similarity was calculated using *Eq. 1*. These similarity values were fed into a stochastic gradient descent (SGD) regressor as features, in order to predict the workflow similarity. We experimented with other regression models like support vector machine but SGD regressor was found to give least mean average error.

To evaluate the model, we compared the workflow similarity obtained from the regressor, to the ground truth aggregated from the responses of workers collected earlier. We found that the mean average error (*MAE*) observed was 0.63 on the 7-point Likert scale (approx. 9% error). Thus, the proposed model can efficiently capture workflow similarity of a task pair.

Table 3: Predicting workflow and topic similarity using supervised machine learning models. *MAE* represents the mean average error in prediction on a 7-point scale.

Dimension	Workflow	Topic
Model	SGD Regressor	SGD Regressor
Parameters	Task type, effort, time required, reward, time available, link count, image count, title length, description length, goal clarity, role clarity	top−5 topics related to task, title length, description length, title quality, description quality, language quality
MAE	0.63	0.7

Table 4: Glossary of features used to predict workflow and topic similarity between a task pair.

Feature	Definition
Task type	Goal-oriented task type of a given task [12]
Effort	Cognitive load experienced by workers during the course of the task [26]
Time Required	Time required to successfully complete the task
Reward	Monetary reward obtained after successful completion of the task
Time available	Time alloted by requester to complete the task
Link count	Number of web links in the task definition
Image count	Number of images associated with the task
Title length	Number of characters in the title of the task
Description length	Number of characters in the description of the task
Goal clarity	Extent to which the objective of a task is clear to workers [14]
Role clarity	Extent to which the steps or activities to be carried out in the task are clear [14]
Quality	Measure of understandable information carried by given entity with respect to the task

Topic similarity: We propose a supervised model to compute the topic similarity of a task pair. We rely on the topic similarity labels acquired from workers for each task pair as described earlier, to learn the model. We adopted the same approach as in case of determining *workflow similarity* between a task pair, using the different parameters shown in Table 3. We found that the proposed model is capable of predicting topic similarity with a mean average error (*MAE*) of 0.7 on 7-point scale (approx. 10% error).

Task clarity similarity: We use the task clarity model proposed in previous work by Gadiraju et al. to obtain task clarity scores for all tasks in the dataset [14]. We compute the task clarity similarity using *Eq.* 2.

Task complexity similarity: Similarly, we use the task complexity model proposed in previous work by Yang et al. to obtain task complexity scores for all tasks in the dataset [26]. We compute the task complexity similarity using *Eq.* 2.

Datatype similarity: Data associated with tasks can be of different media types; audio, image, video, textual in nature, or a combination of these media types. Therefore considering tasks P and Q, similarity (DS) with respect to data associated with tasks can be calculated as follows –

$$DS(P,Q) = 1 + 6J(u,v) \quad (3)$$

where u and v denotes set of type of data associated with P and Q respectively and $J(u,v)$ denotes Jaccard similarity of set u and v. The use of Jaccard similarity in this context is supported by the fact that the type of data involved in a particular task can be expressed as a set.

Similarity w.r.t. other dimensions: The absolute value of other dimensions like reward, batch size, task completion time, available time, etc. of the 61 tasks in the AMT dataset, were available along with the data used to compute similarities based on the aforementioned dimensions. Thus, similarity of a task pair with respect to these objective dimensions can be calculated using *Eq.* 2.

4.2 Computing the Overall Task Similarity

Based on our findings from Study I addressing **RQ1**, the overall similarity of a task pair can be computed as the weighted mean of similarities with respect to each of the individual dimensions described earlier (with their corresponding influences as weights). Considering two tasks P and Q, the overall task similarity S between the task pair can be obtained as follows –

$$S(P,Q) = \frac{\Sigma_i w(i) d(i)}{\Sigma_i w(i)} \quad (4)$$

where $d(i)$ denotes the similarity score with respect to i^{th} dimension and $w(i)$ denotes its corresponding influence (as shown in Table 1).

4.2.1 Results. By computing the overall similarity of task pairs as described above, and leveraging the workflow and topic similarity prediction models, we were able to predict overall similarity of a task pair with an mean average error (*MAE*) of 0.68 on a 7-point Likert scale. This suggests that our proposed model can efficiently capture overall similarity of a task pair.

5 STUDY III : IMPACT OF TASK SIMILARITY

In this section, we aim to investigate the impact of task similarity in microtask chains on aspects such as worker retention, worker satisfaction, boredom, and fatigue.

