Bulletin of the Technical Committee on

Data Engineering

December 2019 Vol. 42 No. 4

IEEE Computer Society

Letters

Opinions

Special Issue on Imagine All the People and AI in the Future of Work

Conference and Journal Notices

The TC on Data Engineering

Membership in the TC on Data Engineering is open to all current members of the IEEE Computer Society who are interested in database systems. The TCDE web page is http://tab.computer.org/tcde/index.html.

The Data Engineering Bulletin

The Bulletin of the Technical Committee on Data Engineering is published quarterly and is distributed to all TC members. Its scope includes the design, implementation, modelling, theory and application of database systems and their technology.

Letters, conference information, and news should be sent to the Editor-in-Chief. Papers for each issue are solicited by and should be sent to the Associate Editor responsible for the issue.

Opinions expressed in contributions are those of the authors and do not necessarily reflect the positions of the TC on Data Engineering, the IEEE Computer Society, or the authors' organizations.

The Data Engineering Bulletin web site is at http://tab.computer.org/tcde/bull_about.html.

Letter from the Editor-in-Chief

Machine learning and AI will play a critical role in our data-driven future. This explains the surge of interest in many machine learning and AI related courses, conferences, and publications. The question here is, in our data-driven future, what is the role of data management?

Michael Stonebreaker and El Kindi Rezig tried answered this question in their opinion piece "Machine Learning and Big Data: What is Important?" Drawing from their experience working with big data in real-life medical and business applications, they highlighted the importance of data preparation, including data integration, data enrichment, and data augmentation (for training). They argued that an ML data management system that is capable of tracking and optimizing the interactive steps of data preparation and machine learning is long overdue.

Lei Chen et al put together an interesting issue consisting of 5 papers on the topic of "Future of Work." They envisioned a future where AI and machines automate most of the work, but humans are in the loop to provide intellectual self-actualization. The papers looked into today's crowdsourcing platforms and investigate problems such as using active learning to make data labeling more efficient and effective. In addition, the issue highlighted ethical challenges in the future of work by using examples of today's self-employment platforms such as Uber and Amazon Mechanical Turk.

Haixun Wang
WeWork Corporation

Letter from the Special Issue Editor

We envision the Future of Work (FoW) to be a place where humans are empowered with the ability to rely on AI machines in an on-demand fashion, with ability to juggle diverse job opportunities that provide intellectual self-actualization, and enhance their capabilities by continuous knowledge and skill acquisition with a variety of onboarding and training tools. We recognize that the current view of "humans-in-the-loop" tends to see humans as machines, robots, or low-level agents used or exploited in the service of broader AI goals. Our hope is that people in their workplace in the future should be treated as fully human, with respect and dignity with the right to productive employment and the goal of bringing them back in the "frontier" of "humans-in-the-loop" systems. Such an environment will allow everyone everywhere to get a job online and offline, train for a new job, and get help from a mix of humans and AI machines. Ensuring portability between FoW marketplaces, and guaranteeing the protection of workers' rights, will play a major role in providing a rewarding and safe work environment to all. Existing platforms must rethink their design to empower humans and be at the frontier of FoW.

Today, humans' relationship to work is changing as online job platforms are blurring the boundaries between physical and virtual workplaces. That is witnessed in freelancing platforms such as CrowdWorks in Japan, TaskRabbit and Fiverr in the USA, and Qapa and MisterTemp in France, crowdsourcing platforms such as Crowd4U in Japan, gMission in Hong Kong, Wirk, and Prolific Academic in Europe, Amazon Mechanical Turk and Figure Eight in the USA, and platforms that provide help with entrepreneurship such as de Asra in India. Prospective employees can find temporary jobs in the physical world (e.g., a plumber, an event organizer, a gardener, can offer their jobs online), or in the form of virtual gigs (e.g., logo design, web programmer). Job providers can hire one or many individuals to achieve a task. The same person can take on those roles at any point in time. An employer can be a regular citizen who needs to hire a plumber, a social scientist needed to conduct population studies to verify some theories, a data scientist needing to validate a new algorithm, a domain expert seeking to verify how much interest a new product generates. The diversity of needs has given rise to a variety of platforms, all of which act as intermediaries between job providers and job seekers. Platforms differ in their ability to manage physical and virtual jobs, in their support for onboarding, socializing, training, and credentialing for employees, in automating the matching between jobs and workers. They also differ in the tools they offer to workers and job providers to express needs and requirements, and in their compliance with labor-related regulations and their handling of ethical concerns.

In this issue, to follow the trend that FoW will be increasingly technology-driven and will have the potential to bring human concerns to the center of the design and deployment of job platforms, we discuss the raised intellectual challenges.

The first paper addresses the challenge related to the ability to understand different types of human roles in future jobs, modeling the inherent uncertainty in human behavior by understanding their evolving characteristics and be able to propose jobs to them by adapting to their changing perceptions and needs.

The second paper addresses the challenge in the development of appropriate interaction methodologies and support for the new kind of workplace. This includes social processes relating workers, requesters and platform managers like onboarding, socializing, recognition and training, as well as ways to communicate and delegate work between humans and machines.

The third paper addresses the challenge on how to design platforms that maximize the satisfaction of various stakeholders.

The fourth paper addresses the challenge on design benchmark and metrics to measure computing capabilities as well as human aspects such as satisfaction, human capital advancement, and equity.

The last paper discusses various of ethics issues raised in FoW design and architecture, including privacy, compensation mechanisms and fairness.

We would like to thank all the authors for their insightful contributions.

Sihem Amer-Yahia, Lei Chen, Atsuyuki Morishima, Senjuti asu Roy

Machine Learning and Big Data: What is Important?

Michael Stonebraker and El Kindi Rezig
Massachusetts Institute of Technology

1 Introduction

At MIT, we have been collaborating on two real-world projects dealing with Machine Leaning (ML) and large amounts of data. The first project deals with performing ML on a 30T data set of EEG (electroencephalogram) sleep study data, in collaboration with Massachusetts General Hospital (MGH). They have collected EEG data on about 2000 patients with time durations varying from a couple of hours to more than 24 hours. In all, they have recorded 21 channels of data, with a total volume of about 30T. Furthermore, they have decomposed this data into about 450M segments, each 2 seconds long. Their project goal is to use ML to classify each 2-second segment "snippet" into one of 12 categories that make sense to doctors. Two papers on this project were presented at the recent VLDB conference [1, 6].

The second project is also in conjunction with MGH and deals with the spread of infections in the hospital. MGH has made 11 years of patient data available, including exactly which room each patient occupied from admission to discharge. For infected patients, they wish to infer how an infection was spread, i.e. by sharing a nurse, by sequentially inhabiting the same room, etc. Because MGH changed patient software in 2016, we first have a data integration problem, which we are addressing using MIT-built ML integration tools [1, 7]. Then, we propose to infer infection pathways from their integrated data.

Finally, one of us has been a principal at an ML data integration company (Tamr) that has done more than 300 ML projects primarily for Fortune 2000 customers. Tamr is in the business of data integration at scale. Typically their projects deal with integrating several-to-many data sets by normalizing the data to a common format, correcting data errors, merging and deduplicating the result, and creating "golden records" for each cluster of records thought to represent the same entity. Typically, input data sets represent parts, suppliers, customers or other entities, and this end-to-end process is usually called "mastering" for a particular entity. The guts of the Tamr system is a collection of ML algorithms to perform schema matching, deduplication and golden record construction.

This paper summarizes our thoughts, based on our experience with these use cases. The remainder of this paper is divided into sections containing our observations.

2 At Scale is Where all the Hard Problems Reside

For example, Toyota Motor Europe (TME) is a Tamr customer. Historically they had country-specific distribution and a country-specific customer database. When a Toyota customer moved countries (say from Spain to France), TME developed amnesia. To correct this situation TME is using Tamr to "master" all European customers. There are 30+ million customer records in 250 databases in 40 different languages. At this scale any N^2 deduplication algorithm is a non-starter, and Tamr has spent a lot of effort on cheaper computations. Obviously, one also needs a parallel multi-core, multi-node execution environment.

Similarly, Glaxo Smith Kline (GSK) is proposing to perform mastering on research data sets often of pharmaceutical compounds. They have many, many data sets with 10+M attributes. Traditional mastering

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

products perform attribute matching by manually drawing lines in a GUI between attributes which represent the same thing. A human obviously cannot draw 10+M lines, and a different (automatic) approach is needed.

In our opinion, if your data set is 1000 records, then any algorithm running "on your wrist watch" will work fine. At scale is where all the demons reside, and traditional solutions fail.

3 Deep Learning is Not for Everyone

So called "deep learning" based on neural networks is the current rage in the research community. However, deep learning is not appropriate for all problems. For example, Tamr customers have very little interest in this technology for two very simple reasons. First, deep learning requires 1-2 orders of magnitude more training data than conventional ML. With structured data in the enterprise, training data required to fit an ML model is always "the high pole in the tent", i.e. there is never enough of it, and getting more is very expensive.

At scale, it is hopeless to obtain training data by asking a human whether given pairs of records are a match or a non-match. That will work fine for 1000 pairs but not 100K pairs. The focus has to be on generating training data using rules or other "weak" methods [8]. For example, if one is mastering enterprise customers, then a rule might specify that two customers with an edit distance between their names below a threshold are probably the same customer. This rule will generate a significant amount of training data, but there will be errors and some sort of validation must be employed. This validation requires skilled enterprise personnel and cannot be done with unskilled contract workers using, say, Mechanical Turk.

For example, the rule in the paragraph above will say that two records for Merck, one with an address in Europe and one in the US, are the same entity. Is this decision correct? Deciding this question can realistically only be entrusted to financial personnel in your enterprise, and these employees are both expensive and busy.

Hence, constructing and validating training data is the most expensive part of the ML process. Any strategy that requires a lot more of it is a non-starter. For the record, German Merck is a different company than US Merck.

The second problem with deep learning is that it is currently an unexplainable "black box". Many Tamr customers want explanations for why a given outcome was selected. If a company is using ML for classifying customers for obtaining credit or other financial reasons, then explanations are necessary. Currently deep learning decisions cannot be explained, and this is an area of active research. Until there is some maturity in this area, conventional ML will be the tool of choice.

In summary, issues with training data and explainability make Tamr customers prefer conventional ML. Perhaps this will change at some point in the future, but for now we expect conventional ML (decision trees, Bayesian methods, etc.) will be the norm for structured enterprise data.

4 Supporting ML at Scale is a Data Management Issue

Obtaining enough training data for many deep learning applications is very expensive. For instance, in our MGH collaboration [6], we built a neural network that required labeling 30M EEG segments for training purposes. Labelling this amount of data manually is infeasible. Hence, the experts (MGH physicians) labeled 8346 EEG segments manually, and then we used clustering techniques to find similar EEG segments from the unlabeled data. This has generated around 37 million EEG segments of training data. However, automatic clustering is error-prone and the training data is also somewhat dirty. Obviously, a good tactic is to iteratively clean the data, train the ML model, and run it, until the results are "good enough". In our demo [1], we showed that we could boost the model's accuracy by 7% just by applying automatic data cleaning tools on the training data.

As a result, improving an ML model can occur in at least three different ways: (1) obtaining more training data; (2) parameter-tuning of the model; and (3) cleaning the source data.

Obviously, one must iterate between these three tactics. In addition, there is a fourth option: data enrichment. The Tamr engineers are often given ML problems where there is simply not enough signal to perform deduplication. Sometimes only the product name is available as a shared attribute, and there is a lot of noise associated with the various ways this gets recorded. Another example is a corporate customer name along with an address. For companies like Staples with many locations, the address has only a small signal value.

In these cases a good tactic is to perform signal enrichment by joining the source tables to other data sources. For example, a successful join of customer name to the Dow Jones Intelligent Identifier (DJII) or the Dow Jones International Securities Number (ISIN) may generate an additional attribute with signal value. Likewise, a join of product name to the Universal Product Code (UPC) may prove helpful. Even if the join can only be performed on a subset of the data, the data network effect (i.e., the value of the added data will benefit all the entities linked or related to it) will cause this to broadly improve results.

Note clearly that enrichment is mostly a data discovery problem, which data management researchers should make a priority research area [4, 5].

Hence, successful ML projects may run hundreds of iterations before they obtain a successful model. Obviously, a good support system is needed to perform model management. Specifically, for each iteration we need to keep track of: the training data, the cleaned data, the state of cleaning, the parameters of the ML model being run, the model's features, the intermediate data, the input and the output data.

One needs a model management system (MMS) to keep track of the "state" noted above, to ensure that everything done during model development can be clearly explained and repeated in production. This is a straight data management problem, and initial work in this area is reported in [9]. A great deal more work is needed in this area, and readers should consider this section as a "call to arms". Moreover, there must be a "human in the loop" to decide what to do next and to back up from failed options when necessary. Optimizing ML platforms for such iteration seems long overdue.

5 ML is not really about ML

A year ago, one of us attended a talk by an iRobot data scientist. (iRobot makes the automatic vacuum cleaners that run around your living room). I asked her how she spent her work week, and here is her answer. "I spend 90% of my time finding data of interest, normalizing it and cleaning it", i.e. data discovery and data integration. "Then, I spend 90% of the remainder fixing my data cleaning errors." In other words, she spends one hour a week or less on data science and 39 hours a week on "data prep". We have talked with many data scientists, and all report similar behavior. We have never met anybody who claimed the answer was less than 80%.

Hence, data scientists are drowning in data prep. Put differently, if you want to help a data scientist do his/her job better, then work on data discovery, data cleaning and data integration. There are many efforts in this area, but much more needs to be done, since this is far and away "the high pole in the tent".

Our favorite complaint is that research institutions appear to spend 90% of their ML effort on algorithms and 10% on data preparation. In our opinion, it should be 90/10 the other way. As real world applications of ML expand, we expect the feedback from enterprises to confirm this. A three star general said exactly this in a recent briefing.

6 ML System Usability

Another issue that often seems overlooked is the usability of ML/data pipeline management systems. Most organizations have their own in-house scripts that perform data preparation and ML. Unfortunately, to benefit from their features, most ML frameworks require users to modify their code to adapt it to those frameworks. We believe this requirement should be reversed: we should strive to build data systems that adapt to users' tools/code and not the other way around. Average organizations often do not have the resources or the manpower to adapt

their tools to specific data frameworks, so they often resort to using their in-house scripts in an ad-hoc fashion. Obviously, frameworks need to become easier for "mere mortals" to use.

Human-in-the-loop data systems have been around for decades. However, we are still a long way from end-to-end data systems that put the human at the forefront [2]. For example, there are few efforts that explore the cognitive cost when recommending a task to a human, i.e., how hard is task X vs. Y. For instance, it is easier to ask a human if two company names refer to the same organization than asking him/her if one company is the parent of another (which requires researching the companies). However, one high-effort answer could lead to greater gains than asking dozens of low-effort questions. This area should be explored to find the best strategies to involve humans and optimize their overall cost.

Lastly, human feedback has been incorporated to solve several point problems in data pipelines; however, it is still unclear how humans should be involved in end-to-end data systems, e.g., from data preparation all the way to analytics. There are many opportunities to incorporate the human feedback at different points in the data processing pipeline, and more research in this area is warranted.

7 Real World Studies Often Deal with Application Specific Data

One of the applications areas where Tamr is used is "oil well mastering". In other words, oil producers like Hess or Exxon have data for wells in various locations. In general, wells are identified by coordinates, but the accuracy of the coordinates as well as the coordinate reference system may have significant uncertainty. As a result, the identifying key for oil wells is imprecise. Therefore, "oil well mastering" requires approximate spatial clustering, to identify records that correspond to the same well. ML in the real world must deal with these kinds of extended types. Many years ago DBMSs began to support type extension, and ML systems need similar techniques. Again, this is a fertile area for extension of ML capabilities.

8 ML/Big Data Papers Should be Required to Use Real Data

ML is sometimes applied to data integration, and often the experimental section justifies the value of the paper's algorithms on "fake" data. In other words, good data is algorithmically perturbed to create incorrect data. It is not surprising that the authors can remove errors that they inserted into a data set!!!

As a community, we should simply reject any paper which uses fake data in its experimental section. Researchers usually argue that real data is not available to them. In our opinion this is a cop out for "I am lazy", and don't want to do the necessary leg work.

9 Debugging ML

ML frameworks should strive to make it easy for users to find out what went wrong with models and to mitigate potential problems. We are still a long way from a full-fledged system that allows users to fully understand what the ML model did and did not do and why. In other words, a full function data debugger would be incredibly valuable, and a first step in this direction appears in [11].

More generally, we need a lot of effort to democratize ML for non-ML-expert users, and to be able to use ML responsibly. It is still very easy to produce garbage results from ML systems! Additionally, humans are not always the best source of training data, as human bias is often transferred to ML models. This will obviously produce biased results. Fairness is emerging as a critical area of research to make sure ML systems are not rigged to unfairly favor one decision over another [3].

10 Conclusions

This paper has presented several areas where we believe current efforts are missing the mark. There should be more efforts devoted to conventional ML a lot more effort to supporting human-driven iteration of an ML pipeline, and a lot more effort on data preparation. Moreover, at scale is where all the demons reside. Data debugging and ML system usability should be research foci. Obviously, researchers in this area should partner with a real-world use case. Otherwise, there is no guarantee that "the rubber will meet the road instead of the sky".

References

[1] El Kindi Rezig, Lei Cao, Michael Stonebraker, Giovanni Simonini, Wenbo Tao, Samuel Madden, Mourad Ouzzani, Nan Tang, Ahmed K. Elmagarmid: Data Civilizer 2.0: A Holistic Framework for Data Preparation and Analytics. PVLDB 12(12): 1954-1957 (2019)

[2] El Kindi Rezig, Mourad Ouzzani, Ahmed K. Elmagarmid, Walid G. Aref, Michael Stonebraker: Towards an End-to-End Human-Centric Data Cleaning Framework. HILDA@SIGMOD 2019: 1:1-1:7.

[3] Babak Salimi, Luke Rodriguez, Bill Howe, Dan Suciu: Interventional Fairness: Causal Database Repair for Algorithmic Fairness. SIGMOD Conference 2019: 793-810

[4] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, Michael Stonebraker: Aurum: A Data Discovery System. ICDE 2018: 1001-1012

[5] Erkang Zhu, Dong Deng, Fatemeh Nargesian, Renée J. Miller: JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. SIGMOD Conference 2019: 847-864

[6] Lei Cao, Wenbo Tao, Sungtae An, Jing Jin, Yizhou Yan, Xiaoyu Liu, Wendong Ge, Adam Sah, Leilani Battle, Jimeng Sun, Remco Chang, M. Brandon Westover, Samuel Madden, Michael Stonebraker: Smile: A System to Support Machine Learning on EEG Data at Scale. PVLDB 12(12): 2230-2241 (2019)

[7] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Nan Tang: The Data Civilizer System. CIDR 2017

[8] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, Christopher Ré: Snorkel: Rapid Training Data Creation with Weak Supervision. PVLDB 11(3): 269-282 (2017)

[9] Manasi Vartak, Joana M. F. da Trindade, Samuel Madden, Matei Zaharia: MISTIQUE: A System to Store and Query Model Intermediates for Model Diagnosis. SIGMOD Conference 2018: 1285-1300

[10] Tim Kraska:Northstar: An Interactive Data Science System. PVLDB 11(12): 2150-2164 (2018)

[11] El Kindi Rezig, Lei Cao, Giovanni Simonini, Maxime Schoemans, Samuel Madden; Nan Tang; Mourad Ouzzani; Michael Stonebraker: Dagger: A Data (not code) Debugger. CIDR (2020)

Capturing Human Factors to Optimize Crowdsourced Label Acquisition through Active Learning

Senjuti Basu Roy
New Jersey Institute of Technology
senjutib@njit.edu

Abstract

The goal of this article is to propose an optimization framework by acknowledging human factors to enable label acquisition through active learning . *In particular, we are interested to investigate tasks, such as, providing (collecting or acquiring) and validating labels, or comparing data using active learning techniques. Our basic approach is to take a set of existing* active learning techniques for a few well known supervised and unsupervised algorithms, but study them in the context of crowdsourcing, especially considering worker-centric optimization (i,e., human factors). *Our innovation lies in designing optimization functions that appropriately capture these two fundamental yet complementary facets, performing systematic investigation to understand the complexity of such optimization problems, and designing efficient solutions with theoretical guarantees.*

1 Introduction

Human workers or crowd can be used as a building block in data-intensive applications, especially for acquiring and validating data. One such fundamental crowdsourcing task is *labeling*, where a human worker is asked to provide one or multiple labels for the underlying observation. A plethora of applications in cyber human system directly benefit from such efforts, where the objective is to design automated classification/prediction algorithms but require *labeled data* for that. Even though human intelligence provides substantial benefit to computation, incorporating humans in the computational loop incurs *additional burden* - it becomes time-consuming, monetarily expensive, or both in such cyber-human systems.

Active learning principles [34, 36] are proposed to optimize system-centric criteria in classification problems, by employing human workers judiciously only for a few tasks. When crowd is involved in data acquisition, additional challenges emerge: (1) contribution from crowd is potentially noisy, (ii) to ensure higher engagement and productivity, one has to understand *worker-centric criteria* [39], such as, *worker skill, motivation*, that are referred to as *human factors* in the literature [2, 48, 31, 8]. In a hybrid human-machine computational environment, an opportunity exists in *laying a scientific foundation for predictive analytics* that combines system-centric optimization derived from active learning [34, 36] principles and worker-centric optimization through human factors modeling.

