©ACM 2015. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in i-KNOW.

MicroTrails: Comparing Hypotheses about Task Selection on a Crowdsourcing Platform

Martin Becker
University of Würzburg
Germany
becker@informatik.uni-wuerzburg.de

Kathrin Borchert
University of Würzburg
Germany
kathrin.borchert@informatik.uni-wuerzburg.de

Matthias Hirth
University of Würzburg
Germany
matthias.hirth@informatik.uni-wuerzburg.de

Hauke Mewes
University of Würzburg
Germany
hauke.mewes@stud-mail.uni-wuerzburg.de

Andreas Hotho
University of Würzburg and
L3S Research Center
Germany
hotho@informatik.uni-wuerzburg.de

Phuoc Tran-Gia
University of Würzburg
Germany
trangia@informatik.uni-wuerzburg.de

ABSTRACT

To optimize the workflow on commercial crowdsourcing platforms like Amazon Mechanical Turk or Microworkers, it is important to understand how users choose their tasks. Current work usually explores the underlying processes by employing user studies based on surveys with a limited set of participants. In contrast, we formulate hypotheses based on the different findings in these studies and, instead of verifying them based on user feedback, we compare them directly on data from a commercial crowdsourcing platform. For evaluation, we use a Bayesian approach called HypTrails which allows us to give a relative ranking of the corresponding hypotheses. The hypotheses considered, are for example based on task categories, monetary incentives or semantic similarity of task descriptions. We find that, in our scenario, hypotheses based on employers as well the the task descriptions work best.

Overall, we objectively compare different factors influencing users when choosing their tasks. Our approach enables crowdsourcing companies to better understand their users in order to optimize their platforms, e.g., by incorparting the gained knowledge about these factors into task recommentation systems.

CCS Concepts

•Information systems → Crowdsourcing;

Keywords

Human Trails, HypTrails, Crowdsourcing

1. INTRODUCTION

Crowdsourcing plattforms are a relatively new type of large scale Internet services and represent a specific type of online labour markets. In contrast to traditional forms of organizing work, the crowdsourcing paradigm is characterized by the fact that tasks are not assigned to a specific person. Instead, employers define *campaigns* consisting of a set of *tasks* which are made available on the crowdsourcing platforms. Then, the users of this platform — so called *workers* — freely choose from the pool of available tasks. The granularity of work on crowdsourcing platforms is smaller than in traditional forms of work organization [4]. This results in pools of hundreds to thousands of different tasks. [5].

Problem setting The large number of tasks and campaigns on crowdsourcing platforms poses challenges: On the one hand, *workers* face the issue of efficiently finding tasks fitting their profile, e.g., according to their skills or their interest. On the other hand, *employers* need all their tasks to be completed. Both interests have to be addressed by the crowdsourcing companies. To solve these issues, mechanisms like task recommendation systems have been identified as a relevant research topic [8].

However, recommendation systems as well as similar mechanisms need prior knowledge about task selection preferences. Unfortunately, there is only little information about how workers choose tasks on crowdsourcing platforms, yet: Current studies are based on surveys which only cover small subsets of workers and are also highly subjective. Thus, in this work we objectively evaluate the influence factors involved in the selection process of tasks.

Method To this end, we interpret the set of tasks a user has completed as ordered trails. Then, we can formulate hypotheses about how these trails emerge based on conditional transition probabilities between campaigns. This allows us to embed the observed transitions as well as the formulated hypotheses in a first order Markov chain model. We then use the Bayesian approach HypTrails to objectively evaluate these hypotheses and derive a relative ranking. Based on results from related papers, we formulate our hypotheses and compare them directly on data from the commercial crowdsourcing platform Microworkers.com[1] including the work history of $39,100$ workers over 6 years. Among others, the hypotheses, considered in this work, are based on campaign categories, monetary incentives, or description similarity of campaigns.

Findings and contribution We objectively evaluate a considerable set of hypotheses and find that, in our scenario, those based on work categories and employers as well as campaign descriptions work

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

i-KNOW '15, October 21 - 23, 2015, Graz, Austria

© 2015 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ISBN 978-1-4503-3721-2/15/10…$15.00

DOI: http://dx.doi.org/10.1145/2809563.2809608

[1] https://microworkers.com/ (accessed: Aug. 2015)

best. Our approach enables crowdsourcing companies to better understand their users in order to optimize their platforms, e.g., by incorporating the gained knowledge about these factors into task recommendation systems.

Structure The reminder of this paper is structured as follows. Section 2 gives a brief overview of related work on influence factors on task selection in commercial crowdsourcing environments. The methodology applied in this paper and the underlying dataset are described in Section 3. The considered hypotheses are presented in Section 4, the results of our experiments in Section 5. Section 6 discusses the results and Section 7 concludes the paper.

2. RELATED WORK

The motivation of working on crowdsourcing platforms in general and the preferences of selecting tasks are the subject of several studies. Most of this research is based on user surveys leading to varying answers depending on the way questions are asked, and consquently limiting the understanding of the respective influence factors [10]. However, the influence factors derived form such studies can be used for formulating hypotheses about how users choose their tasks, which are objectively evaluated in this paper.

Aris [1] reviews research results of motivational factors of participation in the area of mobile crowdsourcing. In contrast to the Microworkers platform investigated in this paper, the platforms and services analyzed by Aris are from the field of creative tasks. This includes for example, participating in innovation contests, generating news content, or even more specialized social tasks like assisting foreign visitors in Japan. The main influence factor in the reviewed studies was found to be "personal benefit", which can be categorized into intrinsic and extrinsic motivation. Intrinsic motivation is given if a task is fun, a new experience is gained, or because it is challenging. Whereas, extrinsic motivation describes participation based on awards, like points or a monetary reward. Overall, Aris finds that intrinsic aspects are more important than the extrinsic. Furthermore, Aris recognizes that the results about monetary rewards are not consistent. It can be assumed, that similar to the case of mobile crowdsourcing, users on micro tasking platforms, like the Microworkers platform we study in this paper, are also affected by intrinsic and extrinsic factors.