5.1 Methodology and Task Design

To address **RQ3** and understand the impact of task similarity on microtask chains, we considered three different conditions;

- *Similar* – In this condition, tasks are chained such that each subsequent task has a high overall task similarity with respect to the preceeding task.
- *Dissimilar* – In this condition, tasks are chained such that each subsequent task in the chain has a low overall task similarity with respect to the preceeding task.
- *Random* – In this condition, tasks are randomly chained irrespective of their overall task similarity with respect to each other.

First, we determined *similar* and *dissimilar* tasks by using the similarity scores obtained from Study II, with the help of *k–means* clustering. Next, we created distinct chains of 10 tasks each according to the three conditions described earlier. Tasks in the first chain were *similar* to each other, those in the second chain were *dissimilar* to each other while the third chain consisted of tasks selected at *random*. Table 5 presents example excerpts of task chains in the three conditions; similar, dissimilar and random).

With an aim to study the impact of task similarity in microtask chains on several aspects pertaining to workers, we deployed three different microtask chains (*similar*, *dissimilar*, and *random*) in otherwise identical fashion on CrowdFlower. Workers were asked to complete as many tasks in the chain as they wished to, with a constraint of the first 4 tasks being mandatory. We did so to ensure a bare minimum number of tasks being completed in the microtask chains by workers, that we could still reliably analyze in case of poor worker retention. After completing at least the mandatory first 4 tasks in the chain, workers were asked to answer questions regarding their *satisfaction*, *boredom*, and *fatigue* on a 7-point Likert scale. Finally, workers were also asked to rate the *overall similarity* of tasks in the chain. To ensure reliability of responses and promote high quality, we restricted the participation to *level-3* workers on CrowdFlower in each of the three microtask chain conditions. Responses from 100 distinct workers were collected and analyzed for each microtask chain condition. Workers were compensated according to the number of tasks they completed at an hourly rate of 7.5 USD.

Table 5: Example excerpts of *similar*, *dissimilar*, and *random* task chains. In this example, the first task in each of the three chains is selected to be the same, to better illustrate the difference in task similarity along the task chains in the different conditions.

Condition	**Tasks in the Chain** (first 4 of 10 tasks are shown)
Similar	(1). Find a consumable product given on a particular website (2). Find duplicate buisness names from a website (3). Find official URL of an organization (4). Find public URL of a given image . . . (10)
Dissimilar	(1). Find a consumable product given on a particular website (2). Find punctuation errors in a sentence (3). Transcribe an audio clip (4). Find address of a particular store or boutique . . . (10)
Random	(1). Find a consumable product given on a particular website (2). Find punctuation errors in a sentence (3). Rate a story with respect to its description quality (4). Find official URL of organization . . . (10)

To understand the overall impact of chaining tasks based on task similarity on the perception of workers, we acquired responses about their *satisfaction* with the tasks on a 7-point Likert scale from *1: Highly Disappointed* to *7: Highly Satisfied*. Similarly, we acquired self-reported assessments of worker *boredom* (from *1: Not Bored At All* to *7: Highly Bored*) and *fatigue* (from *1: Not Tired At All* to *7: Very Tired*), to better analyze the impact of task similarity in microtask chains. Prior works in other domains have investigated boredom [5, 23] and fatigue [16] of workers from the physical and cognitive standpoint. These works found that underutilization of cognitive resources is related to misdirection of attention resources, leading to boredom. Authors showed that fatigue can affect workers doing batches of monotonous tasks, risking a reduced well-being, and creating lower quality, unreliable data as a result. This has also formed the basis of several works in crowdsourcing that have focused on improving worker retention/engagement by reducing boredom/fatigue [7, 18, 24]. While prior works dealing with retention in microtask crowdsourcing have focused on batches of similar tasks with respect to a single dimension of similarity (that of the task objective, for example a batch of image tagging tasks), we investigate boredom, fatigue and worker retention based on our holistic definition of *overall task similarity*.

To verify the authenticity of task similarity in our task chain conditions, we also acquired responses from workers regarding their perceived task similarity of the set of tasks they completed in the chain (from *1: Highly Dissimilar* to *7: Highly Similar*).

5.2 Results

5.2.1 Worker Retention. Worker retention can be defined as the fraction of tasks that workers complete on average, from a given batch of tasks available to them. Since the first 4 of the available 10 tasks were made mandatory, we analyze worker retention in the remaining tasks in the batch. Figure 1 presents the percentage of workers who completed a given number of tasks in the chain across the three conditions.