Imagine that a computational ecologist wants to design a binary classifier [15] that accurately predicts the presence or the absence of species given environmental covariates (such as geographical coordinates, elevation, soil type, average precipitation, etc). In order to learn the classifier, the ecologist needs annotated data (i.e., identify the presence or absence of a species in a given location). Indeed, as shown in Figure 1, there exists

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

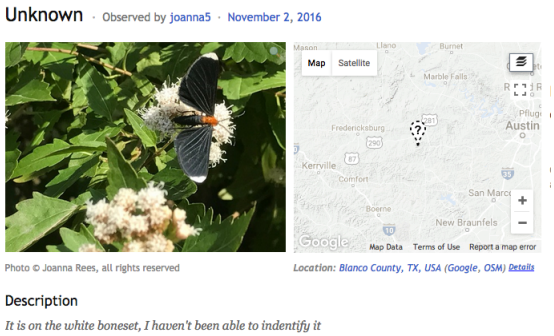

Figure 1: Unidentified Species at iNaturalist Website

unidentified species at large-scale citizen science based platforms like iNaturalist that are to be crowdsourced to be labeled by their registered workers. While doing that, we however need to find the most "suitable" worker taking into account worker-centric criteria, such as, her geographic distance from the observation location, her expertise and/or motivation to identify insects, by considering her past activities/tasks. The goal is to focus on applications, where the workers (e.g., registered citizen scientists of iNaturalist) are involved in completing tasks that are repetitive in characteristics and do not evolve much over the time. We propose to develop an *optimized human-machine intelligence framework for such cyber-human systems for single and multi-label classification problems [27, 55] through active learning*.

Our effort will be investigate and adapt existing *active learning techniques for a few well known supervised algorithms for single-label and multi-label classification in the context of crowdsourcing, especially considering worker-centric optimization through human factors modeling*. The general idea behind active learning is to select the instances that are likely to be most informative. Then the selected instances are annotated by human, and the computation loop is repeated. Our innovations lie in appropriately capturing these two fundamental yet complementary facets in a single optimization function, performing systematic investigation to these problems and designing innovative solutions.

Example 1: Motivating Example: The iNaturalist platform contains photo vouchered independently verified species occurrence records by the citizen scientists across the world and is one of the fastest growing source of contributions to the global biodiversity information facility. We focus on such platforms, where tasks are repetitive in nature, do not change over the time, and one can retain all the past tasks undertaken by the registered workers (citizen scientists). A computational ecologist makes use of such platforms to develop species distribution models using environmental covariates. Thus, she aims to design a crowdsourcing effort to judiciously obtain single and/or multiple label(s) to annotate some of the unidentified images (refer to Figure 1). A single label acquisition is about identifying the insect (which will then be augmented with the location information), whereas, multiple labels will require identifying the Kingdom, Phylum, Class, Order, Sub-Order, Family, Sub-family of the insect in the image.

For these examples, label acquisition involves human workers. On the other hand, domain expert may not have many human workers at her disposal who are qualified for the task - even if there are workers, they may not be *motivated* to undertake these tasks. Furthermore, workers may even have constraints (e.g., only likes to watch birds, can not travel more than 25 miles from her location). Therefore, which task(s) are to be selected and assigned to which worker(s) remain to be a problem.

Desirable properties: To generalize, the above scenarios call out for the following desirable characteristics: (1) An "activized" label acquisition is desirable - i.e., acquire more data only when it optimizes the underlying computational task considering system-centric criteria. (2) At the same time, select workers and assign tasks to enable *worker-centric optimization*. (3) We argue the necessity of studying these two facets together as a single

optimization problem, as a staged solution (first select sub-tasks based on active learning, then assign those to the workers to enable worker-centric optimization) may be inadequate, as tasks selected by active learning techniques may end up having a very low worker-centric optimization, resulting in poor outcome overall.

High level approach: (1) We propose *worker-centric optimization by characterizing human factors* in the crowdsourcing context by adapting well-known theories of behavioral psychology [48, 31, 8]. (2) We propose system-centric optimization by adapting a set of well-known *active learning techniques* for supervised algorithms [34, 36, 51, 52, 10, 21, 17, 23, 55, 24, 35, 61] and augment them by combining worker-centric optimization through human factors modeling. (3) We propose systematic investigations on how the two *aforementioned optimization problems could be combined* and propose effective solutions.

Additionally, we will design both retrospective as well as prospective studies. We will perform these evaluations using publicly available citizen science datasets.

Novelties: To the best of our knowledge, no existing work has studied what we propose. The closest related works for active learning through crowdsourcing are for single label acquisition [64, 62, 41, 16, 18]. Worker-centric optimization is not considered there. Active learning research for multi-labels remains in a rather nascent state [55, 24, 35, 61]. These aforementioned studies do not investigate the problem in the context of crowdsourcing. Therefore, the necessity of worker-centric optimization does not even arise there. Our designed prototypes on iNaturslist platform will bear a long-lasting effect to understand global bio-diversity.

1.1 Research Objectives

Our long term vision is to optimize knowledge discovery processes for cyber-human systems. We are interested in designing effective solutions and support both worker and system criteria through active learning and human factors modeling. The research proposed here puts us on track to achieving this vision by addressing a first series of concrete challenges on a very important application domain.

(1) Optimized *single-label acquisition* involving crowd: In this research aim, we strive to propose optimization guided hybrid human-machine algorithms considering active learning for single label acquisition. Active learning is popular in *single label supervised algorithms*, where the key idea is that a machine learning algorithm can achieve greater accuracy with fewer training labels, if it is allowed to choose the data from which it learns [34, 36, 51, 52], where each data has a single label.

We will study and adapt a set of well-known active learning techniques, such as, *uncertainty sampling [34], query-by-committee [52, 42], or expected-error reduction[47]* that are popularly used in well-known classification algorithm, such as, *Naive Bayes' Classifier, Support Vector Machine [58, 50], Decision Trees [37], or ensemble classification algorithms [32].* Similarly, we will characterize *human factors* of the workers [48, 31, 8], such as *skill, motivation* and then design principled optimization functions that combines task-centric and worker-centric optimization. These complex optimization functions will guide the selection of the right training sample (i.e, task) for further labeling and request the appropriate workers to undertake that task. Using Example 1, this is akin to selecting the most appropriate observation site and select the most appropriate workers to observe the presence or absence of the species there. When multiple workers with varying level of expertise are involved to undertake the same labeling task, we will study how to aggregate their annotations to infer the truth considering *weighted majority voting or iterative approach* [20, 25]. We will formalize *stopping conditions* - i.e., *when to terminate this labeling process by exploiting the confidence [63] of the classification tasks, available budget, or availability of the human workers.* We will investigate effective scalable algorithms to solve these problems by exploiting discrete optimization techniques.

(2) Optimized *multi-labels acquisition* involving crowd : In this aim, we will investigate how to enable active learning principles for *multiple-labels classification tasks* involving crowd (Recall 1). Multi-labels classification is different from multi-class classification, where only a single label needs to be predicted per data point for the latter, albeit there are more than two possible labels. Unlike its single-label counterpart, multi-labels classification using active learning is far less studied, except for a few recent works [55, 24, 35, 61]. In fact, to acquire multiple labels, we are unaware of any related work that attempts to design active learning like techniques involving crowd.

Akin to previous aim, we will adapt a few known active learning algorithms for multi-labels classifications using Support Vector Machine (SVM), Naive Bayes, or Ensemble classifiers [55, 24, 35, 61]. Using this, our objective is to select tasks that will be maximally informative for the classifier. Alternatively, task selection can be guided by a version space analysis such that it will give rise to maximum reduction in the version space of the classifier [57]. We will then augment them with *worker-centric optimization through human factors modeling*, such as worker skill or motivation and design a combined optimization function. This function will dictate which task is to be selected for which worker. Using Example 1, this is akin to selecting the most appropriate unidentified image of the species and select the most appropriate workers to label it. Since a task could be labeled by multiple workers, we will study how to aggregate multiple responses and infer the correct labels (truth inference problem) of a task. We will design an iterative algorithm to effectively infer each task's correct labels. We will also explore the use of correlations among different labels to improve the inference quality. Finally, we will investigate the stopping condition of multi-labels acquisition tasks based on various *convergence criteria*.

We first introduce and characterize different variables (Section 2) pertinent to workers and tasks to describe human factors, then propose worker-centric optimization (Section 3). Both of these are pivotal to investigate crowdsourced single and multi-label tasks through active learning (Sections 3.3 and 4).

2 Data Model

We introduce different variables and characterize human factors [48, 40, 26, 31, 8, 54, 11]. A crowdsourcing platform typically comprises of workers and tasks that serve as the foundation of the framework we propose. We also note that not all the variables are pertinent to every application domain (for example, citizen science applications are usually voluntary contributions). Our effort is to propose a generalization nevertheless.

Domains/types: A set $D = \{d_1, d_2, \ldots, d_l\}$ of given domains is used to describe the different types of tasks in an application. Using Example 1, a particular species may construe a domain.

Workers: A set of m human workers $\mathcal{U} = \{u_1, u_2, \ldots, u_m\}$ are available in a crowdsourcing platform.

Tasks and sub-tasks: A task $\mathcal{T}$ is a hybrid human-machine computational task (classification for example), with a quality condition $Q^{\mathcal{T}}$ and an overall monetary budget $B^{\mathcal{T}}$ that decide its termination. Using Example 1, $\mathcal{T}$ is a classification task which terminates, when $Q^{\mathcal{T}} = 80\%$ accuracy is achieved, or $B^{\mathcal{T}} = \$100$ is exhausted.

Without loss of generality, $\mathcal{T}$ comprises of a set of n subtasks, i.e., $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$. These sub-tasks are of interests to us, as workers will be involved to undertake these sub-tasks. Each sub-task can either be performed by human workers or computed (inferred) by machine algorithms. We consider *pool based active learning*, where a finite pool of sub-tasks exists and given.

Sub-tasks: For single label, a sub-task is an unlabeled instance of the data that requires labeling. Considering Example 1, this is analogous to confirming the presence or absence of a species in a particular geographic location. For multi-label scenario, a sub-task requires multiple labels to be obtained. Using Example 1, this is analogous to obtaining Kingdom, Phylum, Class, Order, etc of the insect.

Worker Response: We assume that a worker u's *response to a particular sub-task* t *may be erroneous, which is used by the machine algorithm in one or more rounds of interactions*. Our framework may ask multiple workers to undertake the same task to reduce the error probability, and may decide which questions to ask in the next round to whom based on the answers obtained in the previous round.

Human Factors: These are the variables that characterize the behavior of the workers in a crowdsourcing platform [48, 40, 26, 31, 8, 54, 11].

Skill (Expertise/Accuracy): Worker's skill in a domain is her expertise/accuracy. Skill of a worker in a domain d is quantified in a continuous $[0, 1]$ scale (to allow a probabilistic interpretation). A worker u may have skills in one or more domains (e.g., different species observation accuracy).

Wage: A worker u may have a fixed wage w_u, or may have to accept the wage a particular task offers. u's may have different wage for different types of tasks.

Motivation: Motivation aims at capturing the worker's willingness to perform a task. A related work [31] proposes a theoretical foundation in motivation theory in crowdsourcing platform and characterizes them in two

different ways:

(a) Intrinsic motivation: Intrinsic motivation exists if an individual works for fulfillment generated by the activity (e.g. working just for fun). Furthermore, related works [31, 8, 54] have identified that intrinsic motivation emerges in the following ways: (1) skill variety (refers to the extent to which a worker can utilize multiple skills), (2) task identity (the degree to which an individual produces a whole, identifiable unit of work, versus completion of a small unit which is not an identifiable final product), (3) task significance (the degree to which the task has an influence over others), (4) autonomy (the degree to which an individual holding a job is able to schedule his or her activities), (5) feedback (the extent to which precise information about the effectiveness of performance is conveyed).

(b) Extrinsic motivation: Extrinsic motivation is an instrument for achieving a certain desired outcome (e.g. making money).

The challenge however is, either the values of these factors have to be explicitly given or they have to be estimated. Related works, including our own, have proposed solutions to estimate skill [46, 30] by analyzing historical data. Nevertheless, we are not aware of any effort that models motivational factors or design optimization involving them.

Worker specific constraints: Additionally, a worker may specify certain constraints (e.g., can not work more than 6 hours, or travel farther than 10 miles from her current location).

Characterizing sub-tasks considering human factors: It is easy to notice that the motivational factors described above are actually related to tasks (i.e, sub-tasks).

Formally, we describe that a set A of attributes or meta-data is available to characterize each sub-task t. They are its required skill-domain[1] s^t , cost/wage w^t, duration $time^t$, location $location^t$, significance sig^t, identity $iden^t$, autonomy $auto^t$, task feedback fb^t. Each t, if performed correctly, contributes by a quantity q^t to $Q^{\mathcal{T}}$. These contributions are purely dictated by the active learning principles, such as how much it reduces the uncertainty.

3 Worker-Centric Optimization through Human Factors Modeling

Recall Section 1 and note that worker-centric optimization is a common theme across single and multi-labels tasks, which we first examine here.

Objectives: Our objective here is to explore mathematical models for worker-centric optimization in crowdsourcing platforms. Specifically, given an available pool of tasks and workers where workers perform repetitive tasks, we first obtain human factors of the workers by analyzing their past tasks and then study the problem of task assignment to enable worker-centric optimization. A recent work performs an ethnomethodological study at *Turker Nation*[2] and argues [39] that it is critical to enable worker-centric optimization. Our effort here is to make a formal step towards that goal, independent of any specific system-centric optimization (i.e., the active learning principles). Therefore, such a study has a much broader applicability that goes beyond active learning. Of course, our framework will ultimately combine both system and worker-centric criteria.

Challenges: While the significance of human factors is well-acknowledged in crowdsourcing, the challenge is to be able to estimate them effectively and propose appropriate models that could capture them during task assignment. Added to the fact is the dependence of the underlying crowdsourcing domain, which makes some of these factors more important than the rest (e.g., unlike AMT, there is no monetary pay-offs in citizen science activity, but skill variety is acknowledged to be critical to avoid monotony).

3.1 Proposed Directions

First, we propose how to model and estimate human factors [48, 40, 26, 31, 8, 54, 11] that are pertinent to capture motivation using the variables that are described in Section 2. Then, we describe mathematical models that

[1] for simplicity, we assume that each sub-task requires one skill, whereas, in reality, multiple skills may be needed for a sub-task. The latter assumption is trivially extensible by our framework.

[2] http://www.turkernation.com/

leverages these estimated human factors to explicitly assign tasks to workers.

Estimating human factors: We leverage the past task completion history of the workers as well as the new tasks to compute a Boolean task completion matrix T, where the rows are the workers and the columns are the (sub)-tasks. If a worker u has completed a (sub)-task t successfully in the past, the corresponding entry gets a 1, it gets a 0 otherwise. We assume that the factors that capture intrinsic motivation, i.e., skill variety, task identity, task significance, autonomy, feedback are independent yet latent variables. The second matrix we consider is the task factor matrix $\mathcal{F}$, where the rows are the tasks and the columns are the motivation related latent variables. The final matrix is the user factor matrix $\mathcal{U}$ where rows are the factors and columns are the users. This matrix could be fully latent or observed. In case it is latent, we minimize the error function, as defined below:

$$\sum_{i,j}(t_{ij} - \mathcal{U}_i F_j)^2 + \lambda(\|\mathcal{U}\|^2 + \|\mathcal{F}\|^2) \tag{1}$$

Here, λ is the regularization parameter. The goal is to find $\mathcal{U}$ and $\mathcal{F}$ such that it minimizes the error. For any new worker and new task, the predicted task completion score is calculated by multiplying U_i with F_j. Here, the important thing is to notice that the optimization function only minimizes the error for which ratings are present. We apply the alternating least square approach [56] to solve this problem. This is an iterative approach, where at each iteration, we fix the tasks' latent factor matrix $\mathcal{F}$ in order to solve for $\mathcal{U}$ and vice versa. We have designed a similar solution for predicting tasks to workers considering implicit workers' feedback[45].

Worker-Centric Task Assignment: The solution above only estimates the intrinsic motivational factors, but does not describe how to aggregate them together or combine with extrinsic motivation to perform worker-centric task assignment.

Psychologists Hackman and Oldham [19] have combined factors associated to intrinsic motivations defined *motivating potential score (MPS)* :

$$MPS = \frac{\text{skill-variety} + \text{task-identity} + \text{task-significance}}{3} * \text{ autonomy } * \text{ feedback} \tag{2}$$

Considering this aforementioned formulation, we study the worker-centric task assignment as a global optimization problem to maximize the *aggregated intrinsic and extrinsic motivation*. For a given set of tasks S^{t_u}, $V(S^{t_u})$ represents the overall motivation for worker u, by combining her extrinsic motivation (EXTM) (recall Section 2 that EXTM could be modeled using wage w^t) and intrinsic motivation, i.e, *motivating potential score (MPS)*(refer to Equation 2) [19]. In our initial effort, we combine them linearly, as that allows us to design efficient algorithms. Assigning a set of tasks per worker is reasonable as well as desirable from worker's perspective, because workers in a typical crowdsourcing platform intend to undertake multiple tasks as opposed to a single task. Workers may also have constraints, such as, not spend more than X^u hours, or the aggregated wage must at least be b^u dollars.

Technically, we want to assign tasks to the workers *to maximize the aggregated motivation, such that the assignment satisfies each worker-specific constraints*. One such optimization function is described in Equation 3 (Recall Section 2 where $time^t$ and w^t are the duration and wage of sub-task t, respectively).

$$\text{Maximize } \sum_{u \in \mathcal{U}} [V(S^{t_u}) = EXTM(S^{t_u}) + MPS(S^{t_u})] \tag{3}$$

$$V(S^{t_u}) = \begin{cases} \text{if } \sum_{t \in S^{t_u}} time^t \leq X^u \text{ and } \sum_{t \in S^{t_u}} w^t \geq b^u \\ 0 \qquad \text{otherwise} \end{cases}$$

As a simple example, given two tasks i and j, we can add the individual significance $sig^i + sig^j$, identity $iden^i + iden^j$, autonomy $auto^i + auto^j$, or feedback $fb^i + fb^j$. Similarly, the wage of two tasks could also be added and normalized to compute EXTM. Alternative problem formulation is explored below.

3.2 Open Problems

Solving the optimization problem: How to design an effective solution to maximize worker motivation based on the aforementioned objective function formulation is challenging. We observe that the proposed optimization problem is NP-hard [14], using a reduction from the assignment problems [49]. In a recent work, we have modeled motivation using *only skill-variety* and we have proved that the problem is NP-hard using a reduction from the Maximum Quadratic Assignment Problem [3]. For our problem, we note that an integer programming based solution is simply not scalable. We will explore greedy heuristic strategies that are effective and efficient. For example, we will assign tasks to the workers greedily based on the marginal gain [49].

Complex modeling for estimating intrinsic motivation & task assignment: In our preliminary direction, we have assumed that variables associated with intrinsic motivations are independent and could be combined as suggested by Hackman and Oldham [19], or intrinsic and extrinsic motivation could be combined linearly. In reality, that may not be the case. In this open problem, we will study the feasibility of a probabilistic model [67], namely a *hierarchical Bayesian framework* [38] for this problem. If the worker is completely new in the platform, we will bootstrap to collect a small set of evidence. We will consider each of the variables associated with worker motivation as a random variable and present a model using hierarchical Bayesian Networks [29] by encoding a joint distribution of these variables over a multi-dimensional space. This model will first establish the relationship among the intrinsic motivational variables themselves and then between intrinsic and extrinsic motivation to capture a workers' "preference" to a given task. We will apply Constraint Based, Score-Based, and Hybrid methods to learn the structure of the network [59]. We will leverage *Bayesian Parameter Estimation as well as Maximum Likelihood Estimation techniques* to learn the parameters of the constructed network. For efficient parameter estimation considering this complex joint distribution, we will use Gibbs sampling [7].

3.3 Optimized Single-Label Acquisition Involving Crowd

We now investigate our proposed optimization framework for single-label acquisition. This problem is examined by augmenting active learning principles with worker-centric optimization (refer to Section 3).

Objectives: We are assuming a setting where single-label acquisition is difficult, expensive, and time consuming (such as, Example 1). We adapt a set of popular as well as well-known active learning principles[34, 52, 42, 47] that are proposed to optimize system-centric cirteria, such as, *minimizing uncertainty or maximizing expected error-reduction* that are known to be effective in supervised (classification) algorithms [58, 50, 37, 32]. We augment these active learning principles with worker-centric optimization. Given a pool of unlabeled instances (of sub-tasks) and an available set of workers, the objective is to select sub-tasks for further labeling and assign workers for annotations, such that, the assignment optimizes both system and workers. The same sub-task may be annotated by multiple workers.

Challenges: An oracle, who knows the ground truth, no longer exists in crowdsourcing; instead, multiple workers, with varying expertise (skill), are available. Under this settings, how to realign traditional active learning goals that are system-centric (i.e., optimizes underlying computational task) requires further investigations. How to systematically design *optimization function*, i.e., one that combines worker-centric optimization in traditional active learning settings [64, 62] is the second important challenge. An equally arduous challenge is the efficiency issue which is mostly overlooked in the existing research. Finally, when to terminate further label acquisition also needs to be examined.