Indeed, a model for the workers' motivation by Kaufmann et al. [7]. has confirmed the importance of intrinsic aspects on Amazon Mechanical Turk (mturk). At the same time, extrinsic factors have been found to be relevant. This includes task related factors as well as motivation based on learning and training skills. Regarding extrinsic factors, Chilton at al. [3] also found that task related properties and characteristics, like the creation date or the overall number of tasks provided by a campaign, influence the selection of tasks. The results are based on the analysis of data scraped from mturk and a survey about the workers' task searching behavior.

In contrast to Aris and Kaufmann, the user study of Yuen et al. [12] shows that a high monetary reward is the most important task selection criterion. In addition, the workers answered that they choose their work based on the nature and the difficulty of the task.

Finally, Schulze et al. [10] show that the preferences and influence factors differ with respect to the location of the workers: For workers from the United States, the most important aspect for selecting a task is their interest in it. This is followed by payment, the simplicity of tasks and a high reputation of the employer. In contrast, Indian workers prefer well payed and simple tasks.

Schnitzer et al. [9] confirm this observation by an user study about worker demands on task recommendation. Here, the similarity of tasks is the most important task property for workers from the United States whereas Asian and European workers are most interested in tasks offering the most money.

Overall, there are many factors influencing the selection of new tasks in crowdsourcing environments. Some results are even contradicting. However, in contrast to our work, none of the papers cited, conduct an objective comparison of the proposed factors.

3. BACKGROUND, METHODOLOGY AND DATA

The main goal of this paper is to study how users choose tasks on crowdsourcing environments. To this end, we apply an approach called HypTrails [11] and utilize data from the crowdsourcing platform Microworkers. In this section, we first introduce the problem setting and corresponding terminology. Then, we briefly review the HypTrails approach and establish how it is applied to the problem setting. Finally, we characterize the data we are working with.

3.1 Background

Commercial crowdsourcing environments usually involve three actors, i.e., (i) platform users submitting work to the platforms, so called *employers*, (ii) users completing work submitted to the platform, so called *workers*, and (iii) the *platform operators*. As mentioned before, unlike in traditional forms of work organization, employers do not choose dedicated workers for completing the submitted work. Instead they define certain tasks and make them available through the crowdsourcing platform. The workers can then freely choose from the currently available work. Usually employers never communicate directly with workers. Instead, the platform and its operators are responsible for providing means to publish work on the platform, submit completed work, and transfer remuneration between worker and employer.

While a wide variety of crowdsourcing platforms exists, microtasking platforms, such as Microworkers, focus on highly repetitive tasks which can be completed in a short amount of time (a few minutes up to an hour). Micro-tasks include, e.g, tagging a series of images or categorizing the sentiment of a set of short text messages. Due to the repetitive nature of micro-tasks, employers publish a *campaign* which describes a class of tasks and set a number of task instances for workers to complete for that particular campaign. A campaign ends when all tasks have been completed. Each task can only be chosen once by a single worker. On the Microworkers platform, which we consider in this work, workers also cannot choose more than one task from the same campaign. Thus, in the following, the notion of *campaign* and *task* are used interchangeably.

Our main goal is to study how workers choose their tasks in crowdsourcing environments. Since we use HypTrails as our method of choice, we need to model this process as a set of transitions between campaigns. As each task is associated with a campaign, we can also derive a "trail" of campaigns for each user. That is, each trail consists of the campaigns associated with the tasks she has completed consecutively. At the same time, these trails define the transitions we need for the HypTrails approach.

3.2 Methodology

HypTrails is an approach for expressing and comparing hypotheses about human trails. Since we know the sequence of tasks a user has completed and since each task is associated with a campaign, we can derive a sequence of campaigns for each user. Then, these sequences of campaigns, called *microtrails*, are the trails required by HypTrails. Technically, HypTrails is based on Markov chain modeling and Bayesian inference. In the following, we very briefly outline the main concepts and ideas of HypTrails and explain how it is applied to our scenario. For more details about HypTrails, we refer the reader to the original paper [11].

Fundamentally, HypTrails models trails as a first-order Markov chain, a stochastic system that incorporates transition probabilities between states, which are in our scenario given by campaigns. Hypotheses can be expressed as belief in Markov transitions, i.e., assumptions on common and uncommon transitions at individual states. Section 4 presents several such hypotheses about how users choose their next campaign. To obtain insights into the relative plausibility of a set of hypotheses given data, HypTrails resorts to Bayesian inference, i.e., the corresponding marginal likelihood (*evidence*) denoting the probability of the data given a hypothesis H. The main idea of HypTrails is to utilize the influence of the prior on the evidence for comparing hypotheses with each other. In particular, expressed hypotheses get incorporated into the inference process by eliciting Dirichlet priors for each hypothesis. This is done via the hyperparameters of the Dirichlet distribution which can be interpreted as pseudo counts, i.e., transition counts before observing the data. These Dirichlet prior pseudo counts reflect the assumptions for a given hypothesis—simply said, the higher the belief in a given transition is, the higher the pseudo counts should be. Additionally, the strength of overall belief in a hypothesis can be influenced by a parameter K that increases the overall number of pseudo counts (concentration) assigned to the prior. The higher the overall number of pseudo counts, the more concentrated the probability mass of a sample from a Dirichlet distribution is. This results in higher evidences for correct hypotheses and less leeway for imperfect hypotheses. Low values of K add some "tolerance", allowing for hypotheses which are based on imperfect assumptions but capturing prominent factors of the underlying processes to also reach high evidence values. This helps in explaining to what extent the different hypotheses capture the data. The elicitation for each K—i.e., the proper construction of Dirichlet priors from expressed hypotheses—is done automatically by HypTrails.