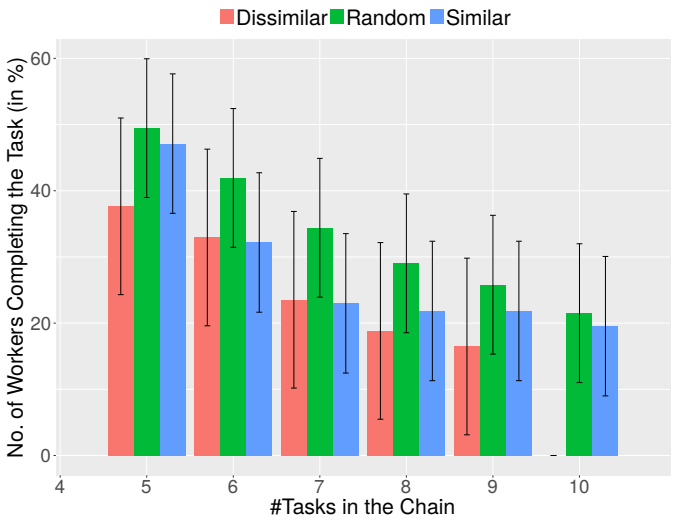

Figure 1: Average worker retention across the different task chains. Tasks 1–4 in each task chain were mandatory.

Interestingly, we found that on average worker retention is greatest in the task chain with the *random* task similarity condition (*M=33.69, SD=10.48*), when compared to that in the task chain with *similar* condition (*M=27.59, SD=10.53*), and the task chain with *dissimilar* task similarity condition (*M=21.57, SD=13.34*). We conducted a one-way between workers ANOVA to compare the effect of task similarity in microtask chains on the worker retention across the three conditions. We did not find a significant effect of task similarity in the chains on worker retention.

We note that in the *random* condition 21.5% of the workers completed all the tasks available to them, compared to 19.54% of workers in the *similar* condition. In contrast, not a single worker completed all the 10 tasks available to them in the *dissimilar* condition; 16.47% of workers completed 9 tasks. We reason that the *random* condition resulted in greater worker retention due to a balance between arousing worker interest through dissimilar tasks and maintaining continuity through similar tasks. Prior works that have suggested micro-diversions or breaks to increase worker engagement lend support to this finding [6, 24].

5.2.2 Worker Accuracy. We analyzed the accuracy of workers across the different task chain conditions, and conducted a one-way between workers ANOVA to measure the effect of task similarity in the microtask chain on the accuracy of workers. We found a significant effect of task similarity in the chain on the accuracy of workers across the three conditions at the $p < .001$ level; *F(2,261)=30.35.* Post-hoc comparisons using the Tukey-HSD test revealed that workers in the *similar* condition performed with a significantly higher accuracy on average *(M=79.68, SD=16.96)*, in comparison to workers in the *dissimilar* condition (*M=61.42, SD=16.68*), and those in the *random* group (*M=70.32, SD=12.11*) with $p < .01$. We also found that workers in the *random* condition performed significantly more accurately than those in the *dissimilar* condition with $p < .01$.

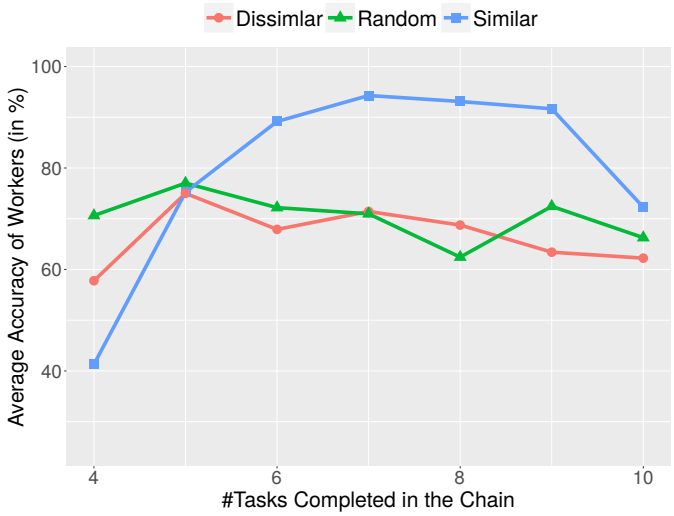

Figure 2: Average accuracy of workers with respect to the number of tasks they completed, across the different task chain conditions. Tasks 1–4 in each chain were mandatory.