3.4 Proposed Directions

Our overall approach is iterative, where, in each round a set of sub-tasks are selected for annotation and a set of workers are chosen. Once annotations are received, the underlying classification model is retrained. After that, either the process terminates or we repeat. It has three primary directions: (1) *in a given round, which sub-tasks are to be selected for annotation and assigned to which workers?* (2) *how to aggregate multiple annotations to obtain the "true" label?* (3) *when to stop?*

Which sub-tasks are to be selected and assigned to which workers? We take a set of well-known active learning techniques, such as, *uncertainty sampling [34], query-by-committee [52, 42], or expected-error reduction [47],*

used in popular classification algorithms, such as, Naive Bayes [53], SVM [58, 50], Decision Trees [37], or ensemble classification[32] and study them in crowdsourcing.

When a single classifier with a binary classification task is involved and the classifier is probabilistic (such as Naive Bayes), we consider existing uncertainty sampling [34] techniques. We use entropy [53] to model uncertainty to choose that sub-task for labeling whose posterior probability of being positive is closest to 0.5. For non-probabilistic classifiers (such as SVM or Decision Tree), we explore *heterogeneous approach* [33], in which a probabilistic classifier selects sub-tasks for training the non-probabilistic classifier. We also study existing expected-error reduction [47] techniques that select the sub-tasks to minimize the expected future error of the supervised algorithm, considering *log-loss or* $0/1$*-loss*. We study the query-by-committee[52, 42] technique, we choose that sub-task for further labeling which has the *highest disagreement*.

Active learning principles mentioned above are too *ideal* to be useful in a crowdsourcing platform. A simple alternative is to design a *staged solution*, where we first select the tasks and then the workers [64]. For us, we can take the task-selection solution from [64] and then plug in our worker-centric optimization (Section 3) to compose tasks for the workers. We, however, argue that such a staged solution is *sub-optimal*, simply because, tasks selected by *active learning* techniques may end up having a very low worker-centric optimization, resulting in poor outcome overall. We therefore propose a global optimization that combines (1) worker-centric goals (recall Equation 3). (2) active learning principles considering workers with varying expertise.

Recall Section 2 and note that q^t represents sub-task t's contribution towards a given active learning goal (for example, how much t reduces uncertainty or expected-error) at a given iteration. Let S^{t_u} represent the sub-tasks assigned to u with value $V(S^{t_u})$ (recall Equation 3). Considering worker's skill s^{u_t} as a probability, u's *expected contribution* to t is $s^{u_t} * q^t$ [9]. One possible way to combine them is as a multi-objective global optimization function where the objective is to select sub-tasks and workers that maximize a weighted linear aggregation of worker and task-centric optimization (Equation 4, where W_1, W_2 are specific weights). While linear aggregation is not the only way, it is more likely to admit efficient solutions, where the weights are tunable by domain experts (by default, $W_1 = W_2 = 0.5$).

$$\text{Maximize } \mathcal{V} = \sum_{u \in \mathcal{U}} [W_1 * V(S^{t_u}) + W_2 * \sum_{t \in S^{t_u}} (s^{u_t} * q^t)] \tag{4}$$

Additionally, if a task has a cost budget associated that could be assigned either as a constraint to this optimization problem, or we could use cost as another objective as part of the optimization function, akin to one of our recent works [49]. Nevertheless, we acknowledge that designing the "ideal" optimization model that suffices the need of every application is practically impossible. We address this in the open problems.

Aggregating multiple annotations: Another challenge is how to combine annotations from multiple workers with varying expertise to obtain the "true" label. We apply weighted majority voting types of approach [20], where the weights are chosen according to the skills of the workers. We also consider iterative algorithm for this purpose. Examples of iterative techniques include EM or Expectation Maximization[25]. The main idea behind EM is to compute in the E step the probabilities of possible answers to each task by weighting the answers of workers according to their current expertise, and then to compute in the M step re-estimates of the expertise of workers based on the current probability of each answer. The two steps are iterated until convergence. We explore Bayesian solution [9] to probabilistically obtain the true label, i.e., given workers' annotations and skill, compute $Pr(t = 0)$ and $Pr(t = 1)$ and choose the one which has the higher probability.

3.5 Open Problems

Solving the optimization problem Solving the optimization problem described above is challenging. In a very recent work, we have formalized task assignment as a linear combination of task relevance (based on a Boolean match between worker expertise and the skill requirements of a task) and skill-diversity [43] and proved the problem to be NP-Complete [13, 12]. We use Maximum Quadratic Assignment Problem (MAXQAP in short) [3] to design an efficient algorithm with approximation factor $1/4$. For our problem, we will examine if it is at

all possible to design an objective function (perhaps as a special case) to exploit its nice structural properties, such as, *sub-modularity or cancavity*. Such an effort is made for active learning problems recently [22] without considering human workers. We will also study the possibility of staged algorithms and heuristic solutions, as described above. To make the algorithm computationally efficient, we will examine how to design incremental active learning strategies [44], such as finding the new classification model that is most similar to the previous one, under a set of constraints.

Complex function design and stopping condition We note that the formulation described in Equation 4 is rather *simple* - a linear function may not be adequate to combine worker and task-centric optimization. We will explore non-linear multiplicative functions. Another possible way is to formalize this as a bi-criteria optimization problem and design pareto-optimal solution that does not require us to assign any specific weight to the individual functions [6, 2, 4]. Finally, we will examine *when to terminate this iterative process*. For the overall classification task $\mathcal{T}$, when quality threshold is not reached or budget is not exhausted (these are two hard stopping conditions), we will design stopping condition by measuring the confidence [63] of the classification model, or availability of suitable workers.

Develop a number of optimization models that are likely to cover a variety of scenarios We realize that what constitutes the "ideal" optimization model is an extremely difficult problem and highly application dependent (e.g., Which factors are important? Should we add or multiply different human factors? In the case of linear weighting, what should be the weighting coefficients?). Even a domain expert who is very knowledgeable about the specific application may not be able to shed enough light on this. We hope to develop a rich set of different models that will cover the various types of applications. This idea of developing a set of optimization models draw parallels from Web Search and Information Retrieval - where a set of alternative criteria, such as relevance, diversity, and coverage, are considered [5]. In our case, this is analogous to developing models that only consider workers skills/expertise, or cost, or motivation, or includes a subset of human factors that we are interested to study in this project.

4 Optimized Multi-Labels Acquisition Involving Crowd

We now investigate the multi-labels acquisition scenario. We are unaware of any related work that performs multi-label acquisition in an active learning settings involving crowd. Although one can transform a multi-label task to several single-label tasks, this simple approach can generate many tasks, incurring a high cost and latency. Akin to the previous section, our effort is to design solutions that adapt a few recent active learning works [55, 24, 35, 61] for multi-label acquisition and combine that with worker-centric optimization, described in Section 3.

Objectives: We will adapt a few known active learning algorithms for multi-label classifications using Support Vector Machine (SVM), Naive Bayes, or Ensemble classifiers [55, 24, 35, 61]. We will combine and augment them with *worker-centric optimization through human factors modeling*. Using Example 1, this is akin to selecting the most appropriate unidentified image of the species and select the most appropriate workers to provide multiple labels. Since a task could be labeled by multiple workers, we will study how to aggregate multiple responses and infer the correct labels (truth inference problem) of a task. We will also explore the use of correlations among different labels to improve the inference quality. Finally, we will investigate the stopping condition or *convergence criteria*.

Challenges: Workers may exhibit different characteristics in multi-label tasks: a conservative worker would only select labels that the worker is certain of, while a casual worker may select more labels. To determine the multi-label tasks' results, the key is to devise the so-called "worker model" to accurately express the behavior of the worker in answering multi-labels. Furthermore, different from single-label tasks, correlations among labels inherently exist in multi-label tasks. For Example 1, consider one pairwise label dependency: if the insect in the image is labeled as Papilionidae (Family name) , then it is highly probable that it also has label Swallowtail (Sub-family name). Therefore, how to understand and leverage label correlation is another challenge. Finally, how to systematically design *optimization function*, i.e., one that combines worker-centric optimization in active

learning settings [55, 24, 35, 61] is the final important challenge.

4.1 Proposed Directions

Our overall approach is iterative here as well, where, in each round a set of sub-tasks are selected to be annotated with multi-labels and a set of workers are chosen. Once multiple labels are acquired, the underlying classification model is retrained. After that, either the process terminates or we repeat. It has three primary directions: (1) *Task assignment* (2) *Truth Inference, i.e., aggregate multiple annotations to obtain the "true" labels.* (3) *Label Correlation*.

Task Assignment: In our preliminary investigation, we have studied the active learning problem for the multi-label scenario considering the widely popular SVM classifier using the *Maximum-Margin Uncertainty Sampling*. Uncertainty sampling [34] is one of the simplest and most effective active learning strategies used for single-label classification. The central idea of this strategy is that the active learner should query the instance which the current classifier is most uncertain about. For binary SVM classifiers, the most uncertain instance can be interpreted as the one closest to the classification boundary by selecting the sample with the smallest classification margin. Multi-label active learning methods simply extend this binary uncertainty concept into the multi-label learning scenarios by integrating the binary uncertainty measures associated with each individual class in independent manners, such as taking the minimum over all classes, and taking the average over all classes.

In our initial direction, given the active learning principle, we combine that with worker-centric optimization and design an objective function akin to Equation 4, as described in Section 3.3. Obviously, exploring alternative optimization models, or how to design a set of optimization functions that can handle a variety of scenarios, or when to stop the iterative process are additional challenges. Once we understand these challenges for the single-label acquisition problem in Section 3.3, we believe they will extend for the multi-label scenarios.

Truth Inference Problem: The truth inference problem, i.e, how to aggregate the annotations provided by multiple workers and generate the actual set of labels requires deeper attention for the multi-label scenario. As the correct set of labels associated with each sub-task is unknown (ground-truth is unknown), the accuracy or expertise of a worker can only be estimated based on the collected answer. To model worker expertise, we compute the following two measures, *True Positive (TP)* and *False Positive (FP)*. TP is the number of labels that a worker selected correctly and FP is the number of labels she selected incorrectly. Unlike a prior work [67], False Negative and True Negative are not relevant, if the workers annotate the labels. In the case where workers validate the given labels, these latter two measures are also relevant. Once these measures are computed, we design a worker's contingency table and calculate her expertise. After that, we design an iterative approach, which can jointly infer the correct labels associated with the tasks and the expertise of the workers. Our iterative solution is motivated by the Expectation Maximization (EM) algorithms and comprises of the following two steps: (step 1), we assume that the worker expertise is known and constant, and infer the probabilistic truth of each object and label pair. (step 2), based on the computed probabilistic truth of each object and label pair, we re-estimate workers expertise.

Label correlation: Since the annotated labels of an object are not independent (Recall Example 1 and note that Papilionidae (Family name) and Swallowtail (Sub-family name) are highly correlated), we study how label correlations can be inferred and facilitate truth inference. In our initial direction, we leverage the existing label correlation techniques [65, 66] to generate the *pairwise label correlations* and regard them as prior input to our problem. For example, the conditional dependency of two labels defines the probability that one label is correct for an object under the condition that the other label is correct. Capturing the higher order label correlations requires computing the joint probability which could be computationally expensive. Once label correlation is computed, we shall explore how to use that information for improved truth inference.

4.2 Open Problems

Alternative Active Learning Strategy Design In our initial direction, we have discussed how to adapt uncertainty sampling to design active learning strategies for SVM classifier for multi-label scenario. The average number of

correct labels assigned to each instance in a multi-label data set is called its label cardinality. Thus the number of predicted labels of an unlabeled instance is expected to be consistent with the label cardinality computed on the labeled data. For an unlabeled instance, this inconsistency measure could be defined as the distance between the number of correctly predicted labels so far and the label cardinality of the labeled data. We will study this **label cardinality inconsistency** [60] to select that sub-task where the label inconsistency is highest. Additionally, we will also study the active learning strategies known for other classifiers, such as Naive Bayes and Ensemble methods could be adapted to our problem [55, 24, 35, 61]. Alternatively, task selection can be guided by a version space analysis such that it will give rise to maximum reduction in the version space of the classifier [57].

Truth Inference with Label Correlation We will study how to use the information obtained from label correlation to improve the truth inference. Intuitively, our truth inference problem could benefit from label correlation in the following way: using Example 1, if label correlation infers high correlation among two labels, let's say, Papilionidae and Swallowtail (family and sub-family of butterflies), it is likely that Papilionidae and Mimic Sulfurs (which is a sub-family of butterflies, but Mimic Sulfurs belong to a different family (Pieridae) will have a very low correlation. Therefore, the probabilistic truth of the labels which have Mimic Sulfurs should be downgraded to reflect that fact. It has been shown in Information Retrieval that the more frequent two words occur together in text corpus, the more similar their vectors are [5]. We will regard each label as a word and compute the similarity (e.g., cosine similarity) between the vectors of two labels. We will explore widely popular Sigmoid function [28] to map a probability value to a real value, re-scale the value based on label correlation, and then revert the re-scaled correlation back to a probability score using the Sigmoid function again.

5 Conclusion

The goal of this article is to propose an *an optimized human-machine intelligence framework for single and multi-label tasks through active learning*. We conceptualize an iterative framework that judiciously employs human workers to collect single or multiple labels associated with such tasks, which, in turn are used by the supervised machine algorithms to make intelligent prediction. Our basic approach is adapt a few existing *active learning techniques for single and multi-label classification, but study them in the context of crowdsourcing, especially considering worker-centric optimization, i,e., human factors*. Our innovation lies in systematically characterizing variables to model human factors, designing optimization models that appropriately combine system and worker-centric goals, and designing effective solutions.

6 Acknowledgment

The work of Senjuti Basu Roy is supported by the National Science Foundation under Grant No.: 1814595 and Office of Naval Research under Grant No.: N000141812838.

References

[1] S. Amer-Yahia and S. B. Roy. Human factors in crowdsourcing. *Proceedings of the VLDB Endowment*, 9(13):1615–1618, 2016.

[2] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 839–848. ACM, 2012.

[3] E. M. Arkin, R. Hassin, and M. Sviridenko. Approximating the maximum quadratic assignment problem. *Information Processing Letters*, 77(1):13–16, 2001.

[4] A. Asudeh, G. Zhang, N. Hassan, C. Li, and G. V. Zaruba. Crowdsourcing pareto-optimal object finding by pairwise comparisons. *arXiv preprint arXiv:1409.4161*, 2014.

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

[6] V. Bilò, M. Flammini, and L. Moscardelli. Pareto approximations for the bicriteria scheduling problem. In *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, page 83. IEEE, 2004.

[7] C. K. Carter and R. Kohn. On gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994.

[8] D. Chandler and A. Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133, 2013.

[9] R. T. Clemen and R. L. Winkler. Aggregating probability distributions. *Advances in decision analysis: From foundations to applications*, pages 154–176, 2007.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[11] E. Estellés-Arolas and F. González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

[12] T. Feo, O. Goldschmidt, and M. Khellaf. One-half approximation algorithms for the k-partition problem. *Operations Research*, 40(1-supplement-1):S170–S173, 1992.

[13] T. A. Feo and M. Khellaf. A class of bounded approximation algorithms for graph partitioning. *Networks*, 20(2):181–195, 1990.

[14] M. R. Garey and D. S. Johnson. Computers and intractability: a guide to np-completeness, 1979.

[15] M. B. Garzón, R. Blazek, M. Neteler, R. S. De Dios, H. S. Ollero, and C. Furlanello. Predicting habitat suitability with machine learning models: the potential area of pinus sylvestris l. in the iberian peninsula. *ecological modelling*, 197(3):383–393, 2006.

[16] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu. Corleone: hands-off crowdsourcing for entity matching. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 601–612. ACM, 2014.

[17] N. Grira, M. Crucianu, and N. Boujemaa. Active semi-supervised fuzzy clustering for image database categorization. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 9–16. ACM, 2005.

[18] D. Haas, J. Wang, E. Wu, and M. J. Franklin. Clamshell: Speeding up crowds for low-latency data labeling. *Proceedings of the VLDB Endowment*, 9(4):372–383, 2015.

[19] J. R. Hackman and G. R. Oldham. Motivation through the design of work: Test of a theory. *Organizational behavior and human performance*, 16(2):250–279, 1976.

[20] C.-J. Ho, S. Jabbari, and J. W. Vaughan. Adaptive task assignment for crowdsourced classification. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 534–542, 2013.

[21] T. Hofmann and J. M. Buhmann. Active data clustering. *Advances in Neural Information Processing Systems*, pages 528–534, 1998.

[22] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424. ACM, 2006.

[23] Y. Huang and T. M. Mitchell. Text clustering with extended user feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 413–420. ACM, 2006.

[24] C.-W. Hung and H.-T. Lin. Multi-label active learning with auxiliary learner. In *ACML*, pages 315–332, 2011.

[25] N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*, pages 1–15. Springer, 2013.

[26] P. G. Ipeirotis. Demographics of mechanical turk. 2010.

[27] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[28] Y. Ito. Approximation of continuous functions on r d by linear combinations of shifted rotations of a sigmoid function with and without scaling. *Neural Networks*, 5(1):105–115, 1992.

[29] F. V. Jensen. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.

[30] M. Joglekar, H. Garcia-Molina, and A. G. Parameswaran. Evaluating the crowd with confidence. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 686–694, 2013.

[31] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, pages 1–11, 2011.

[32] C. Körner and S. Wrobel. Multi-class ensemble-based active learning. In *Machine Learning: ECML 2006*, pages 687–694. Springer, 2006.

[33] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156, 1994.

[34] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[35] X. Li and Y. Guo. Active learning with multi-label svm classification. In *IJCAI*. Citeseer, 2013.

[36] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

[37] C. X. Ling, Q. Yang, J. Wang, and S. Zhang. Decision trees with minimal costs. In *Proceedings of the twenty-first international conference on Machine learning*, page 69. ACM, 2004.

[38] R. Liu, K. Zolfaghar, S.-c. Chin, S. B. Roy, and A. Teredesai. A framework to recommend interventions for 30-day heart failure readmission risk. In *2014 IEEE International Conference on Data Mining*, pages 911–916. IEEE, 2014.

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

[40] W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

[41] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden. Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment*, 8(2):125–136, 2014.

[42] K. Nigam and A. McCallum. Employing em in pool-based active learning for text classification. In *Proceedings of ICML-98, 15th International Conference on Machine Learning*, 1998.

[43] J. Pilourdault, S. Amer-Yahia, D. Lee, and S. B. Roy. Motivation-aware task assignment in crowdsourcing. In *EDBT Conference*, 2017.

[44] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10):1880–1897, 2009.

[45] H. Rahman, L. Joppa, and S. B. Roy. Feature based task recommendation in crowdsourcing with implicit observations, 2016.

[46] H. Rahman, S. Thirumuruganathan, S. B. Roy, S. Amer-Yahia, and G. Das. Worker skill estimation in team based tasks. *PVLDB*, 2015.

[47] N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[48] S. B. Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Crowds, not drones: Modeling human factors in interactive crowdsourcing. In *DBCrowd 2013-VLDB Workshop on Databases and Crowdsourcing*, pages 39–42. CEUR-WS, 2013.

[49] S. B. Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal*, pages 1–25, 2015.

[50] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846. Citeseer, 2000.

[51] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[52] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[53] C. E. Shannon. A note on the concept of entropy. *Bell System Tech. J*, 27:379–423, 1948.

[54] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 275–284. ACM, 2011.

[55] M. Singh, E. Curran, and P. Cunningham. Active learning for multi-label image annotation. In *Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science*, pages 173–182, 2009.

[56] S. M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, pages 465–474, 1981.

[57] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[58] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[59] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[60] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*, 2006.

[61] D. Vasisht, A. Damianou, M. Varma, and A. Kapoor. Active learning for sparse bayesian multilabel classification. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 472–481. ACM, 2014.

[62] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision*, 108(1-2):97–114, 2014.

[63] A. Vlachos. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312, 2008.

[64] Y. Yan, G. M. Fung, R. Rosales, and J. G. Dy. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1161–1168, 2011.

[65] Y. Yu, W. Pedrycz, and D. Miao. Multi-label classification by exploiting label correlations. *Expert Systems with Applications*, 41(6):2989–3004, 2014.

[66] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 999–1008. ACM, 2010.

[67] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5(6):550–561, 2012.

Interaction Management in Crowdsourcing

Yansheng Wang [†], Tianshu Song [†], Qian Tao [†], Yuxiang Zeng [‡], Zimu Zhou [#],
Yi Xu [†], Yongxin Tong [†], Lei Chen [‡]
[†]BDBC, SKLSDE Lab and IRI, Beihang University, China
[‡]The Hong Kong University of Science and Technology, Hong Kong SAR, China
[#] Singapore Management University, Singapore
[†]{arthur_wang, songts, qiantao, xuy, yxtong}@buaa.edu.cn
[‡]{yzengal, leichen}@cse.ust.hk, [#]zimuzhou@smu.edu.sg

Abstract

Crowdsourcing is anticipated as a promising paradigm for Future of Work (FoW), where groups of humans are engaged for problem-solving, services and innovation which are usually difficult for machines. During the entire workflow of crowdsourcing, intensive interactions take place between workers and the crowdsourcing platform, as well as among groups of workers. For sustainable crowdsourcing, the design and management of these interactions should not regard human workers as machines, but rather as individuals and social beings. In this article, we highlight the critical interactions in typical crowdsourcing ecosystems, summarize past efforts on human-centric interaction management in crowdsourcing, and discuss emerging interaction management research towards cross-platform crowdsourcing.