Different hypotheses then lead to different evidence values which are extracted from the observed data via Bayesian inference. When we compare two hypotheses, a higher evidence for a given hypothesis indicates higher plausibility. The fraction of the evidence of two hypotheses (priors), called Bayes factor [6], is then used for determining the strength of evidence for one hypothesis over the other. A Bayes factors can be directly interpreted as the Bayesian equivalent to a frequentist's significance value. In this article, all Bayes factors for reported results are decisive. Therefore, we refrain from explicitly reporting them individually and refer to [11] for further details. Instead we can directly compare evidences which we express on a log scale—higher evidence means higher plausibility.

Given a set of generic hypotheses about how users choose their next campaign, the following steps are necessary to apply HypTrails to our scenario: (i) Given a set of properties for each campaign and based on the ideas the hypotheses are grounded on, we specify a hypothesis matrix Q for each hypothesis. Q quantifies our assumptions about the transitions between campaigns observed in the data. Higher values correspond to a stronger belief in a transition. No negative values are allowed. We describe this process as well as our hypotheses in detail in Section 4. (ii) Next, we pass these matrices as well as the observed campaign transitions (see Section 3.3) to HypTrails which then subsequently elicits the Dirichlet priors for each hypothesis with varying values of the belief concentration factor K. (iii) Based on this elicitation, HypTrails determines the evidences for each hypothesis and each parameter K.

As mentioned above, for simplicity, we can state that one hypothesis H_1 is more plausible compared to another hypothesis H_2, if the evidence of H_1 is higher than the one of H_2 for the same value of K. Thus, the partial ordering based on the plausibility of respective hypotheses $\mathbf{H} = \{H_1, H_2, \ldots, H_n\}$ can be determined by ranking their evidences from largest to smallest for single values of K. In this work, we express evidences on a log scale. We present corresponding results in Section 5.

3.3 Data

We use a dataset from the crowdsourcing platform Microworkers. The data includes anonymized information about campaigns and users in a time period between the founding of the platform in May 2009 and January 2015. We will first explain two restrictions on the dataset we need in order to apply HypTrails and then introduce several campaign features we will use for defining hypotheses.

Data restrictions With respect to some special characteristics and features of Microworkers we have to limit the utilized data for our HypTrails computation. Microworkers offers the possibility for employers to restrict their campaigns to workers from specific countries. Since HypTrails cannot distinguish between different user groups we cannot model this directly. Thus, we choose to focus on US workers because they have access to most campaigns. In order to do so, we remove campaigns which place restrictions on US workers as well as campaign transitions from non-US workers.

Additionally, instead of releasing all tasks of a campaign at once, employers can set its tasks to be released successively at a certain speed. However, since we define hypotheses based on transition probabilities between campaigns, we need to model which campaigns are available after finishing a task. The task release speed feature complicates this process. Thus, we only consider campaigns with a large enough speed in order to guarantee that this does no restrict the workers artificially.

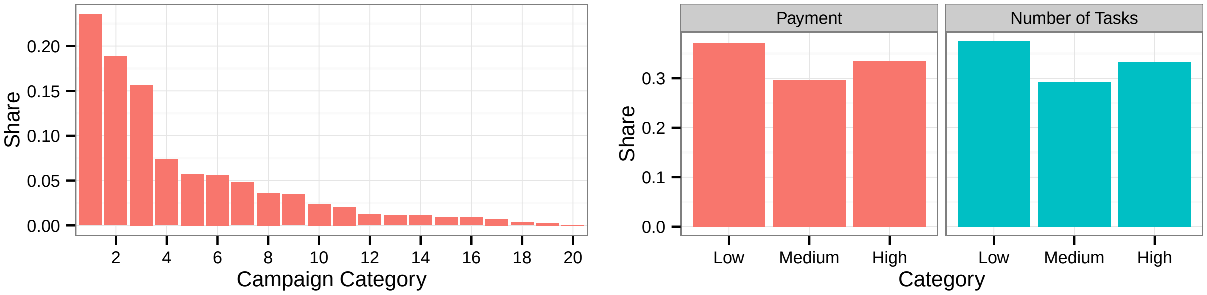
In spite of these restrictions, our final dataset still contains $81,544$ campaigns and $3,415,119$ completed tasks. This includes 95% of the US workers and corresponds to 55% of the campaigns available to them as well as 60% of their completed tasks.

Campaign features For defining hypotheses in Section 4 we use several features based on campaign properties. These include campaign categories, payment, the time required to finish a task, payment per hour and the number of tasks offered by a campaign. These properties are introduced in the following.

On the Microworkers platform each campaign is associated with one of 20 campaign *categories*, e.g., "promotion" or "writing". The distribution of the campaigns per category is shown in Figure1(a). The campaigns are not uniformly distributed, i.e., three categories are very prominent. Each of these categories contains between 15% and 25% of the campaigns, whereas most of the other categories only contain 5% or less. On the one hand the popular categories include simpler tasks with lower requirements to complete these tasks successfully. On the other hand, new categories were added over time, which is also responsible for this imbalance.

Besides the categories, campaigns differ concerning their *payment*. However, payments are strongly skewed towards small amounts of money. Since we later want to gauge if workers generally tend to choose campaigns which are better payed, we divide the payment range into three classes, i.e., low, medium and high. We choose these intervals so that campaigns are as equally distributed as possible. Figure 1(b) shows the resulting distribution. The intervals are: \$0.0 to \$0.15 for low, \$0.15 to \$0.30 for medium, and amounts of more than \$0.30 for highly paid tasks.