Figure 2 presents the average accuracy of workers with respect to the number of tasks they completed in the chain across the three different task similarity conditions. We found that the average accuracy of workers in the *similar* condition is comparatively higher than those in the other conditions, when considering workers who completed 5 or more tasks in the chain of 10 tasks.

5.2.3 Task Completion Time. To compare the effect of task similarity on the task completion time of workers across the different conditions, we conducted a one-way between workers ANOVA. Results confirmed a significant effect of task similarity on the task completion time of workers at the $p < .001$ level; $F(2, 297) = 18.99$. Post-hoc comparisons revealed that workers take significantly more time on average to complete the task chains corresponding to the *dissimilar* (*M=15.67 mins, SD=6.57*) and *random* (*M=15.51 mins, SD=1.83*) conditions when compared to the *similar* condition (*M=11.77 mins, SD=5.55*) at the $p < .01$ level. Since worker retention varied across the different conditions (i.e., workers completed different number of tasks in the different conditions on average), we also computed the average time workers took to complete a single task in the chain. Once again, we found that workers in the *dissimilar* condition took more time to complete a single task (*M=2.21 mins, SD=1.04*) than workers in the *similar* (*M=2.13 mins, SD=1.21*), or the *random* condition (*M=2.02 mins, SD=1.08*). These differences were not found to be statistically significant.

5.2.4 Overall Task Similarity and Worker Satisfaction. Next, we analyzed the responses of workers to the questions regarding their perception of the overall task clarity of the task chains, and their corresponding satisfaction on 7-point scales. Our findings are presented in Figure 3. We found that workers rated the overall task clarity of tasks in the *similar* chain to be higher (*M=3.82, SD=1.67*) than that in the *dissimilar* (*M=3.61, SD=1.65*) or *random* chain (*M=3.67, SD=1.47*). Workers exhibited a higher satisfaction with tasks in the *random* chain (*M=5.49, SD=1.31*) in comparison to those in *similar* (*M=5.06, SD=1.62*) and *dissimilar* chains (*M=5.16, SD=1.41*). However, multiple T-tests revealed a lack of statistical significance between these comparisons.

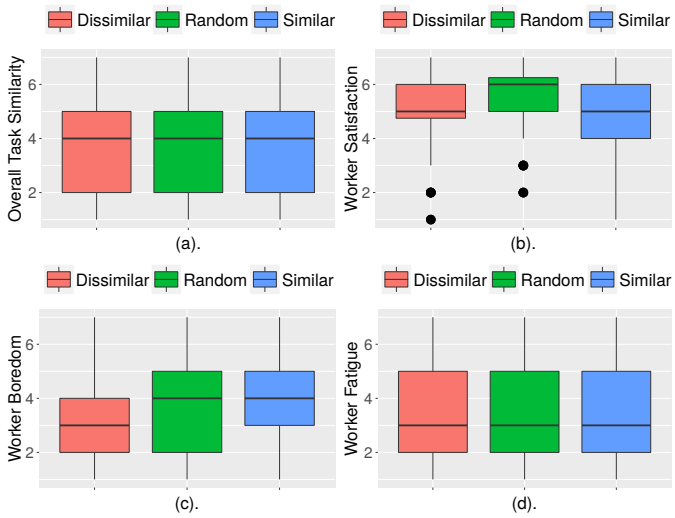

Figure 3: A comparison of the perceived (a). overall task similarity, (b). worker satisfaction, (c). worker boredom, and (d). worker fatigue across the three task chain conditions.

5.2.5 Boredom and Fatigue. We found that workers in the *similar* task chain (*M=3.38, SD=1.78*) experienced the most fatigue while those performing the *dissimilar* tasks experienced the least fatigue (*M=3.16, SD=1.80*). Workers in the *random* chain corresponded to a fatigue of (*M=3.37, SD=1.74*). We did not find a statistically significant difference between these comparisons using a one-way between workers ANOVA.

We conducted another one-way between workers ANOVA to compare the effect of task similarity in chains on worker boredom. Results confirmed a significant effect of task similarity on boredom at the $p < .05$ level; $F(2, 297) = 3.068$. Post-hoc comparisons using the Tukey-HSD test confirmed that workers experience most boredom while completing a chain of *similar* tasks (*M=3.90, SD=1.64*), followed by chain of *random* tasks (*M=3.30, SD=1.75*) and least while doing a chain of *dissimilar* tasks (*M=3.58, SD=1.72*). The difference between the boredom experienced by workers in the *similar* and *dissimilar* chain conditions was found to be statistically significant at the $p < .05$ level. This is in line with our intuitive expectations as discussed earlier. Prior studies in behavioral psychology prove that repeatedly performing tasks which are similar to each other reduces the interest of workers and induces boredom [5, 23].