1 Introduction

Some computation tasks are intractable for computers (or machines) but easy for human beings, such as image recognition, text translation, entity resolution, etc. Although the new breakthroughs of artificial intelligence and machine learning in recent years have relieved the difficulty of solving computer-hard problems to some degree, it still relies on large scale of training data which needs to be labeled manually. Crowdsourcing is a calculation paradigm to solve computer-hard tasks effectively and efficiently. It organizes large scale of human beings as workers through the Internet to solve problems. For decades, crowdsourcing has been applied in various areas. Many crowdsourcing platforms have achieved business success. A typical example is Amazon Mechanical Turk (AMT) , which is the biggest general-purpose crowdsourcing platform. According to a blog published by AMT , in a typical week there are millions of tasks that are processed by over a million workers all over the world. With the aid of the productive forces of AMT, the famous data set of image recognition, ImageNet , is labeled and has a huge impact. Besides, there are many people making a living on the income earned from AMT.

Despite the above achievements of crowdsourcing, the welfare of crowd workers is often neglected. The workers are usually regraded more as machines with some kinds of computing functions, which seems not a wise choice. On the one hand, although workers are viewed as machine, they are not as reliable or stable as computers. Thus, much effort has to be put to tame the uncertainty of workers by modeling their accuracy rates

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

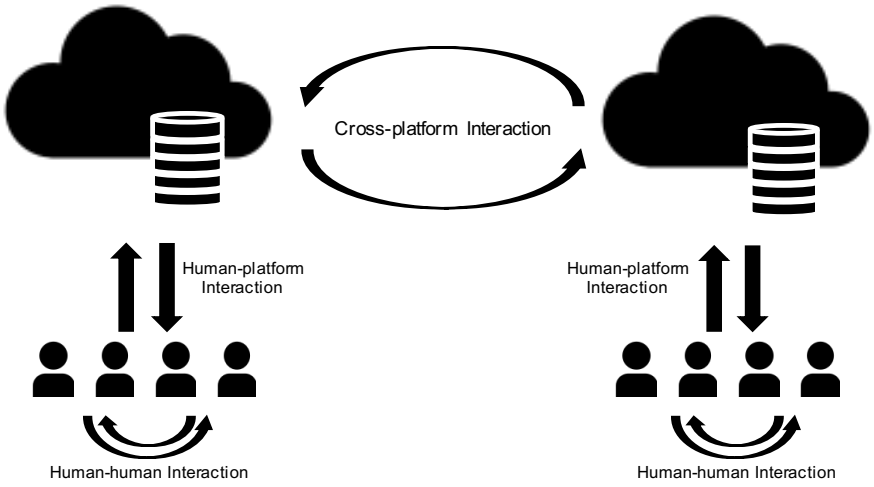

Figure 1: Categories of interactions.

and aggregating feedbacks. On the other hand, workers are more flexible and intelligent than machines. If their initiative can be stimulated, crowdsourcing platforms will be more productive and workers' welfare can also be enhanced.

To achieve the goals of stimulating the initiative of workers to further improve the effectiveness of crowdsourcing platforms and helping the workers achieve self-actualization to make crowdsourcing a form of the future of work, this paper targets at the optimization of critical interactions in typical crowdsourcing ecosystems. To optimize the interaction between human workers and the platforms, *i.e.* the human-platform interaction, the *freewill* of human beings should be considered. The main part of human-platform interaction is task assignment, which assigns crowd tasks to suitable workers. Most existing studies on task assignment only consider how to maximize the utility and ignore whether the workers would like to perform the assigned tasks. To make task assignment more humanized, tasks should be assigned to workers based on the workers' will. More friendly methods of task assignment should be designed. To optimize the interaction between human workers, *i.e.* the human-human interaction, the *sociality* and *selfishness* of human beings should be considered. When the tasks are simple or well-paid, workers may complete them more efficiently. Thus, incentive mechanisms should be designed based on workers' rationality and selfishness. When the tasks are complex and hard to complete, the platforms need the workers to corporate with each other. Thus, the social relationships of workers should be considered too.

The rest of the paper is organized as follows. We categorize the interactions on crowdsourcing platforms in Sec. 2. We discuss how to optimize the human-platform interaction and human-human interaction in crowdsourcing in Sec. 3 and Sec. 4 respectively. We envision the future directions in Sec. 5 and conclude the paper in Sec. 6.

2 Types of Interactions in Crowdsourcing

An interaction process usually involves human workers and platforms as necessary participants. Therefore, we can divide interactions on crowdsourcing platforms into three categories: human-platform interaction, human-human interaction and cross-platform interaction, as shown in Figure 1.

Human-platform interaction. Interaction between human workers and platforms is an essential process for accomplishing crowdsourced tasks. Most human-platform interactive behaviors happen in the task assignment

stage, where the platform assigns tasks to appropriate human workers in order to achieve some specific targets such as maximizing the total profits. We have recognized that assigning tasks to human workers is rather different from assigning tasks to robots or AI agents, because human workers have free wills and they have rights to reject the assigned tasks or to choose whatever tasks they would like to carry out. Therefore, we will focus on the following two kinds of interactions during task assignment process. The first is called *task selection by workers*, where the platform displays all available tasks with their specifications, constraints and payments, and the workers can choose the expected tasks freely. This scenario is more common and can be found on many existing crowdsourcing platforms like Amazon Mechanical Turk (AMT) and Figure Eight. The second is called *task recommendation by platforms*, where the platform recommends tasks to workers actively and each worker can reject the assignment in which case the platform has to continuously recommend new tasks to them until the task is accepted. We will give a detailed introduction to them in Sec. 3.

Human-human interaction. Apart from the communication between human workers and platforms, there also exist interactions between different workers due to their social activities, which can influence the process of performing the tasks. For example, workers with close social relationships like friends or relatives can work together as a team to accomplish complicated tasks. Considering that human workers are rational and selfish, the platform cannot always expect them to stick together for the sake of finishing all the tasks more effectively and efficiently. Therefore, proper guidance of human-human interaction like incentive mechanisms can help improve the overall utility meanwhile respecting each individual's interests. In a way, we can encourage workers to form groups based on their social relations, common likes and dislikes. They will have *collaborative interaction* between each other, which can be more efficient than working alone. In another way, we can also allow friendly competitions between workers or groups and give the winners additional rewards. The *competitive interaction* can improve the efficiency of both sides and can also help the human workers find the value and fun in their efforts. We will introduce the above two human-human interactions in Sec. 4.

Cross-platform interaction. So far, we have discussed about the interactions inside the cycle of a single crowdsourcing platform and its workers. But we can also imagine a new type of interaction that exists between different platforms, which we call cross-platform interaction. Even though today's crowdsourcing platforms are isolated from each other, it is prospective that they cooperate with each other as a federation in the future to serve more tasks and give opportunities to more human workers. In such a scenario, the cross-platform interaction will play a rather important role. The challenges of designing proper cross-platform interaction methodology not only lie in dealing with heterogeneous systems, but also in taking care of the unique requirements of human workers: the *privacy concerns*. For example, suppose there is a supply constraint on Figure Eight and many tasks on the platform cannot be finished in time. Therefore, Figure Eight wants to federate with AMT which has sufficient worker supply. But workers on AMT do not always like to share their profiles and working histories with Figure Eight, which can influence the utility of assignment. The two platforms have to interact with each other to form a federation meanwhile no privacy leakage of workers will happen during the interaction process. We will discuss about such cross-platform interaction as future directions in Sec. 5.

3 Human-Platform Interaction: Task Assignment

In this section, we focus on the interaction management in task assignment between human workers and platforms. Specifically, we identify two types of human-platform interactions during the task assignment in crowdsourcing. In Section 3.1, we introduce the first type, *i.e. task selection by human workers*, where the workers can freely choose the desired tasks and a task is assigned to the first worker who selects it. In Section 3.2, we discuss the second type, *i.e. task recommendation by platforms*. In this type, the platform learns the preferences and profiles of the workers from the history records and then recommends the suitable tasks to the workers based on different methods. After that, the human workers can either accept the tasks or reject the tasks.

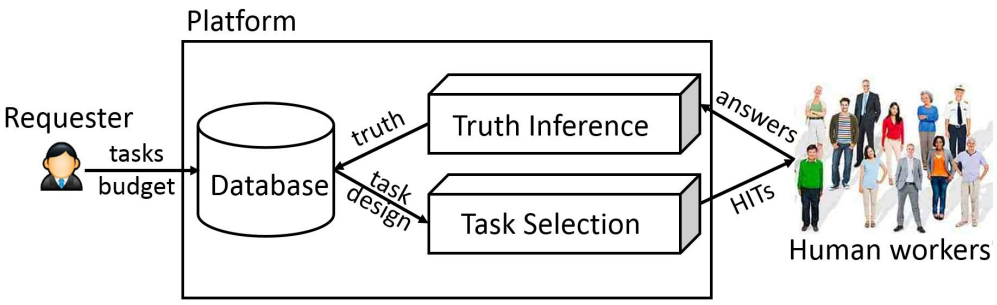

Figure 2: An illustration of the workflow in task selection [2].

3.1 Task Selection By Human Workers

In the process of task assignment, many crowdsourcing platforms apply the strategy of *task selection by human workers* in the interaction management, *e.g.* Amazon Mechanical Turk, Wikipedia, Quora, Yahoo!Answer. Specifically, these platforms usually first collect the tasks from the requesters, then decompose them into micro-tasks [1], and finally pack several micro-tasks into an HIT (*a.k.a.* task bin [1]). After the platforms release these HITs into well-designed user interface, human workers start to interact with the platforms in the process of task assignment. As shown in Figure 2, there are mainly two stages in this workflow: *task selection* and *truth inference*.

Task Selection. In this stage, human workers can freely choose the desired HITs and an HIT is assigned to the first worker who picks it. This mode is also known as *worker selected tasks (WST)* mode in the early studies [3, 4]. To attract the human workers, existing studies usually focus on *task design* in this type of interaction. Specifically, [2, 5] introduce *task types* and *task settings* as the fundamental problems in the task design. When designing the task types, the crowdsourcing platforms need to carefully choose the format of the tasks. For example, a problem with single choice is usually easier to be answer by the workers than the problem with multiple choices. As a result, more studies focus on the tasks with single choice (*e.g.* [6, 7, 8]) than the tasks with multiple choices (*e.g.* [9]). However, the tasks with multiple choices are relatively more profitable for the workers and more helpful for inferring the truth (*i.e.* the second stage). A good task design combines these two types in practice. For instance, Vasilis *et al.* [10] propose the approach called Waldo to detect the difficult tasks with multiple choices and only resolve them by single choice. When considering the task settings, the crowdsourcing platforms mainly determine the *pricing* and *timing* of the tasks. Usually, a high price can intuitively attract more workers, thereby reducing the total time to complete a task (*i.e.* low latency). However, a high price does not always indicate a good quality of the answer [11]. Therefore, a good task design needs to consider the trade-off between the pricing and timing. For instance, Yihan *et al.* [12] applies Markov Decision Process to determine the price and deadline of the tasks. Vasilis *et al.* [13] designs a dynamic-programming based method in the task setting while simultaneously choosing the proper task type.

Truth Inference. In this stage, the crowdsourcing platform will collect the answers from the workers after they accomplish their selected tasks. The challenge in this stage is how to infer the truth of each task based on the different tasks' answers collected from human workers [5]. Many solutions have been proposed to solve the problem of truth inference. A simple solution is to use majority voting, where the truth is viewed as the majority answer of the workers. However, this method overlooks the fact that the workers usually have different qualities in practice. To obtain the worker's quality, one way is to use the hidden test, where a golden question is injected in an HIT. Since the platform knows the truth of this golden question, it can estimate the workers' quality more accurately. However, the hidden test takes more money since the workers need to complete these extra tasks. Moreover, as reported in [14], the hidden tests may not improve much quality. Therefore, existing studies usually

focus on inferring the truth without the hidden tests. The most classic framework is Expectation Maximization (EM) . The EM method is firstly proposed in [15], which leverages all the collected answers from the workers, and updates each worker's quality and each task's true answer until convergence [2]. Other studies use the Graph-based methods [16, 17], where a worker's quality is represented as a node and is derived by the graphical model inference [18].

3.2 Task Recommendation By Platforms

There are many crowdsourcing platforms which take another mode *i.e.* task recommendation by platforms to interact with crowd workers during task assignment, such as Uber, Didi Chuxing [19, 20] and Facebook Editor. Early studies formulate the task recommendation problem as bipartite graph matching problem. Specifically, workers and tasks can be represented by the vertices in the bipartite graph and the score to evaluate the suitability of a worker and a task can be denoted by the weight of the edges. Then the problem is to obtain an optimal matching in the bipartite graph.

Statics Scenario. When the information of tasks and workers is known in advance (or before task recommendation), exact algorithms (*e.g.* Hungarian algorithm) can optimally solve the problem. Alternatively, Kazemi *et al.* [3] reduce the bipartite graph into an instance of the *maximum flow* problem and use the Ford-Fulkerson algorithm to obtain the exact result. They also consider some practical issues. For example, if for a task there are fewer workers who are suitable to perform it, the task should be recommended with higher priority. The idea of entropy can be borrowed to represent this priority. Another heuristic strategy is to iteratively assign the task to its most suitable workers. To reduce the computation cost of exact solutions, various greedy based methods are proposed. For example, [21] maximizes the total number of tasks which are accepted by the workers while considering a budget constraint.

Dynamical Scenario. In reality, the time when new tasks are submitted to the platform and the content of the tasks are unknown. Besides, the workers also log in the platforms dynamically. Thus, some studies address the task recommendation problem while considering the dynamics of tasks and workers. They formulate the problem as online bipartite graph matching. Karp *et al.* [22] first propose the RANKING algorithm which yields a competitive ratio $1 - 1/e$ under the *adversarial order model* and the ratio is proven to be the lower bound of any online algorithm. In [23], Tong *et al.* borrow the idea of secretary problem and devise a two-phase based framework. Their proposed algorithms achieve competitive ratios of 0.25 and 0.125 under the *random order model*. In [24], an comprehensive experimental study has been made and the greedy algorithm is proved to be competitive in real scenarios. More flexible and adaptive online matching algorithms have also been proposed in [25, 26, 27, 28]. In addition, Zhang *et al.* [29] focus on predicting the acceptance ratio of workers in task recommendation via machine learning techniques. They treat the rejected tasks by workers as new tasks that can be re-recommended to other workers. We refer readers to the comprehensive surveys [30] for more details on task assignment approaches.

4 Human-Human Interaction: Incentive Mechanism

Despite the essential interactions between humans and platforms, the sociality from humans has made human-human interactions more and more important in the future of work. These interactions, which we summarize as the incentive mechanisms, depict how workers interact with each other and focus on how to design mechanisms to motivate workers. To be more specific, on one hand, workers tend to be multi-skilled and social as crowdsourcing develops. Those positive interactions (*e.g.* complementary in skills or closeness in social contact) between humans can help improve the satisfaction of the workers and performance of the platforms. On the other hand, the freewill of humans enables selfish workers. Those malicious factors lead to inevitable competition, which

needs to be treated carefully. Correspondingly, we categorize the human-human interactions as collaborative interaction and competitive interaction. In the following of this section, we will discuss about them respectively.

4.1 Collaborative Interaction

An important kind of human-human interaction is the collaborative interaction. On one hand, tasks become more and more complicated and they usually require multiple skills to accomplish. Workers with different skills should collaborate to accomplish the complicated tasks. On the other hand, social factor may be helpful to improve the task quality and latency. For example, two workers which are good friends may accomplish the tasks with higher quality and lower latency. In a word, in this part we introduce two representative collaborative factors in human-human interaction, *i.e.* multi-skill factor and social factor.

Multi-skill Factor. Multi-skill factor is one of the significant factors to depict complicated tasks and has been widely studied in recent years. In early studies like [31, 32], the skills are formulated as the skill level value representing the probability that a worker submits the true answer. [33] utilizes the formulation of expertise level and applies a transfer learning model to estimate the expertise levels in all categories. To overcome the disadvantage that a single value cannot depict multiple skills, [34] extends the formulation to the environment of multiple skills and uses a vector called latent skill category to represent how a worker is familiar with different skills. [35] and [36] study the so-called team formation problem and propose the solutions with the help of social networks and road networks, respectively. In the future of work, despite of existing formulations like skill level and team formation, we may facet a more complicated multi-skill situation where the task requires a time dependency of skills. Under this situation, a task may require a worker with skill A and a worker with skill B sequentially, which requires a novel solution based on the accurate estimation of worker skills.

Social Factor. As human workers have their own social roles apart from workers on crowdsourcing platforms, the social factor can also influence the completion of tasks. Some work has used the social factors like relationships on social networks and communities [37] to help increase the crowdsourcing quality. In [38], the authors address the decision making question answering by crowdsourcing on micro-blog services. One of the most fundamental steps in addressing the problem is to estimate the rating score of different workers. The authors first construct a user-graph with Twitter data according to the workers' forwarding operation retweet, and then rank the users in the constructed graph. Therefore, expertises that are frequently followed by other twitters can be found through the user graph easily. In [39], the authors propose a novel community-based Bayesian label aggregation model, which can predict the labels of different workers by assuming that each worker belongs to a certain community. Some other work models the social relationships with game theory and reveals the importance of social factors in crowdsourcing. In [40], it considers the social team formation problem in crowdsourcing, where a team of socially connected professional workers can work together collaboratively. It focuses on designing truthful mechanisms according to workers' social structure, skill, and working cost to guarantee that a worker's utility is optimized when he behaves honestly. [41] directly studies the social factors in incentive mechanisms in crowdsourcing and finds that collaboration leads to better accuracy and more outputs. It reveals that in pair-work crowdsourcing, when one of the workers tries to exit the task, the partner is very likely to make him/her stay for higher payment and for enjoying the tasks together. Therefore, social furtherance incentives can create a win-win scenario for both the platform and the human workers.

Table 1 summarizes existing works on collaborative interactions. We summarize that both multi-skill and social factors raise attention but their formulation needs to be further completed to suit for more challenging situations.

4.2 Competitive Interaction

Despite the collaboration in human-human interaction, there also exists competitions in the interaction of crowdsourcing. Different from collaboration, malicious competitions may reduce the quality of the tasks or even

Table 1: Related works on collaborative interaction.

Reference	Interaction	Factor	Formulation
[31]	Collaborative	Multi-Skill Factor	Skill Level
[32]	Collaborative	Multi-Skill Factor	Skill Level
[33]	Collaborative	Multi-Skill Factor	Expertise Level
[34]	Collaborative	Multi-Skill Factor	Latent Skill Category
[35]	Collaborative	Multi-Skill Factor	Team Formulation
[36]	Collaborative	Multi-Skill Factor	Team Formulation
[38]	Collaborative	Social Factor	Rating Score
[39]	Collaborative	Social Factor	Label Prediction
[40]	Collaborative	Social Factor	Team Formulation
[41]	Collaborative	Social Factor	Social Furtherance

destroy the human-in-the-loop ecosystem. However, passive competitions can be beneficial if treated carefully. In this part we envision how to utilize competitions to encourage workers to participate in completing tasks with higher quality and lower latency. In details, depending on whether the reward is monetary or immaterial, we categorize the interaction as monetary factor and non-monetary factor.

Monetary Factor. Monetary is one of the most important factors that can influence how workers perform. Current studies usually formulate the monetary factor in competitive human-human interaction based on the game theory. In a word, a carefully designed pricing mechanism should be able to provide a well balance between payments (*e.g.* prices paid to workers) and to promote workers to submit the truthful information. Most works provide auction-based solutions to model the competitive interaction between humans. Workers need to bid their expected prices to the platform, and the platform chooses workers to accomplish tasks based on the proposed mechanisms. Notice that even there is no direct communication among workers, the mechanism which is known by all workers in fact provides an indirect competitive interaction among workers. Because of the selfishness of workers, the mechanisms should be truthful, *i.e.* the mechanism guarantees that workers cannot obtain more reward if they provide false information. Besides, another consideration of designing the mechanism is how to balance various constraints and objectives. For example, [42, 43, 44] aim at maximizing the utility (*i.e.* the value obtained from the workers) while making sure that the total budget is under a given threshold. Different from [42] and [43], [45] focuses on two different objectives: maximizing the number of accomplished tasks with constrained budget and minimizing total payment with least number of accomplished tasks. In future of work, the interaction may facet more challenging situations. Firstly, the private information may be more complicated (comparing with current research where each worker is associated with a private expected value). For example, each worker may privately possess the total order of preference to the tasks, the capability of different kinds of tasks, etc.. Besides, in future of work there can be other practical trade-offs and objectives like trade-off between payment and quality of tasks or between payment and latency of workers. These complicated private information and various objectives can raise more challenging problems in the competitive human-human interaction.