The *time required to finish a task* is also a feature we will use to derive hypotheses. The time is set by the employer and is an estimation about how long a single tasks will take approximately. However, here it is not possible to derive equally sized intervals since one time setting is far too prominent.



(a) Share of campaigns per predefined category on Microworkers.

(b) Share of campaigns per payment and campaign size classes.

Figure 1: Selected statistics of the Microworkers dataset.

Using the payment and the time required to finish a task, we can derive the *payment per hour (pph)* for each campaign. Again, we define intervals so that campaigns are as equally distributed as possible. The intervals are: $0.0 to $2.4 for low, $2.41 to $6.0 for medium, and amounts of more than $6.0 for high.

Additionally, each campaign defines a different number of tasks (also called *positions*) which ranges from 30 up to several hundred. Similar to payment, campaigns often only provide a rather low number of tasks. Consequently, we again choose to define equally sized intervals. The intervals are: only campaigns with 30 tasks for low, 31 to 90 tasks for medium, and 90 tasks and up for high volume campaigns. The resulting distribution is depicted in Figure 1(b).

4. HYPOTHESES

In our case, hypotheses about how users choose their campaigns are based on transition probabilities between campaigns. That is, given a campaign $c_i \in C$, we need to define the probability to choose any other campaign $c_j \in C$:

$$P(c_j|c_i)$$

These probabilities represent the entries of the hypothesis matrix Q required by the HypTrails method as introduced in Section 3. HypTrails uses these matrices to elicit corresponding priors and compare the respective hypotheses.

Note, that for simplicity purposes, instead of conditional probability distributions, we define transition functions $\bar{P}$ which do not sum up to one. However, those can easily be converted:

$$P(c_j|c_i) = \frac{1}{\sum_{c_j \in C} \bar{P}(c_j|c_i)} \bar{P}(c_j|c_i)$$

4.1 Uniform hypothesis and availability

With the uniform hypothesis we assume that, after finishing a task from a campaign, a user will randomly choose a task from any other campaign. Formally:

$$\bar{P}_{uni}(c_j|c_i) = 1$$

However, since campaigns are only available for a user to choose as long as some of its tasks have not been completed, the point in time when a user chooses her next campaign defines the set of campaigns available to choose from. That is, a user can not choose a campaign whose tasks have all been completed. Now, assuming that a user just finished campaign c_i. Then, let $[s_i, e_i]$ denote the time interval in which campaign c_i was active, i.e., s_i is the time campaign c_i was made available for users and e_i is the time the last task has been completed. Thus, the user finished her task sometime between s_i and e_i. In the best case, the user finished her task at s_i, i.e., right after campaign c_i started. If we now assume that a user may wait arbitrarily long before choosing her next campaign, all campaigns c_j which end after campaign c_i has started ($e_j \geq s_i$) are available to the user. However, all campaigns $c_{j'}$ which have ended before campaign c_i has started ($e_{j'} < s_i$) are not available to the user. Now, for all users, given a campaign c_i they just finished, we define a set of available campaigns C_i:

$$C^i_{after} = \{c_j \in C | e_j > s_i\}$$

This availability setting assumes that a user may take an arbitrarily long time for choosing her next campaign. However, since tasks usually require a short amount of time to complete and campaigns only pay small amounts of money for theses tasks, a user whose goal is to earn a sensible amount of money will choose her campaigns in quick succession. Thus, it is realistic to assume that users only pick from campaigns available at the time they finish. This can be modelled by assuming that a campaign c_j is only available from another campaign c_i if the active time of both campaigns overlaps, i.e. $[s_i, e_i] \cup [s_j, e_j] \neq \emptyset$. Formally, the corresponding set of available campaigns, given a campaign c_i is defined as

$$C^i_{overlap} = \{c_j \in C | [s_i, e_i] \cup [s_j, e_j] \neq \emptyset\}$$

Thus, given different definitions of availability C^i, the most natural extension of the uniform hypothesis is to set the probability of campaigns, which are not available from a given campaign, to zero:

$$\bar{P}_{av}(c_j|c_i) = \begin{cases} 1, & \text{if } c_j \in C^i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

We formulate two corresponding hypotheses, namely $\bar{P}_{after}$ and $\bar{P}_{overlap}$. The uniform hypothesis $\bar{P}_{uni}$ is equivalent to the availability hypothesis where all campaigns are available from every campaign: $C^i = C$.

4.2 Category and employer

Crowdsouring platforms often define a set of categories in order to group certain types of campaigns (see Section 3.3). Building on the idea that users like certain types of tasks better than others (for example interesting ones as indicated by Aris [1]), we propose a hypothesis which favors follow-up campaigns of the same category. Let the category of a campaign c_i be denoted as cat_i, and let α define the weight for campaigns with a category of the same type and β define the weight for campaigns with a category of a different type, then the corresponding hypothesis is:

$$\bar{P}_{cat}^{\alpha,\beta}(c_j|c_i) = \begin{cases} \alpha \cdot \bar{P}_{av}(c_j|c_i), & \text{if } cat_i = cat_j \\ \beta \cdot \bar{P}_{av}(c_j|c_i), & \text{otherwise} \end{cases} \tag{2}$$

For this, we only consider campaigns c_j *available from* the given campaign $c_i \in C$ as denoted by the factor $\bar{P}_{av}(c_j|c_i)$.