6 CONCLUSIONS AND FUTURE WORK

In this paper, we investigated the role of task similarity in microtask chains and how it affects worker performance. We successfully identified different similarity dimensions and their influences. Using a supervised model, we were able to model overall task similarity of a task pair. Next, we studied the impact of task similarity in microtask chains on worker retention, satisfaction, boredom and fatigue.

Our studies reveal that ordering tasks in a chain according to overall task similarity results in improved accuracy, but at the cost of inducing boredom. This paper points at the necessity of striking a balance between similarity and dissimilarity in a chain of tasks so as to bring forth better engagement of workers. Our findings enrich the current understanding of crowd work and bear important implications on structuring workflow. Further studies into the similarity dimensions will help us in striking a better balance during task chaining and is reserved for future work.

ACKNOWLEDGEMENTS

This research has been supported in part by the European Commission within the H2020-ICT-2015 Programme (*Analytics For Everyday Learning* (AFEL) project, Grant Agreement No. 687916).

REFERENCES

[1] David S Ackerman and Barbara L Gross. 2005. My instructor made me do it: Task characteristics of procrastination. *Journal of Marketing education* 27, 1 (2005), 5–13.
[2] Paul S Adler, Barbara Goldoftas, and David I Levine. 1999. Flexibility versus efficiency? A case study of model changeovers in the Toyota production system. *Organization Science* 10, 1 (1999), 43–68.
[3] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2012. Collaborative workflow for crowdsourcing translation. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1191–1194.
[4] Carrie J Cai, Shamsi T Iqbal, and Jaime Teevan. 2016. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3143–3154.
[5] Jonathan SA Carriere, J Allan Cheyne, and Daniel Smilek. 2008. Everyday attention lapses and memory failures: The affective consequences of mindlessness. *Consciousness and cognition* 17, 3 (2008), 835–847.
[6] Peng Dai, Jeffrey M. Rzeszotarski, Praveen Paritosh, and Ed H. Chi. 2015. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*. 628–638.
[7] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 2015. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 628–638.
[8] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*.
[9] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*. 238–247.
[10] Joseph R Ferrari, Judith L Johnson, and William G McCown. 1995. Procrastination research. In *Procrastination and Task Avoidance*. Springer, 21–46.
[11] Ujwal Gadiraju and Stefan Dietze. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, March 13-17, 2017*. 105–114.
[12] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*. 218–223.
[13] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online

surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1631–1640.

[14] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT 2017, Prague, Czech Republic, July 4-7, 2017*. 5–14.

[15] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

[16] Gerald P Krueger. 1989. Sustained work, fatigue, sleep loss and performance: A review of the issues. *Work & Stress* 3, 2 (1989), 129–141.

[17] Christopher H Lin, Daniel S Weld, et al. 2012. Dynamically switching between synergistic workflows for crowdsourcing. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[18] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why stop now? predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[19] Stephen Monsell. 2003. Task switching. *Trends in cognitive sciences* 7, 3 (2003), 134–140.

[20] Robert R Morris, Mira Dontcheva, and Elizabeth M Gerber. 2012. Priming for better performance in microtask crowdsourcing environments. *IEEE Internet Computing* 16, 5 (2012), 13–19.

[21] Edward Newell and Derek Ruths. 2016. How one microtask affects another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3155–3166.

[22] David Oleson, Alexander Sorokin, Greg P Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Human computation* 11, 11 (2011).

[23] Nathalie Pattyn, Xavier Neyt, David Henderickx, and Eric Soetens. 2008. Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue? *Physiology & Behavior* 93, 1 (2008), 369–378.

[24] Jeffrey M Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. 2013. Inserting micro-breaks into crowdsourcing workflows. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[25] Glenn Wylie and Alan Allport. 2000. Task switching and the measurement of "switch costs". *Psychological research* 63, 3 (2000), 212–233.

[26] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling Task Complexity in Crowdsourcing. In *In Proceedings of The Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016)*. 249–258.

[27] Ming Yin, Yiling Chen, and Yu-An Sun. 2014. Monetary interventions in crowdsourcing task switching. In *Second AAAI Conference on Human Computation and Crowdsourcing*.