Non-Monetary Factor. Despite the dominant role of the monetary factor, none-monetary factor (*e.g.* gamification, volunteering etc.) also plays an important role in the competitive human-human interaction. Specifically, gamification can attract workers to join the procedure [46] and result in positive competitions that motivate workers to provide better services [47]. For example, [46] develops a series of games to attract workers to attain the crowdsourcing tasks. Further experiments show that the proposed games can improve the quality and latency of the accomplished tasks. Compared with [46], [47] combines the luck games (*e.g.* lottery) with crowdsourcing. Each worker is given one or more lotteries and which worker can be paid is related not only to the quality of the accomplished tasks, but also to the lottery results. In the future of work, non-monetary factor may become more and more important. An interesting study direction is to formally model how non-monetary factor like

Table 2: Related works on competitive interaction.

Reference	Interaction	Factor	Formulation
[45]	Competitive	Monetory Factor	Auction
[42]	Competitive	Monetory Factor	Auction
[43]	Competitive	Monetory Factor	Auction
[48]	Competitive	Monetory Factor	Collaboration
[46]	Competitive	Non-Monetory Factor	Relevance Assessments
[47]	Competitive	Non-Monetory Factor	Lottory

gamification produce positive competitive interaction and impact the performance of workers in crowdsourcing.

Table 2 summarizes existing works on competitive interactions. We summarize that most existing works focus on the interaction with monetary factor. Non-Monetary factor, although lacking in systematic study in crowdsourcing, can also play an important role in future work.

5 Future Direction: Cross-Platform Interaction

In the previous sections, we have discussed about human-platform and human-human interaction in existing crowdsourcing platforms. In this section, we are going to introduce a future direction: the cross-platform interaction. Nowadays, there is a large amount of commercial crowdsourcing platforms that divide the online job markets into many pieces. If these platforms can work together as a federation (as illustrated in Figure 3), task requesters will have a higher chance to meet more professional workers and workers will also have more career opportunities. However, sensitive information like worker profiles, professional skills and working histories cannot be shared freely between different platforms, for the purpose of preserving each human worker's privacy. Therefore, the cross-platform interaction should be more carefully designed and the major concern is how to preserve the privacy of human workers from different platforms.

There is rich literature on privacy-preserving schemes, including encryption methods like homomorphic encryption [49] and randomized algorithms like differential privacy [50]. Next, we will discuss about the possible shape of the interactive modules including cross-platform task assignment and cross-platform incentive mechanism, as well as the privacy-preserving techniques behind them.

Cross-platform task assignment. The interaction in cross-platform task assignment can be different from conventional task assignment where the platform has full knowledge of its tasks and workers. In a cross-platform task assignment scenario, workers will have opportunity to complete the tasks from other platforms. We conceive two types of interaction in cross-platform task assignment: interaction by recommendation and by negotiation. In the first type, surplus tasks on each platform can be shared with other platforms in the form of recommendation and the task specifications and the original links will be given. The recommended tasks from other platforms will be visible in the task pool of each platform, and the workers can select the preferred tasks freely. If they are interested in tasks from other platforms, they can have a quick registration through the link and become a cross-platform worker. Considering that there might be a huge number of pending tasks in the whole platform federation, one of the major challenge is to recommend for each worker a personalized ranking list of tasks so that they can locate their favorite tasks in a short time. A federated recommender system can be designed to protect the privacy of each worker while making personalized and accurate recommendations. Randomized algorithms like locality-sensitive hashing are possible solutions to protect private tags of workers in federated recommender systems. In the second type, the platforms will negotiate with each other at first. For example, there are surplus tasks on Figure Eight and the platform would like to borrow workers from AMT to finish these tasks. Figure Eight will negotiate with AMT, about the amount and types of the tasks and the profit shares meanwhile AMT can provide some information on workers to select appropriate tasks. Afterwards, the selected tasks will become

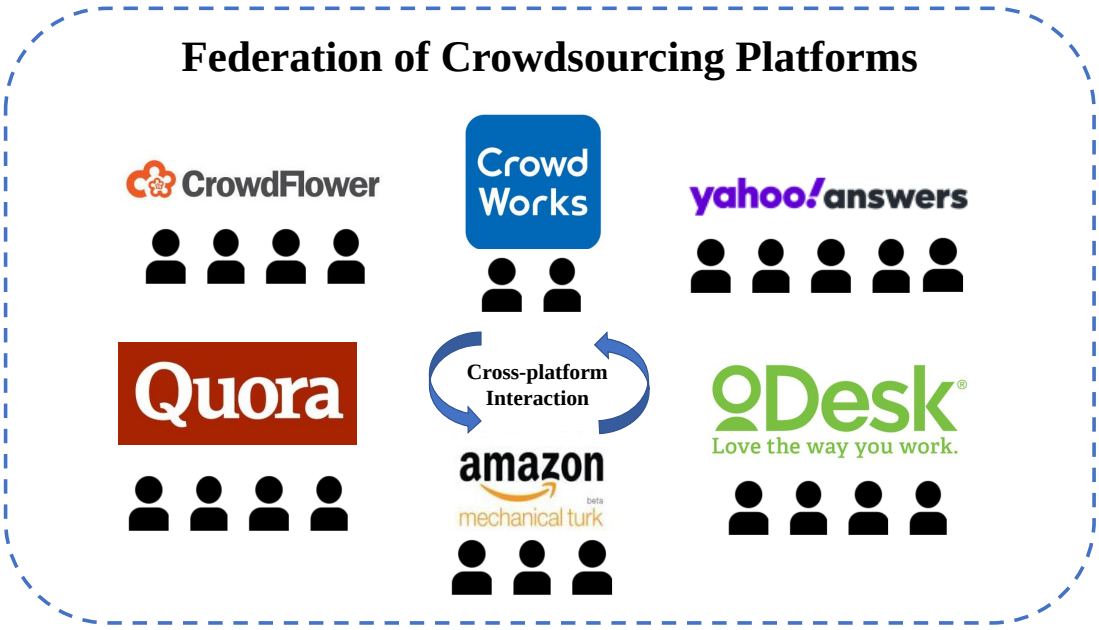

Figure 3: An illustration of the federation of crowdsourcing platforms

available to workers on AMT as new tasks. After the completion of tasks, AMT will return only the results and Figure Eight should know little about the specific workers who complete them. To protect the private profiles and working histories of the workers, there should be anonymization techniques during the negotiation. For example, AMT can only provide the summary or aggregated results of worker abilities for Figure Eight to decide which tasks can be shared. Another possible problem is that AMT can hide the information of profit shares from the workers if the payment by Figure Eight is much higher and how to ensure the trust and fairness remains challenging.

Cross-platform incentive mechanism. Similar to cross-platform task assignment, cross-platform incentive mechanism also faces privacy challenges. We mainly focus on incentive mechanisms in collaborative scenarios, where the privacy leakage is much more likely to happen. In a collaborative scenario, workers from different platforms can have a chance to form groups to perform cross-platform tasks. In one way, they can cooperate to perform a batch of simple tasks respectively and aggregate the results afterwards. For example, the federated platform publishes a task of labeling a large number of CT images. Expertises like doctors from different platforms can form a team to accomplish the tasks. They can divide the tasks according to their areas of expertise so the group work can be more efficient. In another way, each worker can also perform one or several steps in the working pipeline of complicated tasks. For example, the task is to tag all the objects in photos of street views and then to annotate the texts on street signs and buildings. It can be divided into several steps, including marking all the objects, annotating the texts and double-checking the results, where each step can be conducted by workers from different platforms. In both ways, the federated platform should design team formation algorithms based on social networks or ability tags of different workers, which also has risk of privacy leakage. Although there is plenty of work concerning privacy issues in social networks, how to preserve personal information while finding like-minded partners still remains a challenge.

6 Conclusion

Crowdsourcing has attracted much popularity in recent years, due to its high efficiency in dealing with computer-hard tasks by human workers. However, we have noticed the problem that humans are often treated as machines. To deal with such problem, managing the interaction among the workers and platforms is essential. In this paper, we categorize and analyze different interactions on crowdsourcing platforms, and discuss about the challenges and existing work on human-platform interaction and human-human interaction respectively. Finally, we envision a new type of interaction, the cross-platform interaction. We hope in the future of crowdsourcing, there will be a federation of platforms that offer much more opportunities to everyone everywhere and human workers will be treated as fully human with respect and dignity.

References

[1] Y. Tong, L. Chen, Z. Zhou, H. V. Jagadish, L. Shou, and W. Lv, "SLADE: A smart large-scale task decomposer in crowdsourcing," *IEEE TKDE*, vol. 30, no. 8, pp. 1588–1601, 2018.

[2] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, "Crowdsourced data management: A survey," *IEEE TKDE*, vol. 28, no. 9, pp. 2296–2319, 2016.

[3] L. Kazemi and C. Shahabi, "Geocrowd: enabling query answering with spatial crowdsourcing," in *GIS*, 2012, pp. 189–198.

[4] D. Deng, C. Shahabi, and U. Demiryurek, "Maximizing the number of worker's self-selected tasks in spatial crowdsourcing," in *GIS*, 2013, pp. 314–323.

[5] G. Li, Y. Zheng, J. Fan, J. Wang, and R. Cheng, "Crowdsourced data management: Overview and challenges," in *SIGMOD*, 2017, pp. 1711–1716.

[6] D. Firmani, B. Saha, and D. Srivastava, "Online entity resolution using an oracle," *PVLDB*, vol. 9, no. 5, pp. 384–395, 2016.

[7] J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng, "Leveraging transitive relations for crowdsourced joins," in *SIGMOD*, 2013, pp. 229–240.

[8] Y. Zeng, Y. Tong, L. Chen, and Z. Zhou, "Latency-oriented task completion via spatial crowdsourcing," in *ICDE*, 2018, pp. 317–328.

[9] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *PVLDB*, vol. 5, no. 11, pp. 1483–1494, 2012.

[10] V. Verroios, H. Garcia-Molina, and Y. Papakonstantinou, "Waldo: An adaptive human interface for crowd entity resolution," in *SIGMOD*, 2017, pp. 1133–1148.

[11] S. Faradani, B. Hartmann, and P. G. Ipeirotis, "What's the right price? pricing tasks for finishing on time," in *HCOMP*, 2011.

[12] Y. Gao and A. G. Parameswaran, "Finish them!: Pricing algorithms for human computation," *PVLDB*, vol. 7, no. 14, pp. 1965–1976, 2014.

[13] V. Verroios, P. Lofgren, and H. Garcia-Molina, "tdp: An optimal-latency budget allocation strategy for crowdsourced MAXIMUM operations," in *SIGMOD*, 2015, pp. 1047–1062.

[14] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: Is the problem solved?" *PVLDB*, vol. 10, no. 5, pp. 541–552, 2017.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–22, 1977.

[16] Z. Zhao, F. Wei, M. Zhou, W. Chen, and W. Ng, "Crowd-selection query processing in crowdsourcing databases: A task-driven approach." in *EDBT*, 2015, pp. 397–408.

[17] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *NIPS*, 2011, pp. 1953–1961.

[18] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[19] Y. Tong, Y. Zeng, Z. Zhou, L. Chen, J. Ye, and K. Xu, "A unified approach to route planning for shared mobility," *PVLDB*, vol. 11, no. 11, pp. 1633–1646, 2018.

[20] Y. Zeng, Y. Tong, and L. Chen, "Last-mile delivery made practical: An efficient route planning framework with theoretical guarantees," *PVLDB*, vol. 13, no. 3, pp. 320–333, 2019.

[21] H. To, L. Fan, L. Tran, and C. Shahabi, "Real-time task assignment in hyperlocal spatial crowdsourcing under budget constraints," in *PerCom*, 2016, pp. 1–8.

[22] R. M. Karp, U. V. Vazirani, and V. V. Vazirani, "An optimal algorithm for on-line bipartite matching," in *STOC*, 1990, pp. 352–358.

[23] Y. Tong, J. She, B. Ding, L. Wang, and L. Chen, "Online mobile micro-task allocation in spatial crowdsourcing," in *ICDE*, 2016, pp. 49–60.

[24] Y. Tong, J. She, B. Ding, L. Chen, T. Wo, and K. Xu, "Online minimum matching in real-time spatial data: Experiments and analysis," *PVLDB*, vol. 9, no. 12, pp. 1053–1064, 2016.

[25] T. Song, Y. Tong, L. Wang, J. She, B. Yao, L. Chen, and K. Xu, "Trichromatic online matching in real-time spatial crowdsourcing," in *ICDE*, 2017, pp. 1009–1020.

[26] Y. Tong, L. Wang, Z. Zhou, B. Ding, L. Chen, J. Ye, and K. Xu, "Flexible online task assignment in real-time spatial data," *PVLDB*, vol. 10, no. 11, pp. 1334–1345, 2017.

[27] Q. Tao, Y. Zeng, Z. Zhou, Y. Tong, L. Chen, and K. Xu, "Multi-worker-aware task planning in real-time spatial crowdsourcing," in *DASFAA*, 2018, pp. 301–317.

[28] Y. Wang, Y. Tong, C. Long, P. Xu, K. Xu, and W. Lv, "Adaptive dynamic bipartite graph matching: A reinforcement learning approach," in *ICDE*, 2019, pp. 1478–1489.

[29] L. Zhang, T. Hu, Y. Min, G. Wu, J. Zhang, P. Feng, P. Gong, and J. Ye, "A taxi order dispatch model based on combinatorial optimization," in *KDD*, 2017, pp. 2151–2159.

[30] Y. Tong, Z. Zhou, Y. Zeng, L. Chen, and C. Shahabi, "Spatial crowdsourcing: a survey," *The VLDB Journal*, 2019. [Online]. Available: https://doi.org/10.1007/s00778-019-00568-7

[31] C. Ho and J. W. Vaughan, "Online task assignment in crowdsourcing markets," in *AAAI*, 2012.

[32] C. Ho, S. Jabbari, and J. W. Vaughan, "Adaptive task assignment for crowdsourced classification," in *ICML*, 2013, pp. 534–542.

[33] Z. Zhao, D. Yan, W. Ng, and S. Gao, "A transfer learning based framework of crowd-selection on twitter," in *KDD*, 2013, pp. 1514–1517.

[34] Z. Zhao, F. Wei, M. Zhou, W. Chen, and W. Ng, "Crowd-selection query processing in crowdsourcing databases: A task-driven approach," in *EDBT*, 2015, pp. 397–408.

[35] D. Gao, Y. Tong, J. She, T. Song, L. Chen, and K. Xu, "Top-k team recommendation and its variants in spatial crowdsourcing," *Data Science and Engineering*, vol. 2, no. 2, pp. 136–150, 2017.

[36] W. Wang, Z. He, P. Shi, W. Wu, and Y. Jiang, "Truthful team formation for crowdsourcing in social networks: (extended abstract)," in *AAMAS*, 2016, pp. 1327–1328.

[37] L. Zhang, T. Song, Y. Tong, Z. Zhou, D. Li, W. Ai, L. Zhang, G. Wu, Y. Liu, and J. Ye, "Recommendation-based team formation for on-demand taxi-calling platforms," in *CIKM*, 2019, pp. 59–68.

[38] C. C. Cao, J. She, Y. Tong, and L. Chen, "Whom to ask? jury selection for decision making tasks on micro-blog services," *PVLDB*, vol. 5, no. 11, pp. 1495–1506, 2012.

[39] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *WWW*, 2014, pp. 155–164.

[40] W. Wang, Z. He, P. Shi, W. Wu, Y. Jiang, B. An, Z. Hao, and B. Chen, "Strategic social team crowdsourcing: Forming a team of truthful workers for crowdsourcing in social networks," *IEEE TMC*, vol. 18, no. 6, pp. 1419–1432, 2019.

[41] O. Feyisetan and E. Simperl, "Social incentives in paid collaborative crowdsourcing," *ACM TIST*, vol. 8, no. 6, pp. 73:1–73:31, 2017.

[42] A. Singla and A. Krause, "Truthful incentives in crowdsourcing tasks using regret minimization mechanisms," in *WWW*, 2013, pp. 1167–1178.

[43] N. Anari, G. Goel, and A. Nikzad, "Mechanism design for crowdsourcing: An optimal 1-1/e competitive budget-feasible mechanism for large markets," in *FOCS*, 2014, pp. 266–275.

[44] Y. Tong, L. Wang, Z. Zhou, L. Chen, B. Du, and J. Ye, "Dynamic pricing in spatial crowdsourcing: A matching-based approach," in *SIGMOD*, 2018, pp. 773–788.

[45] Y. Singer and M. Mittal, "Pricing mechanisms for crowdsourcing markets," in *WWW*, 2013, pp. 1157–1166.

[46] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan, "Quality through flow and immersion: gamifying crowdsourced relevance assessments," in *SIGIR*, 2012, pp. 871–880.

[47] M. Rokicki, S. Chelaru, S. Zerr, and S. Siersdorfer, "Competitive game designs for improving the cost effectiveness of crowdsourcing," in *CIKM*, 2014, pp. 1469–1478.

[48] S. K. Mridha and M. Bhattacharyya, "Introducing collaboration in competitive crowdsourcing markets," *IEEE Intelligent Systems*, vol. 34, no. 1, pp. 23–31, 2019.

[49] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *STOC*, 2009, pp. 169–178.

[50] C. Dwork, "Differential privacy," in *ICALP*, 2006, pp. 1–12.

Platform Design for Crowdsourcing and Future of Work

David Gross-Amblard
University of Rennes, France
dga@irisa.fr

Atsuyuki Morishima
University of Tsukuba
mori@slis.tsukuba.ac.jp

Saravanan Thirumuruganathan
QCRI, HBKU
sthirumuruganathan@hbku.edu.qa

Marion Tommasi
University of Lille and INRIA
marion.tommasi@inria.fr

Ko Yoshida
CrowdWorks
yoshida@crowdworks.co.jp

Abstract

Online job platforms have proliferated in the last few years. We anticipate a future where there exists thousands of such platforms covering wide swathes of tasks. These include crowdsourcing platforms such as Amazon Mechanical Turk (AMT), CrowdWorks, Figure Eight; specialized services such as ride-hailing; matching markets such as TaskRabbit that matches workers with local demand and so on. It is widely anticipated that a vast majority of human workforce will be employed in these platforms. In this article, we initiate discussions about the under studied aspect of platform design *– how to design platforms that maximize the satisfaction of various stakeholders. We also contribute a novel taxonomy for platform ecosystems that categorizes existing and emerging platforms. Finally, we discuss the need for interoperability between these platforms so that workers and requesters are not tied to a single platform.*

1 Introduction

Online job platforms are a class of alternative job arrangements that have widely proliferated in the last few years. These could be general purpose platforms such as Amazon Mechanical Turk or CrowdWorks that support a wide variety of tasks entirely online. Alternatively, they could be specialized platforms such as for ride hailing services where all tasks are of the same type. Hybrid platforms such as TaskRabbit matches *labor* providers such as plumbers with requestors who need such services. Even platforms such as Etsy could be construed as a job platform as it allows *goods* providers to sell their products. Job platforms acts as an intermediary by matching workers with requestors while taking a cut in the payments. As much as 36% of US workers have an alternative work arrangement in some capacity [12]. This number is steadily increasing and it is in the realm of possibility that the vast majority of human workforce will be employed in such arrangements in the near future.

We anticipate a future where there are thousands of online job platforms for different types of tasks. Due to their bespoke nature, this will not be a winner takes all. Despite the impending Cambrian explosion of such platforms, there has been very limited work on *platform design* – what are the current pain points of current platforms and how can they be better designed to handle the future of work? In this article, we investigate this important problem and make a number of concrete suggestions.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

There are three major stakeholders – workers, requestors and the online job platforms. It is important that the platform design takes into account the requirements of each of them. The current generation of platforms are grappling with how to satisfy each of them. So far, the efforts are ad-hoc and error prone wherein none of them are satisfied with the status quo.

Current generation of platforms support simple microtasks only. Support for complex tasks is often rudimentary at best. We believe that in the future, platforms must be able to support complex knowledge intensive and collaborative tasks that could involve sophisticated workflows. When there are a large number of job platforms, it is very important to have formally defined mechanisms to allow platforms to inter-operate with each other. This is often beneficial for workers (they could work on multiple platforms at once) and for requesters (recruit workers from multiple platforms). A worker should be able to move between platforms without losing any of the skills/ratings from the original platform. Similarly, a requester must be able to post a task where workers from multiple platforms are able to contribute. Achieving this requires tackling a number of technical and policy challenges. The technical challenges include agreeing upon a predefined schema and API interfaces. Furthermore, one must identify a class of queries that allows exchange of information about workers, requesters and tasks. The regulatory policies must ensure that platforms play well together.

Paper Outline. The rest of the paper is organized as follows. In Section 2, we conduct an user study over multiple popular crowdsourcing platforms whose results indicate the need for platform design. In Section 3, we enumerate a number of key functionalities that are either missing or not widely supported in the current generation of crowdsourcing platforms. However, these are also the functionalities that were desired by stakeholders or have the potential to improve their satisfaction. In Section 4, we discuss the important problem of platform interoperability. As more and more online job platforms are created, it is important that they can interact with each other and we identify the key dimensions for enabling this functionality.