Furthermore, Schulze et al. [10] have shown that reputable employers are generally favored. Thus, we define a hypothesis assuming that users are consistent with regard to their employer when choosing their new tasks. Let the employer of a campaign be denoted as emp_i, then we define the employer hypothesis as:

$$\bar{P}_{emp}^{\alpha,\beta}(c_j|c_i) = \begin{cases} \alpha \cdot \bar{P}_{av}(c_j|c_i), & \text{if } emp_i = emp_j \\ \beta \cdot \bar{P}_{av}(c_j|c_i), & \text{otherwise} \end{cases} \tag{3}$$

While both hypotheses can be a good explanation for how users choose their next task, we want to combine the two notions. That is, we assume that users choose from the same employer and at the same time they also like to work on the same type of tasks, thus also staying consistent regarding the category. This hypothesis is then defined as

$$\bar{P}_{cat\&emp}^{\alpha,\beta,\beta',\gamma}(c_j|c_i) = \begin{cases} \alpha \cdot \bar{P}_{av}(c_j|c_i), & \text{if } cat_i = cat_j \\ & \wedge\, emp_i = emp_j \\ \beta \cdot \bar{P}_{av}(c_j|c_i), & \text{if } cat_i = cat_j \\ \beta' \cdot \bar{P}_{av}(c_j|c_i), & \text{if } emp_i = emp_j \\ \gamma \cdot \bar{P}_{av}(c_j|c_i), & \text{otherwise} \end{cases} \tag{4}$$

In this work we set $\beta = \beta'$ and write $\bar{P}_{cat\&emp}^{\alpha,\beta,\gamma}(c_j|c_i)$.

However, as mentioned in Section 3.3, the overall distribution of campaigns within categories and from employers is not equally distributed. As a result, there are significantly more campaigns in some categories. Consequently, more campaigns of this category will be chosen by workers. This possibly favors the category and employer hypotheses mentioned above. In order to investigate if this is true, we also formulate a hypothesis based on overall category frequencies. Let f_i denote the frequency of a category in our corpus. Then we define:

$$\bar{P}_{cat_f}(c_j|c_i) = f_j$$

Equivalently, we define $\bar{P}_{emp_f}$, for employer frequencies.

4.3 Payment, positions and time

As has been shown in several studies [10, 12], the amount of money to be earned from a task can be considered a decisive factor in choosing new tasks. In this context there are two aspects to cover, namely, task payment [12] and hourly earnings [10]. The former is the amount of money to be payed for finishing a task. The latter also takes into account the estimated time required to finish a task. Another factor which has been shown to influence the users preferences to choose tasks is the number of available positions of the campaign [10, 3]. That is, different campaigns provide different numbers of tasks and users seem to favor campaigns which provide more tasks.

Both, the payment and the position factors, have in common that a higher value, i.e., higher payment or more positions, implies a higher probability to choose the corresponding campaign. Let $value_i$ denote the corresponding value. A straight forward formulation of a corresponding hypothesis would be

$$\bar{P}_{value}(c_j|c_i) = value_i$$

However, when looking at the value distributions, we notice that payment as well as positions are strongly skewed towards low values. In order to model a tendency towards higher values, we divide the value range into stratified intervals which consist of an equal number of campaigns. In our case we choose three different classes: low, middle, high. For further details, see Section 3.3. Now, let $class_i$ denote the class, a campaign c_i is assigned to. Then the hypothesis is formulated as:

$$\bar{P}_{lmh}^{\alpha,\beta,\gamma}(c_j|c_i) = \begin{cases} \alpha \cdot \bar{P}_{av}(c_j|c_i), & \text{if } class_i = \text{low} \\ \beta \cdot \bar{P}_{av}(c_j|c_i), & \text{if } class_i = \text{middle} \\ \gamma \cdot \bar{P}_{av}(c_j|c_i), & \text{if } class_i = \text{high} \end{cases} \tag{5}$$

Based on this, we define three hypotheses for payment, payment per hour and positions, namely $\bar{P}_{pay}(c_j|c_i)$, $\bar{P}_{pph}(c_j|c_i)$ and $\bar{P}_{pos}(c_j|c_i)$ respectively.

In Section 3.3 we have also mentioned the time required to finish a task as a factor influencing the user when choosing a new task. A long required time may deter users from choosing the corresponding task [10]. Assuming normalized time values $value_i$ in a range from zero to one we define the corresponding hypothesis as

$$\bar{P}_{time}(c_j|c_i) = 1 - value_i$$

Since we were not able to derive stratified classes for the required time spans, we are not formulating a hypothesis based on intervals.

4.4 Title and Description

The category hypothesis assumes that tasks of the same type are chosen consistently. However, this may not accurately represent similarity of tasks required for example to capture the notion of always choosing interesting tasks [10]. This is especially true considering the skewed distribution of campaigns across categories. Thus, we further investigate this line of thought by comparing the title and the description of the campaigns instead of just their categories. Both the title and the description can be represented as a bag-of-words. Thus, to compare titles and description, respectively, we are employing the cosine distance based on TF-IDF vectors[2] (we use MLlib[2] to calculate the corresponding vectors).

Note, that we do not apply any other pre-processing steps like stop-word removal or stemming. Now, let $tf\text{–}idf_i$ denote the TF-IDF vector of a document, i.e., either a title or a document, then we define the respective hypothesis as:

$$\bar{P}_{cos}(c_j|c_i) = cos(tf\text{–}idf_i, tf\text{–}idf_j)$$

For the corresponding hypotheses for the titles and the descriptions, we write $\bar{P}_{title}$ and $\bar{P}_{desc}$ respectively.

5. EXPERIMENTS

In order to compare the relative plausibility of the hypotheses introduced in Section 4, we apply the HypTrails approach as outlined in Section 3.2 based on the data described in Section 3.3. The results for the individual hypotheses are reported in Section 5.1 through 5.4 and visualized in Figure 2. We start by assessing the performance of the uniform hypothesis and different availability assumptions. We find that the availability assumption based on overlap is the most realistic one. Thus, the following hypotheses are all grounded on the availability based on this assumption. We give a summarizing comparison of hypotheses in Section 5.5. Overall, the hypothesis based on employers works best, directly followed by

[2] https://spark.apache.org/mllib/ (Accessed: Aug. 2015)

the descriptions. This indicates that employers as well as semantic similarity of campaigns play an important role for users when choosing campaigns.

5.1 Uniform and availability

Since campaigns are only available for a limited amount of time, i.e., as long as some tasks have not been completed, we have introduced several notions of availability in Section 4.1. In this section, we compare the corresponding availability hypotheses: $\bar{P}_{uni}$, $\bar{P}_{after}$ and $\bar{P}_{overlap}$. Figure 2(a) shows the results. Generally, the uniform hypothesis is the most unrealistic one since it assumes the availability of all campaigns from every campaign. Thus, as expected, it performs worse than both, the $\bar{P}_{after}$ and the $\bar{P}_{overlap}$ hypothesis. When comparing the latter two, the overlap hypothesis $\bar{P}_{overlap}$ is strongly superior. As outlined in Section 4.1 this is due to the fact that users choose their campaigns in quick succession. Since the overlap availability $\bar{P}_{overlap}$ is the one which is most plausible in our setting, we use it as the $\bar{P}_{av}$ component required by our other hypotheses as outlined in Section 4. Thus, we are looking for hypotheses which improve on $\bar{P}_{overlap}$. Consequently $\bar{P}_{overlap}$ serves as our baseline.

5.2 Category and employer

As introduced in Section 4.2, some straightforward hypotheses are those based on the fact that users may tend to choose campaigns from categories and/or employers which they already know. In this section, we compare different category and employer based hypotheses. First we investigate if users tend to pick campaigns from the same category or the same employer. Afterwards we combine both factors.

Figure 2(b) shows the results for campaigns and employers separately. Independent of the respective parameters, both hypotheses are superior to the availability hypothesis indicating that users tend to stay within the same category and prefer campaigns from the same employer. For both hypotheses we initially set parameters so that moving to the same category or employer is two times as likely as moving to a campaign from another category or employer ($\alpha = 1, \beta = 0.5$). We experimented with different parameter values and found that further favoring the same categories/employers increases the plausibility of the hypothesis up to a certain point. In Figure 2(b) we show the best parameter settings we have found. The results indicate that users strongly favor the same categories and employers. We further note, that staying with the same employer is a more plausible hypothesis, i.e., because we find greater evidence for it and also because a stronger focus on the same employer can be set ($\beta = 0.01$) before the evidence is decreasing again. This is in line with the findings of Schulze et al. [10], who report a tendency to choose campaigns from reputable employers.

However, as mentioned in Section 3.3, the overall distribution of campaigns within categories and from employers is not equally distributed. Thus, we defined hypotheses based solely on employer and category frequencies in Section 4.2 to contrast the category and employer hypotheses evaluated above. We find that, even though both hypotheses are more plausible than the availability baseline, both hypotheses perform worse than the hypotheses favoring the same campaigns or employers.

Since both, hypotheses based on categories and hypotheses based on employers, perform well, we investigate if a combination of both, as mentioned in Section 4.2, can further improve our results. Indeed, we find that strongly favoring campaigns from the same category and the same employer, and setting the weights for choosing only the same campaign or the same employer or a totally different campaign rather low ($\alpha = 1$, $\beta = 0.025$, $\gamma = 0.0046825$), results in the most plausible hypothesis so far. However, these combinations are only marginally more plausible than the hypotheses exploiting only one factor. We have also tried setting different parameters for only favoring one of the two, category or employer, also only resulting in marginal improvements.

Overall, we can conclude that the idea that users choose campaigns from the same category and the same employer can indeed explain parts of the campaign transitions we observe. Particularly, the employer is an important factor.

5.3 Payment, position and time

As introduced in Section 4.3 we are also considering several hypotheses based on payment, available positions per campaign and the time required to complete a task. To this end we directly use the values of payment, payment per hour (pph), positions and the inverse of the required time. The results are shown in Figure 2(c). All of these hypotheses perform badly compared to the overlap hypothesis. This is due to their skewed value distributions which have a strong tendency towards small value ranges, cf. Section 3.3. Thus, we introduced three stratified classes for payment, pph and required time in order to model a tendency to pick low, average or high prices. We weight these classes as follows: $low = 1$, $average = 2$ and $high = 3$. We observe, that all the resulting hypotheses, even if marginal, have a greater evidence than the uniform hypothesis on overlapping availabilities (while the payment classes are hardly distinguishable in the graph, evidence values actually do significantly differ from those of the overlapping availability hypothesis). Thus, confirming the findings of [3, 10, 12], we can conclude that all of these factors play a role in choosing campaigns. However, we were not able to derive a hypothesis based on required time explaining the influence factor found by Schulze et al. [10].