2 Motivation

Crowdsourcing platforms has been extremely well studied by different communities [3]. We believe that they are harbingers of the oncoming shift towards the gig economy brought upon by online job platforms. By studying how workers fare in these platforms allows us to extrapolate the findings to Future of Work (FoW).

Crowdsourcing platforms including Amazon Mechanical Turk (AMT) and CrowdWorks have a number of similarities to online job platforms. Requestors create various tasks that include information such as description of the work required, compensation provided, requestor details etc. The requestor can also filter workers based on previous experience. When a worker logs on, she could see all the available tasks and choose to work on a subset of them. When the worker submits a task, it could be reviewed by the requestor. If deemed satisfactory, the worker is paid. If not, the worker submission is rejected. The crowdsourcing platform takes a cut from the payment made by requestor to the worker. This basic model pioneered by AMT has become ubiquitous in online job platforms as diverse as Uber, TaskRabbit and so on.

A number of studies such as [2] have found that there is a large turnover among workers in Crowdsourcing platforms. Often, the barriers for turnover is much less steeper than in a traditional employment setting. This often causes workers to either completely stop working in a platform such as AMT or have a partial turnover where they just stop working on tasks from a particular requestor. It is important to address both types of turnover. The former could be illustrative of dissatisfaction with the platform whereas the latter could be due to dissatisfaction with requestors.

We collaborated with CrowdWorks, one of the largest crowdsourcing platforms in Japan. It provides support for more than 200 types of tasks. It is used by more than 25000 companies and close to a million active workers. We begin by analyzing the turnover rate of workers in CrowdWorks. Figure 1 shows what fraction of workers persist with the platform over a period of time. One can see that as much as 75% of workers drop off within couple of months. The number of workers who persist for more than 2 years is as little as 5%. The success of any

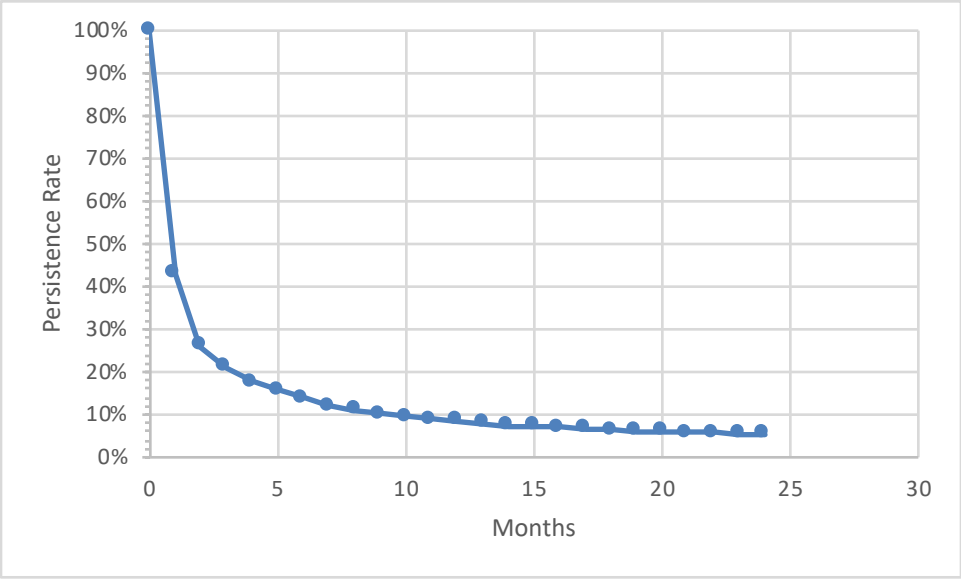

Figure 1: Persistence Rate of Workers in CrowdWorks, according to the log of all workers who made accounts from January 1st, 2017 to December 31th, 2018.

Q1	Platform	Q2					Total
		Very Easy	Easy	Neither	Difficult	Very Difficult	
Yes	AMT	14%	47%	12%	0%	0%	73%
	CrowdWorks	0%	2%	6%	2%	0%	10%
No	AMT	4%	6%	6%	10%	2%	26%
	CrowdWorks	6%	18%	48%	16%	2%	90%

Figure 2: Results of Questions 1 and 2. The result suggests a correlation between the worker mobility and the easiness of switching among crowdsourcing platforms.

crowdsourcing platform hinges on successfully retaining workers. Hence, identifying the factors causing worker turnover and mitigating them is of paramount interest. It is expected that workers who drop-off have a wide variety of reasons for doing so. We are specifically interested in the role of platform design for this phenomenon. By understanding how workers feel about current platforms and their desiderata for new functionalities, one can perform a better platform design.

In December 2019, we conducted a poll of crowd workers on AMT and CrowdWorks, that have different characteristics; AMT is a microtask-based crowdsourcing platform, and CrowdWorks provides a wide variety of task types, focusing on a a bit larger tasks such as writing articles and codes. This involved 101 workers (51 AMT and 50 CrowdWorks workers). Figure 2 summarizes the answers to the following questions: (Q1) Do you envision switching between platforms in the next X months? (Q2) How easy/challenging is switching between the platforms? The result clearly shows that there is a strong correlation between worker mobility and the easiness of switching between platforms. We assume that since AMT is a pure microtask platform and obtaining a good reputation is easier, more workers think switching is easy in AMT than in CrowdWorks. Figure 3 shows that workers would like platforms to have many advanced features that are not necessarily fully supported by the current generation of platforms. It is interesting to see that many workers on CrowdWorks dislike the collaboration feature among workers. We have not pursued the reasons yet, but a possible cause is the difference in the granularity of tasks.

Q3. Preference for the feature?	Platform	Answers		
		Like	Dislike	I don't know
Displaying Credentials	AMT	90%	4%	6%
	CrowdWorks	62%	20%	18%
Specifying/ Quantifying/ Learning Skills	AMT	86%	6%	8%
	CrowdWorks	68%	18%	14%
Anonymity	AMT	75%	18%	10%
	CrowdWorks	66%	18%	16%
Complex Workflows	AMT	67%	20%	14%
	CrowdWorks	42%	12%	46%
Collaboration	AMT	84%	4%	12%
	CrowdWorks	28%	42%	30%

Figure 3: Whether workers would like platforms to support for the features or not

3 Platform Design

Based on the analysis of the poll of crowd workers and extensive discussions with participants of the 2019 Shonan Seminar on Future of Work[1], we have identified platform design as one of the key drivers for ensuring the continued success of online job platforms. We first identify a number of issues with the current platform design and make a number of concrete suggestions.

Lack of Interoperability: The current generation of online job platforms operates in silos. Even platforms that are in a specific domain such as driving (Uber, Lyft, Ola, Didi, etc.) are not interoperable with each other. A driver who has driven more than 10K rides with 4.9 rating in Uber will start as a newbie when moving to a different platform such as Lyft. This is also a problem for requesters. Consider a task that requires 10 experts. It is possible that the labor market has the requisite experts – but they are scattered across multiple platforms. In this case, the requester could not successfully complete the task. By enabling interoperability between platforms such issues and many more could be ameliorated.

Lack of Support for Complex Tasks and Workflows: Currently, there are a limited number of tasks for which crowdsourcing is possible. Even sophisticated platforms such as CrowdWorks only support as little as 200 types of tasks. As more and more task types are being serviced by gig economy, the need for supporting more complex tasks becomes important. These complex tasks often have very different set of requirements. For example, they might require a sophisticated workflow so that output of one stage is passed to another. They could require workers with different types of roles. There might also be a need for specialized requirements such as splitting a complex task into multiple sub-tasks that could then be assigned using a workflow.

Limited Pricing Model: There are four types of pricing models that are widely prevalent in online job platforms. First is the fixed income model where each worker is given a fixed amount of money every month or so. There is also task based pricing where the worker is paid a pre-agreed amount after completing a task. There are some intermediate approaches such as fixed income plus a bonus amount and discounted pricing for completing large number of tasks. Finally, there are competitions pricing models in which we pay to the winner. For online job platforms to thrive, it is important to have a wider variety of pricing models.

For the remainder of the section, we propose a number of improvements to platforms that could either reduce the pain points of the key stakeholders or improve their satisfaction.

[1] ACM Sigmod Blog, Sept. 23, 2019, `http://wp.sigmod.org/?p=2931`

3.1 Platform Design for Workers

Workers are the back bone of a job platform by completing the tasks of requesters. As discussed in the previous section, the current generation of platforms are not very conducive for long term employment.

Ease of On- and Off-boarding: One of the positive things about online job platforms is the flexibility they provide to the worker. The worker can take tasks that are of interest to them while operating flexible hours. Many platforms simplify them with as little as some identity document and bank account. While joining the platform is straightforward, the onboarding process often leaves much to be desired. Once the worker joins the platform, they are not provided enough guidance to contribute productively. The worker is expected to learn how to contribute on their own. Similar to offline employment in a traditional organization, it is important to have a proper onboarding procedure. The current process for off-boarding is also ad-hoc. Typically, workers and requesters can easily stop working in a platform. However, the workers often lose the reputation that they have earned when moving to a different platform. Similarly, the requesters also lose access to valuable and productive employees when moving to a different platform. It is important to have a better off-boarding platforms so that the workers could transfer the reputation and knowledge learned from one platform to another.

Support for Learning Skills. When a worker joins a platform, she often learns "on-the-job". If the worker does not perform well due to inexperience, the job could be rejected by the requester thereby affecting the approval rate of the worker. Since a number of requesters filter workers based on task approval rate, this could limit the number of tasks a new worker could contribute to. This often leads to frustration of new workers and eventual turnover. It is often desirable to have a more formal mechanism for simplifying this process. For example, job platforms could have a collection of previously completed tasks that could serve as an on-ramp for the workers. By working on such tasks, the worker can learn the requisite skill in a low stress environment.

Knowledge Base (KB) for Workers: Currently, there are a number of knowledge repositories in an enterprise so that employees know the practices of the company. It is important that online job platforms provide something similar. Note that this is in addition to the aforementioned set of completed tasks. As workers finish tasks, they must be able to add things to a personalized knowledge base about what they learned from the task. This could be public so that any worker can learn from it or private where it is visible only to the worker. As workers become increasingly knowledgeable, this serves as a repository for what they learned over the years. Of course, this must also be interoperable and associated with the worker and transferable as needed. This will also allow workers to find other mentors or experts to learn from.

Support for Expressing Workers' Preference on Task Assignment The task assignment is usually done by workers themselves, partly because automatic assignment of tasks to workers is not straightforward. There are many reasons for the worker to do the tasks; they did the task because it was easy to do, the task was interesting, they wanted to learn something from the task, it gave them a lot of money, or it was the regular time slot for the worker to do tasks. If the platform has the support for them to express their preferences, platforms will be able to automatically suggest them the tasks more accurately.

3.2 Platform Design for Requesters

In this subsection, we highlight some of the major pain points of requesters and how new functionality from platforms could improve their satisfaction.

Expressive Specification of Task Requirements. Currently, there is limited support from online job platforms for requesters to precisely specify their requirements. For example, in AMT, the requester can filter workers based on approval rate but cannot impose additional sophisticated filtering. It is important for requesters to be able to specify worker requirements such as skills [7], output requirements such as latency, cost and quality. Furthermore, the requester could be open to various tradeoffs such as cost vs quality / latency. For example, one might be willing to pay higher for better quality or faster response. Unfortunately, current platforms do not provide such flexibility.

Supporting Complex Tasks and Workflows. Almost all of the current job platforms support simple microtasks. As more and more tasks are disrupted by the gig economy, it is important to have platforms that can support more complex tasks. Often, complex tasks have a number of distinct requirements. They are often knowledge intensive and collaborative [14] requiring co-ordination with multiple workers. They often are not monolithic and must be split into multiple sub-tasks with different groups of workers completing each of them. They also often require workers to embrace different roles. Finally, they often have a complex workflow where the output of one stage is passed to the next.

Support for Workflow Evolution and Changes during the Execution. Workflows sometimes need to change or evolve during the execution, since completing all tasks requires a long time in general and it is often the case that we find better workflows after we start to execute the original workflow. The platforms should support this kind of evolution and changes of workflows with a minimum amount of effort.

Sophisticated Algorithms for Assigning Workers to Tasks. Currently, most crowdsourcing platforms do not have any sophisticated algorithms in matching workers to tasks. Often, this is done manually by the workers by browsing the list of available tasks. Other than filtering the pool of workers, requesters do not have much control on which workers perform the tasks. Despite extensive research in algorithms for task assignment [3, 8, 14], they are not often incorporated into the platforms. It is important to either have sophisticated algorithms for task assignment so that requester specifications are satisfied or provide an alternate way for the requesters to select workers.

Support for AI Workers. Given the increasing capabilities of AI, it is a matter of time when AI workers become a major part of online job platforms. There are a number of scenarios where AI workers could be useful. If there are urgent tasks, then it is not always possible to use humans to answer them. Often, there is a substantial latency when humans are involved. In these circumstances, AI workers could be a valuable resource. Alternatively, the requester could be cash-strapped and willing to accept less accurate results in exchange for cheaper payments. There are many collaborative situations where AI takes care of the boring work while the human works on the subset that requires human intuition. Finally, AI and humans should be able to replace each other in some situations: an AI can be used as a fallback is a human is answering too late to an urgent task, and conversely, humans can be used as a fallback if an AI fails at recognizing something critical. However, one must be careful in how AI workers are integrated. If not, they could replace human workers causing significant social strife. Furthermore, it is important for requesters to understand the advantages and limitations of the AI workers.

Total Optimization. Currently, most of the work on matching is done in a piecemeal manner where best workers are identified for each task. It is often important to have a holistic optimization that takes the preferences of all the stakeholders into account. This would ensure that good workers are overloaded with work and the workers/tasks are matched in a fair manner.

Algorithm Boutiques: A typical crowdsourcing platform could be improved by incorporating algorithms into the major components including (i) how the task requirements are specified (ii) how tasks are assigned to workers (iii) how the ground truth of tasks are obtained by aggregating worker responses and (iv) how the skills of workers are learned based on their response to tasks. Unfortunately, most of the platforms do not incorporate any of these algorithms. Most of these points are offloaded to the workers and requesters. While experienced requesters often have a concrete mechanisms to effectively achieve each of them, the vast majority of requesters have an ad-hoc and sub-optimal procedure. Finally, even if some platforms implement the algorithms, they are often opaque and the requesters have no say in how they are chosen. It is important for a crowdsourcing platform to have a boutique of algorithms from which requesters can choose the specific variants.

Bespoke Platforms: Most of the crowdsourcing platforms are not very customizable. For example, a domain scientist might need a custom crowdsourcing platform for the task at hand. Currently, the scientist is left with two unappealing choices. Either use an existing platform by approximating the task to suit the constraints of the platform. Alternatively, the scientist could build a new platform from scratch at tremendous cost. In order to unleash the future of work, it is important to enable any requester to create custom online job platforms. It must have a set of default algorithms that could be customized by the requester. The emergence of on-demand

computing frameworks such as Amazon AWS or Microsoft Azure lowered the cost of startups by relieving them of the pressure of managing servers. We believe that the time is ripe to do something analogous for crowdsourcing.

Library of Workflows and AI Workers. In order to help new requesters, it is important for crowdsourcing platforms to provide a large collection of commonly used workflows. Similarly, they could also provide some AI workers that could perform limited set of tasks albeit with the understanding that the work could be of lower quality that of a human.

Open Source Academic Platforms. Most of the online job platforms are closed source and proprietary. This inhibits research on improving various components of the platform and evaluate the potential impact. One natural solution for driving further research in platform design is through open source academic platforms. A well designed and modular platform could allow a researcher to modify certain components, investigate how it impacts the stakeholders and use to improve platform design. Currently, the researcher has to more or less implement the end-to-end crowdsourcing platform which could be prohibitively challenging. There are a number of promising options such as Headwork and Crowd4U. Headwork [2] is a proof-of-concept crowdsourcing platform focusing on skill modeling and complex task workflows, using Tuple Artifacts. Crowd4U [9] is a nonprofit open microvolunteering and crowdsourcing platform for academic and public purposes. It has been widely used for tasks such as identifying building damages during natural disasters, annotative tweets, translation, identifying paths of tornados and so on. The most appealing property is its ability to extend the functionality through a datalog type language called CyLog [10]. For example, when the authors came up with an algorithm [14] for task assignment in collaborative crowdsourcing setting they were able to easily implement it on top of Crowd4U [9]. Such functionality has the potential to dramatically improve crowdsourcing research on platform design.

4 Platform Interoperability

In the previous section, we discussed the various functionalities that could be added to the platform to improve it. In this section, we discuss how additional benefits could be obtained by enabling interoperability between platforms. A vast ecosystem has emerged around online job platforms. We begin by creating a taxonomy of such platforms and discuss how these functionalities could be improved and integrated.

4.1 Taxonomy of Online Job Platforms and Ecosystems

So far, the prior work has extensively discussed online job platforms such as AMT or CrowdWorks. However, there is a thriving ecosystem built around these platforms. They often add features missing in the original platforms or provide value added services. In order to improve platform design, it is important to understand this ecosystem.

Job Platforms. For the sake of completeness, we include online job platforms in the taxonomy. These are the platforms such as AMT, CrowdWorks and others where requesters post tasks and workers complete them *online*. These could be generic platforms where a wide variety of tasks are performed or specific ones such as ride-hailing that perform a single task.

Platforms for Worker Communication. Most of the current platforms do not support communications between workers. Some times, this is desirable to avoid inter-worker collusion. Most of the time, this is very limiting as collaborative tasks become more and more important. Furthermore, there are also many scenarios where workers might want to communicate. These include providing feedback about requesters, providing tips and so on. So there is an emerging set of platforms to enable such communication. The most popular of these are Turkopticon [6] and TurkerNation (which closed in 2018). Turkopticon has almost 500K reviews of around 50K requesters. It also provides granular way to provide requestor feedback on fairness, pay quality and speed of payment. While these are a good start, both Turkopticon and TurkerNation were specific to AMT. It is important

[2] `http://headwork.gforge.inria.fr`

Support for...	Functionalities	Platforms for...					
		Providing Jobs	Worker Communication	Worker/ Requester Profiles	Task-Worker Matching	Workflow management	Worker Training
Worker	Ease of On- and off-boarding	◯		◯			
	Learning Skills		◯				◯
	KB for workers					◯	
	Worker Preference Specification	◯		◯			
	Ease of choosing tasks	◯	◯	◯	◯		
Requester	Task Requirement Specification	◯					
	Complex Workflow					◯	
	Sophisticated Assignment Algorithm				◯	◯	
	AI Worker Support			◯	◯	◯	
	Total Optimization					◯	◯
	Algorithm Butiques				◯	◯	
	Total Optimization					◯	◯
	Bespoke Platforms	◯		◯			◯
	Workflow and AI Worker Library			◯		◯	

Figure 4: Relationship between functionalities and platforms for FoW. This suggests that cooperation between different platforms will be the key to make use of the full potential of ecosystem of FoW platforms.

that most of these platforms provide a mechanism for worker communication. This could be done within each platform or by providing a generic website where workers from different platforms can congregate.

Platforms for Persistent Worker/Requester Profiles. Currently, each online job platform has an internal mechanism for measuring worker/requestor reputation. This could be as simple as approval rate for workers and requesters. It could also be more granular such as how many tasks a worker has completed for each skill. It is important to have persistent profiles for both workers and requesters so that they can build their reputations based on their entire body of work. Currently, if an experienced worker moves from one platform to another, she has to start from scratch. Hence, it is important to have an independent website where workers/requesters can create a profile and share all relevant information in a verified manner.

Platforms for Matching Workers and requesters. The explosion of online job platforms makes it harder for both workers and requesters. Productive workers and fair requesters are often scattered across platforms. Each worker has to login to multiple platforms to find useful work that quickly becomes tedious. It is desirable to have platforms that could match workers and requesters across multiple platforms. For example, the worker could specify task preferences and the matching platform will scour other online job platforms and notify the worker when relevant tasks become available. This could also benefit requesters by finding productive workers across platforms. While there are some preliminary effort such as by CrowdWorks or Figure Eight that act as a layer on top on AMT, there is still lot of work to be done.

Towards Better and Fairer Crowdsourcing Platforms. Recently, there has been increasing interest in improving crowdsourcing. Some of these efforts include providing guidelines for requesters for specific platforms. Popular examples include Dynamo Guidelines [12] for academic requesters and FairCrowdWork.org. These often help well-meaning requesters to treat workers properly by providing clear instructions about the task and fair pay. FairCrowdWork.org provides advice to workers based on their rights. They have extensive survey involving 95 questions that allows workers to rate various crowdsourcing platforms. Specifically, they quizzed workers on 8 topics including demographics, work experience in a platform, payment details, communication with requesters and platforms, reliability of the platform, quality of available tasks and miscellaneous information. Based on the response of the workers, each platform is graded in a scale between 1-5.

4.2 Platform Interoperability

Given the plethora of platforms, it is important they are interoperable to ensure that workers and requesters have a wide variety of options. Currently, interoperability is nearly non-existent. However, we believe that this will become more and more important as an increasing proportion of workers join these online job platforms.