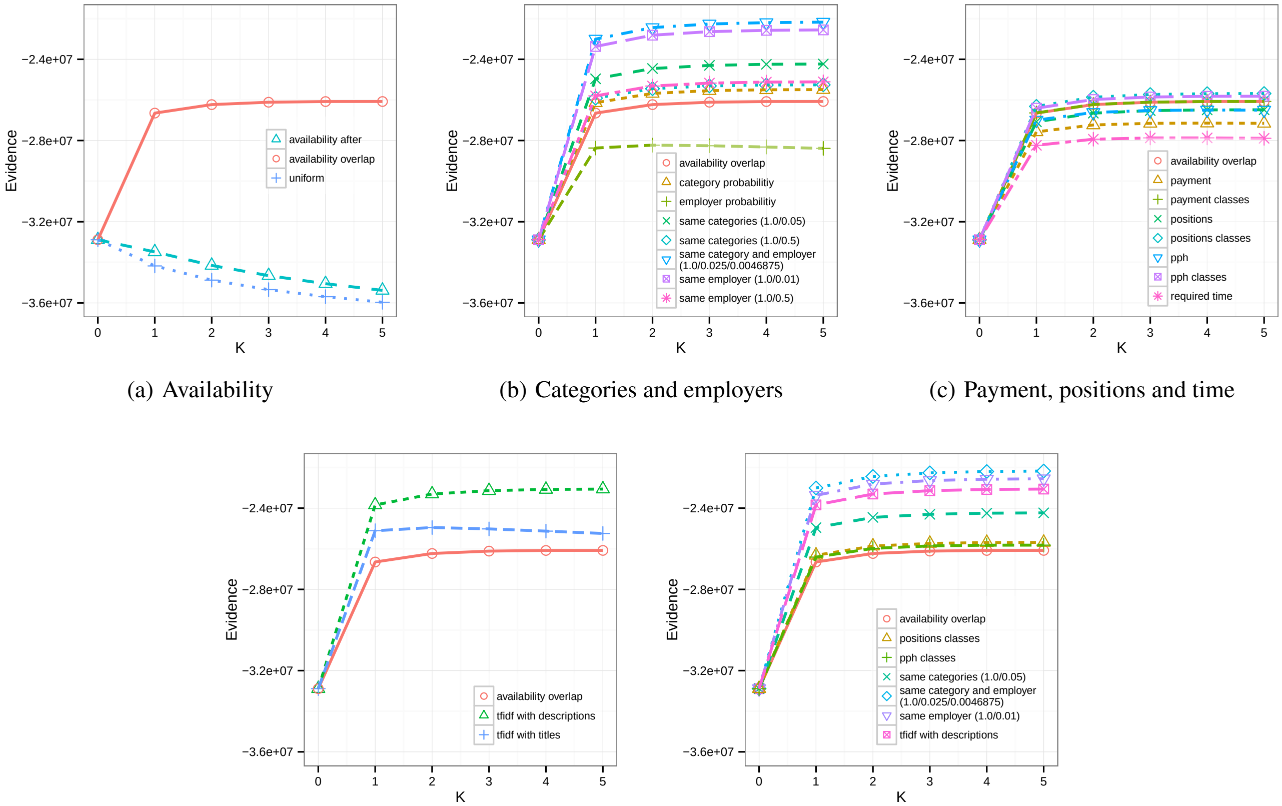
5.4 Title and description

As introduced in Section 4.4, we also compare similarities between the title and the descriptions of the campaigns. The results are shown in Figure 2(d). We can clearly see that both hypotheses perform better than the hypothesis assuming a uniform distribution over all overlapping campaigns. This is a strong indicator that both, title and description, and thus the semantic content of the task to complete, are a decisive factor for choosing campaigns. Note though, that the similarity based on title clearly performs worse than the similarity based on descriptions. This can be interpreted by the fact that while a certain type of campaign will have the same description, the title may differ. For example, consider a campaign whose goal is to annotate a certain corpus of documents. While the description might be the same, the title for one campaign might be “Annotating Literature”, while the other might be “Annotate 10 Research Papers”. The description based hypothesis captures the similarity of those two campaigns, while the title based hypothesis does not.

5.5 Summary

For an overall comparison we show the best hypotheses of each category in Figure 2(e). We can see that three hypotheses yield especially high evidence values, that is, the mixture of category and employer, followed by the hypothesis solely based on employers and the hypothesis based on description similarities. The hypotheses based on payment and positions hardly improve on the uniform overlap hypothesis. The hypothesis based solely on categories performs worse than the description based hypothesis but better than the hypotheses based on payment and positions.

The fact that the category hypothesis performs significantly worse than the employer hypothesis and that combining the employer hy-



(a) Availability

(b) Categories and employers

(c) Payment, positions and time

(d) Title and descriptions

(e) Comparison

Figure 2: Experiments. This figure visualizes the results for our studies on how users choose their tasks on the crowdsourcing platform Microworkers. First, we compare different task availability models in (a). Then, in (b-d) we show the results for different categories of hypotheses. (b) compares different campaign category and employer hypotheses, (c) focuses on hypotheses based on payment, available tasks per campaign (positions) as well as the time required to finish a task, and (d) depicts evidences for description and title based hypotheses. In (e) we compare the best hypotheses of each category. We find that staying with the same employer is a strong influencing factor when choosing campaigns directly, followed by the description based hypothesis.

pothesis with categories only marginally improves the evidence, is an indicator that users primarily choose campaigns from the same employer instead of focusing on categories. Yet when they have chosen their employer they prefer to stay within the same campaign category. Overall, the large evidence values for employer based hypotheses are in line with results found by Schulze et al. [10]. They imply that workers are loyal to reputable employers when choosing campaigns: since employers rate the completed tasks by their quality, and only good ratings result in money being payed, workers prefer employers who rate fairly.

Furthermore, the good performance of the description based hypothesis is an indicator that users not only choose from within the same campaign, but also try to choose similar campaigns with regard to the actual task they have to work on. One reason for this might be that familiar tasks are easier and more quickly to handle than unknown ones. This may also be a result of users choosing similar tasks based on their area of interest as suggested by Aris [1]. Note, that the description hypothesis might strongly overlap with the employer hypothesis since the same employer may often use the same description for her campaigns of a similar type. This needs to be further investigated.

Finally, the result that the payment and positions hypotheses perform badly, seems counter-intuitive. This might be due to the fact that users optimize for certain types of tasks and can earn more money when they stay at the same employer and within the same category accounting for the high plausibility of the corresponding hypotheses. Also, Aris [1] finds that payment in general is not a consistent factor influencing how users choose their tasks. Schulze et al. [10] and Schnitzer et al. [9] even find that at least US workers (which we have focused on here) are not mainly interested in the amount of money they earn for completing a campaign. However, the bad performance may also be due to the way we model the corresponding hypotheses. For example, we have noticed that directly incorporating the payment value into the hypothesis does not perform well (see Section 5.3). Now, while the stratified classes reveal a certain tendency to choose well payed campaigns or campaigns with many positions, the resulting increase in evidence is not as large as could be expected. Thus, we might not capture the influence of payments or positions correctly. That is, different classes or different weighting strategies might result in better hypotheses. Also, the payment may strongly interact with other factors like the tendency to choose similar tasks as mentioned before. Regarding these issues, we will propose ideas for further research in the discussion section.

Overall, we studied a considerable amount of hypotheses partially motivated by related work using statistical methods solely based on data from the crowdsourcing platform Microworkers and without resorting to error-prone and possibly biased user studies. In the process, we focused on US workers and were able to show from observed task transition data that most factors found in literature indeed influence the process of how workers choose campaigns. Employer and description based hypotheses worked best. Whereas hypotheses based on payment only showed a marginal influence on how users choose their tasks.

6. DISCUSSION

We have tested several hypotheses about how users choose their campaigns on the crowdsourcing platform Microworkers. In this section, we give a short overview of particular limitations of our approach and propose possible future work.

Data First of all, the data set we are using is limited to workers from the US. This is because users from the US are free to choose nearly all campaigns. For non-US workers many campaigns are not available. Thus, incorporating non-US workers would introduce restrictions on transitions which can not be directly modelled by our hypotheses. In further studies it might be useful to actively incorporate a user component indirectly into hypotheses or directly into an extended version of HypTrails.

Furthermore, we are evaluating our hypotheses on only one data set. It would be interesting to check if the hypotheses behave similarly on different crowdsourcing platforms. Additionally, for example, Schulze et al. [10] and Schnitzer et al. [9] suggest that there may be strong differences between certain user groups, e.g., from different countries. Further studies on corresponding data sets might be an interesting line of research.

Hypotheses While we have studied quite a few different hypotheses, there are more to consider. For example, it might be interesting to study if users tend to prefer recently created campaigns as suggested by Chilton et al. [3]. Also, Aris [1] implies that intrinsic factors are more important than extrinsic ones. In this work we have manly focused on extrinsic ones.

Furthermore, we have only combined the category with the employer hypothesis. Other combinations might yield better results. Also, we have not checked how the hypotheses are related to each other in a sense that the features used to build them are correlated. An example would be that the same employer will often use similar descriptions for her campaigns. Thus, as mentioned in Section 5.5, the description hypothesis might be correlated to the employer hypothesis. This needs further investigation.

Finally, the payment, positions and time related hypotheses did not yield good results when compared to category or employer based hypotheses. While there are explanations in literature [1, 10], this may also be due to a poor understanding of how these factors influence the choice of campaigns. We have approximated the influence using three stratified classes. Other approaches might be more appropriate.

Availability One limitation of HypTrails is that it is not built to model states that are only available at certain time intervals. We solved this approximately by introducing the notion of availability. In our scenario we used local availability (from a specific campaign to other campaigns) based on time intervals. However this is an approximation. Further research may find a better solution in corporate such time dependent availability into HypTrails and/or in the process of formulating hypotheses.

7. CONCLUSION

In this paper, we have studied how users choose their next task on the crowdsourcing platform Microworkers. To this end, we have formulated different hypotheses about the underlying processes based on properties like the similarity of campaign descriptions, categories, employers or payment information. Then, utilizing campaign transition data from Microworkers, we objectively compared the resulting hypotheses by means of the Bayesian approach HypTrails. While the results highly depend on how hypotheses are formulated, in our scenario, combinations of category and employer as well as the description based hypothesis work best. Overall, instead of using survey based investigation as similar studies do, we successfully apply the Bayesian method HypTrails to objectively compare hypotheses about how users choose their next campaign solely on data already available from the crowdsourcing platform. This is a step forward in providing crowdsourcing companies with the means to gauge the preferences and the behavior of their users in order to optimize their platforms, e.g., by incorparting the gained knowledge about these factors into task recommendation systems.

Acknowledgments

The research leading to these results received funding from the Deutsche Forschungsgemeinschaft (DFG) under Grants HO 2586/2-2, HO 4770/2-1, TR257/38-1.

8. REFERENCES

[1] H. Aris. Influencing factors in mobile crowdsourcing participation: A review of empirical studies. In *Conference on Computer Science and Computational Modelling*, 2014.

[2] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[3] L. B. Chilton, J. J. Horton, R. C. Miller, and S. Azenkot. Task search in a human computation market. In *Workshop on Human Computation*, 2010.

[4] T. Hoßfeld, M. Hirth, and P. Tran-Gia. Modeling of crowdsourcing platforms and granularity of work organization in future internet. In *International Teletraffic Congress*, 2011.

[5] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

[6] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[7] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *Americas Conference on Information Systems*, 2011.

[8] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton. The future of crowd work. In *Conference on Computer Supported Cooperative Work*, 2013.

[9] S. Schnitzer, C. Rensing, S. Schmidt, K. Borchert, M. Hirth, and P. Tran-Gia. Demands on Task Recommendation in Crowdsourcing Platforms - The Worker's Perspective. In *CrowdRec Workshop*, 2015.

[10] T. Schulze, S. Seedorf, D. Geiger, N. Kaufmann, and M. Schader. Exploring task properties in crowdsourcing-an empirical study on mechanical turk. In *European Conference on Information Systems*, 2011.

[11] P. Singer, D. Helic, A. Hotho, and M. Strohmaier. Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *Conference on World Wide Web*, 2015.

[12] M.-C. Yuen, I. King, and K.-S. Leung. Task recommendation in crowdsourcing systems. In *Workshop on Crowdsourcing and Data Mining*, 2012.