There are a number of dimensions in which interoperability must be achieved. Workers should be able to move between platforms without losing any of their skills/ratings etc. This will require the creation of a unified schema of skills. This could be different for each domain. Once this is done, each platform has to fill the worker skills according to this schema. Of course, each platform could have their own additional proprietary list of metric. Nevertheless, it must be possible to create a worker "profile" that gives a holistic perspective of the worker skills in that domain regardless of how many platforms the worker is involved with. Interoperability must also be possible from a requestor perspective. Given a task, it must be possible to recruit workers from multiple platforms. For example, if a requestor needs 10 qualified experts, it must be possible to recruit 4 from Platform 1, 5 from Platform 2 and 1 from a third one. Similarly, it is important to create a requester "profile" that is persistent across multiple platforms so that the worker has a holistic perspective of the requester such as the rate of rejection of tasks.

We can see that achieving interoperability is quite challenging. This requires a commonly agreed standard of skills, which, due to the richness and specialties of the many disciplines that exist, is probably difficult to achieve. For platforms to exchange information, one must design a common, unifying API that will fit for all types of applications.

Standardized API and Schema: The diversity of the crowdsourcing and FoW platforms makes standardization much challenging. However, it is possible that specific components such as task creation, worker selection, truth inference, worker skill estimation could be more amenable for standardization. The creation of the standard API would boost the adoption of crowdsourcing by bringing in more enterprises. Entrepreneurs could also create innovating applications for specific crowdsourcing tasks. This is akin to the explosion of smartphone apps once iOS and Android provided a standardized API functionality. For example, one could identify an important tasks such as Entity Resolution (finding tuples that represent the same real-world entity) and use the API of the platforms to recruit the crowd, perform the tasks and provide the results. There will also be standardization in terms of how metadata about workers and tasks are stored [13, 14].

Interchange Format: There is a need to have a pre-defined interfaces and Schema for data exchanges between platforms. It should provide a simple agreed model that has enough expressive power to achieve total optimization. At the minimum, one must be able to transfer the human factors of workers (such as skill, past tasks, etc) and that requesters (task approval rate, etc). This allows both these types of users to migrate to a different platform without being locked in.

Interchange Queries: When data is not large and is allowed to be exported, platforms can exchange the whole data in the interchange format, but there are cases where this is not a practical solution. An effective approach is to develop common sets of interchange queries so that we can identify only small subsets of data to be exchanged.

5 Conclusion

Online job platforms are becoming increasingly popular. They provide an alternative job arrangement that has the potential to dramatically affect the Future of Work. An increasing proportion of human workforce will be employed in such platforms. Hence, it is very important to study the problem of platform design. In this paper, we have a preliminary attempt at investigating this problem. We provide a taxonomy of platforms and what are the major missing functionalities for both workers and requesters. We also advocate the need for interoperability between platforms. This allows workers and requesters to move their data between platforms without getting locked into a single platform. There are a number of interesting research and policy challenges in achieving the vision of platform interoperability.

References

[1] S. Basu Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal—The International Journal on Very Large Data Bases*, 24(4):467–491, 2015.

[2] A. M. Brawley and C. L. Pury. Work experiences on mturk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior*, 54:531–546, 2016.

[3] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.

[4] C.-J. Ho and J. W. Vaughan. Online task assignment in crowdsourcing markets. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[5] K. Ikeda, A. Morishima, H. Rahman, S. B. Roy, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Collaborative crowdsourcing with crowd4u. *Proceedings of the VLDB Endowment*, 9(13):1497–1500, 2016.

[6] L. C. Irani and M. Silberman. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 611–620. ACM, 2013.

[7] P. Mavridis, D. Gross-Amblard, and Z. Miklós. Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 843–853, 2016.

[8] S. McFeely and R. Pendell. What workplace leaders can learn from the real gig economy. gallup, 2018.

[9] A. Morishima, S. Amer-Yahia, and S. B. Roy. Crowd4u: An initiative for constructing an open academic crowdsourcing network. In *Second AAAI conference on human computation and crowdsourcing*, 2014.

[10] A. Morishima, N. Shinagawa, T. Mitsuishi, H. Aoki, and S. Fukusumi. Cylog/crowd4u: A declarative platform for complex data-centric crowdsourcing. *Proceedings of the VLDB Endowment*, 5(12):1918–1921, 2012.

[11] H. Rahman, S. B. Roy, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Task assignment optimization in collaborative crowdsourcing. In *2015 IEEE International Conference on Data Mining*, pages 949–954. IEEE, 2015.

[12] N. Salehi, L. C. Irani, M. S. Bernstein, A. Alkhatib, E. Ogbe, K. Milland, et al. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1621–1630. ACM, 2015.

[13] D. Schall. *Service-Oriented Crowdsourcing - Architecture, Protocols and Algorithms*. Springer Briefs in Computer Science. Springer, 2012.

[14] D. Schall, B. Satzger, and H. Psaier. Crowdsourcing tasks to social networks in bpel4people. *World Wide Web*, 17(1):1–32, 2014.

On Benchmarking for Crowdsourcing and Future of Work Platforms

Ria Mae Borromeo
Philippines Open University
rhborromeo@up.edu.ph

Lei Chen
HKUST
leichen@cse.ust.hk

Abhishek Dubey
Vanderbilt University
abhishek.dubey@vanderbilt.edu

Sudeepa Roy
Duke University
sudeepa@cs.duke.edu

Saravanan Thirumuruganathan
QCRI, HBKU
sthirumuruganathan@hbku.edu.qa

Abstract

Online crowdsourcing platforms have proliferated over the last few years and cover a number of important domains, these platforms include from worker-task platforms such Amazon Mechanical Turk, worker-for-hire platforms such as TaskRabbit to specialized platforms with specific tasks such as ridesharing like Uber, Lyft, Ola etc. An increasing proportion of human workforce will be employed by these platforms in the near future. The crowdsourcing community has done yeoman's work in designing effective algorithms for various key components, such as incentive design, task assignment and quality control. Given the increasing importance of these crowdsourcing platforms, it is now time to design mechanisms so that it is easier to evaluate the effectiveness of these platforms. Specifically, we advocate developing benchmarks for crowdsourcing research.

Benchmarks often identify important issues for the community to focus and improve upon. This has played a key role in the development of research domains as diverse as databases and deep learning. We believe that developing appropriate benchmarks for crowdsourcing will ignite further innovations. However, crowdsourcing – and future of work, in general – is a very diverse field that makes developing benchmarks much more challenging. Substantial effort is needed that spans across developing benchmarks for datasets, metrics, algorithms, platforms and so on. In this article, we initiate some discussion into this important problem and issue a call-to-arms for the community to work on this important initiative.

1 Introduction

Online crowdsouring platforms have become an unavoidable part of our life. The breadth of these platforms is truly staggering. These include the widely studied crowdsourcing platforms such as Amazon Mechanical Turk (AMT) to worker-for-hire platforms such as TaskRabbit. Even platforms such as Uber that provides ride-hailing services fall under this category. These alternate task arrangements have been inexorably growing over the last few years. The crowdsourcing community has done tremendous work in developing algorithms that enabled productive use of workers/volunteers in important domains such as Wikipedia, data annotation for machine

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

learning, disaster analysis and so on. Hence, it is important that our community also takes the lead research on the future of work (FoW) on these online crowdsourcing platforms. While developing algorithms for specific components is important, it is much more important to take a holistic perspective and develop a framework to evaluate the research on this topic. In this article, we advocate the need for developing benchmarks for crowdsourcing and future of work.

Importance of Benchmarks: Benchmarks and standardization are an underappreciated type of research that has the potential to dramatically boost a research domain. Often, benchmarks have a galvanizing effect on a community by focusing them on the most important things. This is especially important for emerging areas such as crowdsourcing/FoW. There are a number of success stories on how developing benchmarks led to the blossoming of a young field. Two major examples are the fields of databases and deep learning. In databases, TPC benchmark provides a reference environment in which major activities of customers (such as transactions or decision support) can be evaluated. The development of ImageNet benchmark dataset is widely credited as the standard benchmark data to evaluate various of deep learning algorithms on image classification and object identification. Benchmarks allow researchers to fairly evaluate competing algorithms and spurs innovation.

Need for Benchmarks in FoW: Online crowdsourcing platforms have been reshaping the workforce. A number of studies such as [12] report that 36% of US workers have an alternative work arrangement in some capacity. This number has substantially increased over the last few years. Research on FoW is still at its infancy and it is exactly the time to develop benchmarks to ensure that the energy of the research community are spent on the major priorities. There has been extensive work in crowdsourcing with hundreds of papers being published every year by researchers in domains as diverse as computer science, psychology, social science, management and so on. While this allows for cross pollination of ideas in the best case, it could also cause duplication of work and development of mutually incompatible outcomes. Benchmarks have the potential to mitigate such efforts. There are some open data sets in crowdsourcing research such as Figure Eight's Data for Everyone [1], which contains information about the task (instructions, questions, and answers). There have also been efforts to create a directory of crowdsourcing data sets from various sources [11]. However, the lack of common metrics and reference implementations makes the problem of identifying most promising ideas and evaluating competing implementations very challenging.

Challenges: Developing a benchmark is challenging even in a rigorously empirical field such as databases where the fundamental metrics for performance are much more well understood. The biggest strength of crowdsourcing/FoW research, namely its diversity, also causes the biggest challenge. Unlike databases, FoW could span across many diverse domains and each field could have different metrics and requirements. Any benchmark should be able to reflect this fundamental property. Additionally, FoW is uniquely positioned due to the presence of multiple stakeholders – workers, requestors and platforms – who could have different objectives. Finally, FoW is driven by humans and it is important to take human and social factors into account. The presence of these volatile human factors [17] that are different for each human makes developing uniform benchmarks much more challenging.

2 Taxonomy for Benchmarks

Prior approaches such as developing a large dataset such as ImageNet or a synthetic workload such as TPC are clearly insufficient since we need not only a dataset but also a set of metrics to measuring the effectivenss of the crowdsourcing platforms. FoW necessitates a fundamental rethink on how benchmarks must be designed. One of our key contribution is the identification of major dimensions in which benchmarks must be developed. These include:

- *Metrics.* Crowdsourcing/FoW has a number of different components such as task assignment, collecting and aggregating results from the workers, evaluating worker skills and so on. It is important to develop metrics so that crowdsourcing algorithms for these compontents can be fairly evaluated according to these

metrics. Metrics must measure both accuracy and efficiency. These should also reflect the requirements of the stakeholders. Almost all of the current metrics are focused on requestors – and it is important to flesh out metrics for workers that could include human factors [17] such as job satisfaction and worker fairness.

- *Datasets.* There is a paucity of datasets that could be used for crowdsourcing research. These include all facets such as task assignment [3, 8], identifying ground truth from imperfect responses [19], learning skills of workers [15], identifying spammers [16] and so on.
- *Platform Simulations and Synthetic Datasets.* In order to evaluate algorithms for platform management tasks (such as matching workers to tasks), it is important to have a realistic data that could be used to model workers and tasks. These could include their arrival rates, worker demographics and preferences, task characteristics and requirements and so on. By using these information, it is possible to evaluate task assignment algorithms holistically. Of course, these data are often proprietary – so even a realistic looking synthetic data could dramatically improve research.
- *Reference Implementations:* By standardizing the settings, one could develop reference implementations of various algorithms. While there are some early efforts in this direction for truth inference [19], substantial additional work needs to be done. These reference implementations allow a researcher to confidently claim that their algorithm is better than current state-of-the-art.
- *Open Source Online Crowdsourcing Platforms.* Currently, Amazon Mechanical Turk is one of the most popular crowdsourcing platform. However, no open source clone of it exists. It is important to have an open source academic platform that could be used to prototype FoW algorithms. As an example, the development of Postgres triggered an avalanche of research whereby an individual researcher can plug-in their algorithm for specific tasks such as query optimization without implementing an database from scratch. Currently, a researcher working on task assignment has to build a simulator for a crowdsourcing platform to evaluate her algorithm. The presence of a modular open source framework allows one to just modify a single component. For example, when the authors implemented an algorithm for task assignment for collaborative crowdsourcing [14], they were able to implement and evaluate it on the academic platform Crowd4U [9]. Given another example, when authors want to evaluate various of task assignment algorithms [18], they could compare them on the same spatial crowdsourcing platform, gMission [5].
- *Competitions for FoW Tasks.* The Natural Language Processing community has a tradition of conducting yearly competitions (such as SemEval or TREC) for making progress on major research challenges. Such events trigger the competitive nature of researchers who strive to beat prior state-of-the-art. It is important that our community also adopt this important tradition. Every year, one could identify major FoW tasks in important domains such as Citizen science or disaster crowdsourcing and seek to make meaningful progress.

3 Benchmarking Metrics

As mentioned before, the diversity of the tasks and the presence of multiple stakeholders makes the process of designing a comprehensive set of metrics very challenging. In this section, we make an initial attempt in identifying a set of relevant metrics that are relevant for FoW research.

Desiderata for FoW Metrics. Metrics are the primary mechanism by which the performance of an FoW platform is evaluated. Crowdsourcing has been used in a number of diverse domains – so the metrics must be both generic and comprehensive. Currently, most of the metrics are focused on the requestors. It is important to design metrics that take into account the needs of all three major stakeholders – workers, requestors and platforms. It should be able to handle both quantitative (such as computational criteria) and qualitative aspects (such as human and social factors).

3.1 Metrics for FoW Platforms

We begin by describing few metrics that could be used to quantify any given FoW platforms.

Crowd Size, Diversity and Rate of Participation. Workers are an indispensable part of any FoW platform. So one of the most basic metrics for the platform measures the size of the crowd. Typically, requestors would prefer a platform with more workers as it gives a wider pool for recruitment. Of course, size only gives an incomplete perspective of the platform. The diversity of the crowd is often a better indicator of the quality of worker pool especially for knowledge intensive crowdsourcing tasks. A diverse crowd with different background and perspectives often results in more creative solutions. Finally, one metric that is useful to measure how thriving a platform is participation rate that measures the number of workers who are active and perform tasks in a given unit of time such as a week or a month. In a thriving FoW platform, one would expect that this number will be large. An alternate metric could measure the average amount of time that is spent by workers on the platform.

Worker Skill Distribution. Another key aspect of the FoW platform is the skill distribution of the workers. In a generic FoW platform such as AMT, each worker and task could be annotated with a set of skills such as translation, writing, comprehension and so on. Ideally, requestors would prefer a platform with workers having diverse skills. Another key aspect is the alignment between the skill distribution available in the worker pool and the distribution required by the task pool. Such a misalignment often results in the frustration of both workers and requestors.

Task Diversity and Complexity. Similarly, it is important to measure the distribution of the tasks themselves. It is important for the platform to have a wide variety of tasks involving different skills and varying complexity ranging from easy to difficult. A steady stream of monotonous or complex tasks could reduce worker motivation and result in turnover.

Efficiency Metrics: Tail Latency and Throughput. The FoW platform could have a number of quantitative metrics that measure its performance. Two of the key metrics are latency and throughput. Latency measures the time taken between posting of a job and its completion. Of course, some latency is inevitable due to the inherent nature of humans. However, a large latency would preclude certain tasks that require interactive responses from being posted on the platform. In addition to the mean and median latencies, it is also important to measure the tail latencies corresponding to the 95-th or 99-th percentile of task latencies. Similarly, it is also important to measure throughput which could measure the number of tasks completed in any given unit of time. Throughput could also be used to evaluate specific algorithms such as task assignment wherein it measures the number of tasks for which workers were matched with.

Reliability and Robustness to Adversaries. Any major FoW platform attracts a wide variety of adversaries such as workers who are scammers out to make a quick buck by possibly colluding with other workers. This could also include requestors who either maliciously reject tasks completed by workers so as to not pay them or those who use crowdsourcing for illegal or unethical purposes. It is important that the platform has sufficient mechanisms so that it is robust against such adversaries. Crowdsourcing is increasingly being used for major tasks and it is important that the platform is reliable and does not crash.

Usability of the FoW Platform Interface. An under-appreciated aspect of FoW platforms is how user friendly the interface is. This is applicable to both the workers and requesters. Any large FoW platform attracts a diverse group of requestors and workers who may not be fluent in how the platform works. For example, the requestors could be domain scientists such as psychologists or social scientists with limited knowledge of computer science. Similarly, workers could have a wide variety of educational levels. Hence, it is important that the platform's interface is intuitive and allows both requestors and workers to complete the tasks efficiently.

3.2 Metrics for FoW Workers

Workers are a key part of the platform and its success hinges on learning appropriate information about the workers and use it effectively so that all the stakeholders are satisfied. We enumerate below a number of key

facets of human workers that must be systematized and measured. Such metrics have the potential to represent the worker holistically than the current approach of quantifying the worker simply as a number based on the task approval rate.

Human Factors. It is important to model the behavior or characteristics of human workers in any crowdsourcing platform. This has a number of applications such as assigning appropriate workers to tasks to ensure their completion and recommending appropriate tasks to workers to increase job satisfaction. Prior work [2, 17, 7] typically involve identifying appropriate human factors, integrating them into FoW components such as task assignment and estimating them from past interactions with FoW platforms. For the reminder of the section, we discuss the major human factors. Systematically formalizing them and integrating them holistically into FoW platforms is a major open challenge.

Skill, Knowledge and Expertise. Most generic crowdsourcing platforms such as AMT could have a set of domains $D = \{d_1, d_2, \ldots, d_m\}$ that denote the various knowledge topics. Consider a task of translating documents from English to French. Even this task requires multiple skills such as comprehension of English language, writing and editing in French. Generic platforms have a wide variety of task types with large platforms such as Crowdworks supporting as much as 200 types of tasks. Given this setting, it is important to enumerate the set of skills needed by the tasks and possessed by the workers. One could quantify the skill using categorical values (not knowledgeable, novice, knowledgeable and expert) or in a continuous $[0, 1]$ scale. So, a value of 0 for a specific skill such as English comprehension denotes no expertise while the value of 1 could denote complete mastery.

Worker Motivation. This is one of the most important human factors and a key component of the worker to be measured. Understanding worker motivations could be used to improve the performance of the FoW platform through a more informed matching of workers with tasks. However, understanding what motivates workers is not so obvious. Some workers could be motivated by monetary compensation while others are motivated by fun and enjoyment. It has been found that prior work social studies on workplace motivation is also applicable to FoW platforms [10, 13]. In fact, there are as many as 13 factors that are highly relevant for worker motivation [10]. These could be categorized as intrinsic and extrinsic motivation.

Intrinsic motivation include aspects of tasks such as [10] skill variety (preference for tasks requiring a diverse collection of skills), task identity (the worker perception about the completeness of the task), task autonomy (the degree of freedom allowed during the task), feedback during the task completion and so on. Extrinsic motivation includes monetary compensation, human capital advancement and signaling (performing a task to give a strategic signal to environment).

Metrics for Group based Human Factors. The increasing popularity of knowledge intensive crowdsourcing tasks requires collaboration. Hence, it is important to measure the various factors that are relevant to modeling worker collaboration. The most important of those is worker-worker affinity that measures the collaborative effectiveness of any two workers. This could be extended to measure the social cohesiveness of any group of workers. It is often desirable to form a group of workers where the aggregate pairwise affinity is large. Of course, there is a natural diminishing returns when increasing the group size beyond certain threshold dubbed critical mass.

3.3 Metrics for FoW Tasks

Tasks (and requestors) form the final leg of FoW platforms. A number of metrics described for platforms are also applicable for tasks.

Accuracy and Quality. This is often the most important metric for the requestors. If the responses of the workers are accurate, then most popular approaches for aggregating worker responses will provide accurate results. It is important for the requestor that the completed crowdsourcing tasks have a high accuracy rate.

Cost. The requestor often wants to complete a given task with high accuracy while minimizing the monetary payment or the worker effort. There has been extensive work on identifying appropriate workers while satisfying

the quality and cost constraints of the tasks. The requestor has a natural cost-benefit tradeoff and higher cost often deters them without the requisite quality.

Completion Rate and Latency. The human workers create an uncertainty in terms of task completion. It is possible that a worker accepts a task but does not complete it immediately. This creates a straggler effect where some workers could delay the completion of the tasks. This is especially important for longer tasks where workers losing motivation is a key risk factor. This affects the requestor in two ways. First, any task consist of a number of micro-tasks all of which need not be completed. Higher the completion rate, the better it is for the requestor. Second, for complex tasks that often have a binary outcome (task completed or not), it could dramatically increase the latency. Having systematic method to measure these two phenomenon is quite important.

Fairness Related Metrics. There has been intensive research on how to quantify fairness. In our context, it is important to ensure that the platform and the task requestor are seen as fair. For example, workers who did similar tasks should be paid similar compensation. Similarly, the worker submissions must not be rejected without a valid reason and so on.

4 Benchmarking Data

With the proposed metrics, the next task is to design Benchmarking data to test the effectiveness of the FoW platforms. Existing works [18, 6, 4] often use real data set to test. However, it is hard if it is not impossible to derive statistic information of real data. Without knowing the characteristic of real data, it will be difficult to choose the right real data to test. Moreover, even we use all the real data sets, we still do not know if we have tested all the possible worst case scenarios, the robustness of the platform is still unknown. Thus, in this section, we mainly discuss how to design synthetic benchmarking data of workers and tasks to test the FoW platforms.

4.1 Benchmarking data for workers

There are many factors we should consider to design benchmarking data of workers, which are listed as follows.

Worker's expertise. As we can observe from a crowdsourcing platform, given the same question/task, different workers may offer different answers. This is because different workers often have different level of expertise in different domains. Thus, when we design benchmarking data of workers, we should assign workers into different categories (domains) with different expertise levels. Moreover, different category and expertise level distributions should be generated.

Worker's preference. Worker's preferences towards different types of tasks also determine the workers' willingness to accept the tasks. For example, some workers prefer image labelling tasks to language translation tasks, if both types of tasks are available in the platform, with a very high probability, they will choose the image labeling tasks. Thus, benchmarking data of workers should take worker's preferences (categories of tasks) into consideration. Again, we should generate different distributions for category preference of workers.

Worker's activeness. Given the same crowdsourcing platform, some workers are quite active and solve many tasks/question within a short period of time, while some only solve a few questions but spend quite long time. Therefore, to generate benchmarking data of workers, we can use the number of tasks completed with a specified period time as the activeness factor and simulate it with different distributions.

4.2 Benchmarking data for tasks

Similar as workers, for tasks, we also need to consider different factors when we want to generate benchmarking tasks.

Task's category and difficulty. Given a crowdsouring platform, there are many different types of tasks. For the same type of tasks, the difficulties are also different. For example, given an image, an image classification task

is much easier compared to an object identification task. Therefore, we need to consider categories and difficulty levels to generate tasks with different distributions.

Task's budget. Given the same tasks with different budgets, the task with a higher budget is often accepted much faster than the one with a lower budget. Of course, this does not indicate that tasks with higher budget will get higher quality answers. We need to generate budgets with different distributions for the tasks.

Task's completion time. When the requester posts a task on the crowdsourcing platform, she often sets up a completion time, which is the time that the requester expects the answer back. Depends on the urgency of the tasks, different tasks often have different specified completion time. We need to generate completion time with different distributions for the tasks.

Tasks' required number of answers. For some tasks, the requester only needs one answer, such as simple true/false questions, while for complicated tasks, such as object identification, more workers are needed to verify the correctness of labelled objects. Thus, we need to take tasks' required number of workers as another factor to generate task data.

5 Research Directions

Creating a comprehensive benchmark for online job platforms with complex interactions among human workers and requesters, machines, and AI agents is a non-trivial task. Here we list some challenges that need to be addressed when we design such a benchmark.

Broad applicability for use in academia and industry. The benchmarks designed for evaluating the platforms should be able to cover different types of works, platforms, and workers. Real applications have different requirements, especially for many real industry applications, therefore, developing a "leaderboard benchmark" with single dataset and single application may not work. Instead, we need to support tailored benchmarks for different workloads and have sufficient diversity in the requirements within the benchmark to satisfy the needs of different platforms, especially from industry (e.g., as done in TPC benchmarks used to evaluate performances of query evaluation supporting different workloads).

Acceptability in Research Communities. One practical challenge after designing a benchmark is to getting the benchmark widely accepted in the research community and industry. This would need wider discussions and involvement of all the relevant players like the crowdsourcing platforms from industry and academia, research communities, and also possibly the workers and requesters who use these platforms.

Incorporating Metrics capturing human factors that are not easy to measure. One can be seen that a number of metrics like latency, throughput, or cost are easy to measure (although can be measured in multiple ways). On the other hand, some metrics, especially the ones related to human factors such as fairness, equity, and satisfaction, are difficult to measure. Incorporating ideas from the recent advances in research on fairness and ethics from data management, ML/AI, and also non-computational domains like psychology, cognitive science, sociology, laws, and policy would be useful.

Handling multiple criteria and optimizations: Satisfying multiple metrics all at the same time may lead to multi-objective optimization problems, which are typically computationally intractable. As in computational challenges, formulating the problems meaningfully, designing efficient approximation algorithms with formal guarantees, and also developing efficient tools that produce good results in practice will be required. On the other hand, one can explore whether keeping all the criteria separate and giving individual scores to those criteria work better.

Supporting interactions with AI agents: Since AI agents are used in different stages in an online job platform, we need to develop benchmarks that enable comparison of effectiveness and interactions of AI agents and humans on jobs of various categories. Such a benchmark will have to continuously updated as AI technology improves.

Measuring robustness: Measurement of robustness of the platform or of the matching algorithm used in

the platform would require creation of adversarial benchmarks that will enable assessment of effectiveness of integrity checks and anomaly detection mechanisms built into the platforms. For instance, if an adversarial requester attempts to ruin the reputation of a worker with poor ratings, or if a worker attempts to game a system by providing wrong inputs, a good platform would be able to detect or defect such attempts by using other information collected in the process, which might be one of the desired properties of the benchmark.

Reproducibility: Results that use standard benchmarks used in different contexts (e.g., TPC-H or TPC-DS data for evaluating scalability in data management) are repeatable or reproducible. On the other hand, tests involving humans are not always repeatable benchmark tests results could vary for every test run. Allowing some level of differences in the results and considering mean/variances might be needed while developing the benchmarks.

Synthetic benchmark data generation: Based on the data distribution of data logs, such as availability of workers/jobs and dynamics of the platform, collected from real job markets, how can we generate synthetic benchmark data which follows the same data distribution under the similar environment scenarios to test various aspects of users, jobs and platforms, such as effectiveness and scalability.

6 Conclusion

Alternative job arrangements such as online job platforms are becoming increasingly important and a major source of employment in the near future. It is important for the crowdsourcing community to take the lead on researching the Feature of Work. In this article, we argue that a key pre-requisite is the creation of benchmarks for such research. The diversity of crowdsourcing research by various communities is a key challenge. We propose a taxonomy of dimensions for which benchmarks has to be developed. We also have enumerated a list of metrics that are most relevant for effective crowdsourcing. Moreover, we list essential factors that need to be considered during the process of creating benchmarking data to test the effectiveness and robutness of crowdsouring platforms. Benchmarks have had a dramatic impact in the development of various domains. We issue a call-to-arms to the crowdsourcing community to seize the opportunity!

References

[1] Figure eight's data for everyone. `https://www.figure-eight.com/data-for-everyone/`. Accessed: 07-Dec-2019.

[2] S. Amer-Yahia and S. B. Roy. Human factors in crowdsourcing. *Proceedings of the VLDB Endowment*, 9(13):1615–1618, 2016.

[3] S. Basu Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal—The International Journal on Very Large Data Bases*, 24(4):467–491, 2015.

[4] Z. Chen, P. Cheng, Y. Zeng, and L. Chen. Minimizing maximum delay of task assignment in spatial crowdsourcing. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 1454–1465, 2019.

[5] Z. Chen, R. Fu, Z. Zhao, Z. Liu, L. Xia, L. Chen, P. Cheng, C. C. Cao, Y. Tong, and C. J. Zhang. gmission: A general spatial crowdsourcing platform. *PVLDB*, 7(13):1629–1632, 2014.

[6] P. Cheng, L. Chen, and J. Ye. Cooperation-aware task assignment in spatial crowdsourcing. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 1442–1453, 2019.

[7] E. Cullina, K. Conboy, and L. Morgan. Measuring the crowd: a preliminary taxonomy of crowdsourcing metrics. In *Proceedings of the 11th International Symposium on Open Collaboration*, page 7. ACM, 2015.

[8] C.-J. Ho and J. W. Vaughan. Online task assignment in crowdsourcing markets. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[9] K. Ikeda, A. Morishima, H. Rahman, S. B. Roy, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Collaborative crowdsourcing with crowd4u. *Proceedings of the VLDB Endowment*, 9(13):1497–1500, 2016.

[10] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, pages 1–11. Detroit, Michigan, USA, 2011.

[11] G. Li, J. Wang, Y. Zheng, and M. J. Franklin. Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2296–2319, 2016.

[12] S. McFeely and R. Pendell. What workplace leaders can learn from the real gig economy. gallup, 2018.

[13] D. Pilz and H. Gewald. Does money matter? motivational factors for participation in paid-and non-profit-crowdsourcing communities. *Wirtschaftsinformatik*, 37:73–82, 2013.

[14] H. Rahman, S. B. Roy, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Task assignment optimization in collaborative crowdsourcing. In *2015 IEEE International Conference on Data Mining*, pages 949–954. IEEE, 2015.

[15] H. Rahman, S. Thirumuruganathan, S. B. Roy, S. Amer-Yahia, and G. Das. Worker skill estimation in team-based tasks. *Proceedings of the VLDB Endowment*, 8(11):1142–1153, 2015.

[16] V. C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13(Feb):491–518, 2012.

[17] S. B. Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Crowds, not drones: modeling human factors in interactive crowdsourcing. 2013.

[18] Y. Tong, J. She, B. Ding, L. Chen, T. Wo, and K. Xu. Online minimum matching in realtime spatial data: Experiments and analysis. *Proceedings of the VLDB Endowment*, 9(12):1053–1064, 2016.

[19] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.

Ethical Challenges in the Future of Work

Pierre Bourhis
CNRS at CRIStAL
pierrebourhis@univ-lille1.fr

Gianluca Demartini
University of Queensland
demartini@acm.org

Shady Elbassuoni
American University of Beriut
se58@aub.edu.lb

Emile Hoareau
University Grenoble Alpes
Emilie.Hoareau@grenoble-iae.fr

H. Raghav Rao
The University of Texas at San Antonio
hr.rao@utsa.edu

Abstract

The rise of self-employment empowered by platforms such as Amazon Mechanical Turk and Uber has drastically changed our perception of work. The possibility to link requesters and workers from all over the world in a scalable manner has resulted in advancements in the work world that would not have been possible otherwise. However, many ethical concerns related to fairness, transparency, and bias regarding these new forms of work have also been raised. In this paper, we present our vision on these ethical issues, how they can be combated in the future of work, and how this will impact the data management research community and future work platforms.

1 Introduction

The rise of online piecework platforms such as Amazon Mechanical Turk[1] and Uber[2] are profoundly changing the way we conduct work. These platforms typically link requesters to workers from all over the world in a scalable manner, providing the opportunity to accomplish work in an unprecedented way. However, while these transformations offer numerous opportunities to both stakeholders, they also bring threats especially for workers, who are more vulnerable against requesters' demands and platforms' setups. The traditional work environment is highly regulated through law to avoid power imbalance and thus abuses from the more powerful actor, the employer, can be avoided. Online work platforms remain, at this time, less subject to strong regulations, which could then foster unfair situations. Several studies have already pointed out the potential drawbacks of the practices of online work platforms especially towards workers [10]. The different roles and activities offered by online work platforms (e.g., requesters, contract workers, micro-tasking, etc.) make the creation of an ethical and fair work environment challenging. Indeed, while the main role of a platform is to allocate workers to available jobs, it also needs to ensure work payments and the security of both requesters and workers.

In the future, work facilitated by online platforms may become unfair or harmful in the absence of a strong and deep ethical reflection. An ethical approach in this context is defined as stakeholders in the "loop of work" doing the "right things" based upon rationally justifiable standards. It is important to ensure that any artifact in

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

[1] https://www.mturk.com/
[2] https://www.uber.com/

the loop should not be used in ways that can harm people, the environment, and society. As such, each artifact would have to follow certain legal, regulatory, and ethical frameworks. It is important to ask questions about the relationship between humans and computational entities in terms of what are the implications of replacing and supplementing humans with machines, and also try to foresee both utopian and dystopian future-of-work scenarios in order to be prepared for both an optimistic future as well as a pessimistic future. The role of work platform in the relationship between the job providers and the workers has to be better defined and regulated as compared to the current situation. Indeed, platforms could play several roles: putting in relation requesters and workers, transferring payments between them, ensuring a virtual platform where to do work. These roles have to be defined precisely and need to follow relevant legal obligations. The notion of work contract in this context becomes complex and often not well defined, as it puts in relation three parties (i.e., job providers, workers, and platforms) and in different contexts (e.g., online and offline). Requesters, workers, and platforms can also be in different countries and the question of which work legislation should apply is complex.

In this paper, we point out that practices of online work platforms may raise ethical issues which must be handled to aspire for a sustainable Future of Work (FoW). We focus on issues related to privacy, accountability, fairness (with regard to compensation and job allocation for instance), transparency and explainability. We offer a detailed view of each of them, presenting existing research and suggestions about future directions.

We first detail them, presenting how they appear today and how they are addressed by existing scientific literature. Then, we suggest a path forward, for each of them. We focus on ethics in this case, as it would apply to the stakeholders consisting of the "worker", the "requester" and the "platform". The rise of online work platforms raises several major ethical issues: (i) *privacy/access control* - ensuring that the use of the personal data is not used for harmful purposes; loss of personal information can result in negative consequences which could have an effect on physical and mental health and hence access control becomes an important tool for the protection of privacy, (ii) *accountability* - identification of outcomes and procedures and association with the appropriate entities, with the goal of not violating system or organizational policies and rules for compliance, (iii) *fairness* - in terms of equity or equality of decision outcomes such as compensation to workers, (iv) *transparency* of decision making processes including rejected job applications, decision on task assignment, and on reward allocation, (v) and *explainability* of processes involved. These last two focus on the degree of justification and truthfulness regarding explanations and information provided in the process of work as well as the rewards and the assignments.

2 Related Work

Several researchers have started to explore online work platforms especially in the context of crowdsourcing. Crowdsourcing involves the distribution of tasks to various people, often across the world, via the Internet. The strength of crowdsourcing is that a variety of skills are instantly available and work can be conducted online in any language. From the point of view of these piecework workers, while some may be working at hours when they are not in their official work capacity, in order to make some extra money, for others, like Uber drivers, the work may actually be a full time job. In this section we review some of the related work regarding work condition and ethics of crowdsourcing.

Deng et al. [4] explore microtask crowdsourcing as perceived by crowd workers, revealing their values as a means of informing the design of such platforms. Analyzing detailed narratives of 210 crowd workers participating in Amazon's Mechanical Turk, they uncovered a set of nine values they share: access, autonomy, fairness, transparency, communication, security, accountability, making an impact, and dignity. They found that these values are implicated in four crowdsourcing structures: compensation, governance, technology, and microtask. Two contrasting perceptions—empowerment and marginalization—coexist, forming a duality of microtask crowdsourcing. Their study heightens awareness of worker marginalization in microtask crowdsourcing, and offers guidelines for improving crowdsourcing practice. Specifically, they offer recommendations regarding

the ethical use of crowd workers (including for academic research), and call for improving platform design for greater worker empowerment.

Adda et al. [1] discussed the issue of compensation (monetary or otherwise) for completed tasks on crowdsourcing platforms. They also discussed the ethical and legal issues raised when considering work on crowdsourcing platforms as labor in the legal sense. They used crowdsourcing for under-resourced languages as a case study to exemplify the different issues they discussed. Finally, they proposed some specific solutions for researchers who wish to use crowdsourcing in an ethical manner.

Gellman [6] reviewed legal and regulatory issues that federal agencies face when they engage in citizen science and crowdsourcing activities. His report identified relevant issues that most federal agencies must consider, reviewed the legal standards, suggested ways that agencies can comply with or lawfully evade requirements, and discussed practical approaches that can ease the path for federal citizen science and crowdsourcing projects, including procedural activities, cooperative actions, legislative changes, and regulatory adjustment.

Brey [3] provided a critique of mainstream computer ethics and argued for the importance of a complementary approach called *disclosive* computer ethics, which is concerned with the moral deciphering of embedded values and norms in computer systems, applications and practices. Also, four key values were proposed as starting points for disclosive studies in computer ethics: justice, autonomy, democracy and privacy. Finally, it was argued that research in disclosive computer ethics should be multi-level and interdisciplinary, distinguishing between a disclosure level, a theoretical level, and an application level.

Schmidt [16] discussed some of the ethical implications of crowdsourcing in general and of contest-based crowd design in particular, especially in regard to the question of fair payment. He established four different categories of crowdsourcing with separate ethical challenges and argues for the crowd work industry to develop a code of ethics from within, in order to counter the exploitation and abuse that it often enables. In [13] authors proposed and tested such a bottom-up approach to defining a code of practice for crowdsourcing platform use which resulted in a set of guidelines for requesters to provide fair work conditions to Mechanical Turk workers.

Williamson [18] examined the ethics of crowdsourcing in social science research, with reference to her own experience using Amazon's Mechanical Turk. She pointed out that many people work long hours completing surveys and other tasks for very low wages, relying on those incomes to meet their basic needs. She resented her own experience of interviewing Mechanical Turk participants about their sources of income, and she offered recommendations to individual researchers, social science departments, and journal editors regarding the more ethical use of crowdsourcing. In [8], authors present a data-driven study of crowd worker wages showing how, on average, Mechanical Turk workers earn about $2 per hour while only 4% earns more than the US federal minimum wage.

Ford et al. [5] reviewed the various ways organizations employ non employees to overcome human resource limitations. They then focused on crowdsourcing as a novel source of external labor. After presenting key questions that every organization considering the use of crowdsourcing must address, they offered specific recommendations for those organizations who choose to employ a crowd to meet their needs.

Saxton et al. [15] provided a practical yet rigorous definition of crowdsourcing that incorporates crowds, outsourcing, and social web technologies. They then analyzed 103 well-known crowdsourcing websites using content analysis methods and the hermeneutic reading principle. Based on their analysis, they developed a "taxonomic theory" of crowdsourcing by organizing the empirical variants in nine distinct forms of crowdsourcing models. They also discussed key issues and directions, concentrating on the notion of managerial control systems.

Garber et al. [7] demonstrated that the crowdsourcing model of research has the potential to cause harm to participants, manipulates the participant into continued participation, and uses participants as experimental subjects. They concluded that protocols relying on this model require institutional review board (IRB) scrutiny.

Harris [9] explored the potential for which crowdsourcing may be used to bypass commonly-established ethical standards for personal or professional gain. Adda et al. [2] demonstrated that the situation in crowdsourcing is far from being ideal, be it from the point of view of quality, price, workers' status or ethics. Their goal was threefold: 1- to inform researchers, so that they can make their own choices with all the elements of such reflection

in mind, 2- to ask for help from funding agencies and scientific associations, and develop alternatives, 3- to propose practical and organizational solutions in order to improve new language resources development, while limiting the risks of ethical and legal issues without letting go of price or quality.

In this paper we present a summary of the current and envisioned future forms of work taking a stand on the ethical challenges that will need to be faced, including issues of transparency, fairness, and bias.

3 Open Issues in Current Platform-based Work

Empowering workers and protecting their rights and privacy should be at the heart of the Future of Work (FoW). This is a critical challenge as, while work platforms have a global reach, policies and regulations remain local for the most part. Advances in cybersecurity can be used to address privacy and access control mechanisms to guarantee that the right actors have visibility of the right data. Platforms should provide different privacy settings and be transparent about what data about workers is exposed and to whom. Requesters should be transparent about what the purpose of work is, and how the work outcome will be utilized. They should also be able to protect their confidential information when needed and to protect the copyrights and intellectual property of the work done by workers. Fair compensation for workers, including base payments, bonuses, benefits and insurance should be guaranteed and regulated by law. Workers should have the freedom to choose the compensation type they deem acceptable. Finally, job allocation should be transparent, fair and explainable by design. Worker's sensitive attributes that might bias the job allocation process should be protected. Auditing mechanisms to ensure compliance with fair, transparent and explainable job allocation and compensation need to be developed and adopted. Even though these ethical approaches are well known in the context of traditional work settings, these are more complex when applied through virtual platforms such as crowdsourcing. Workers are not considered as employees of the platforms but rather as self-employed workers that are paid for a one-off service. The question of which type of relation exists between workers and platforms is complex as different countries and states legislate differently on this issue. For example, California would consider drivers of Uber as employees. Platforms have also some other legal obligations such as the security of workers or requesters. For example, Uber has been banned from London because it was not able to guarantee the security of the drivers and customers.

In terms of fairness, an interdisciplinary approach will be required to develop novel methods to assess and quantify algorithmic fairness in job allocation practices. For example, looking at bias trade-offs between fully-algorithmic vs human-in-the-loop job allocation methods where algorithmic bias could lead to different issues as compared to implicit bias in humans. This will also result in higher levels of algorithmic transparency for job allocation where decisions should be easy to explain independently of whether they have been made by humans or by AI models. Processes and procedures should be in place to specify how to best address unfair cases, e.g., by means of additional rewards for workers or novel/better job opportunities. We also envision novel methods to make job allocation distribution (i.e., the long tail effect where few workers complete most of the available jobs) and time spent on jobs more transparent to workers and external actors like compliance agents. For example, visual analytics dashboards that communicate to workers how much time they spent and how much money they have earned on a platform, with warnings on risks for addiction or on unfair payments, or how transparent the requesters are with regard to the rules and procedures regarding compensation, for instance, are important in the FoW space. The power of the platform to delete the account of a worker or to prioritize particular workers over others creates difference in the work relations among the platform. These decisions should not only be explained, but processes should be put in place for workers and work providers to contest these unilateral platform decisions.

To summarize, in most of online work platforms, workers are currently poorly protected from the pressure of platforms and work providers, thus creating a power imbalance that puts workers in an unfavourable condition with limited or no contractual power. This situation raises a growing number of ethical issues related to privacy/access control, fairness and compensation.

First, the level of privacy is currently quite low since workers do not have control over their own data and their final work output. As IP regulations are non clear or almost non-existent, transfer of property rights is currently in favour of requesters which use the results/outputs of the crowd as they wish. Moreover, data relating to workers (personal information and online behaviour, for instance) are under the complete control of the platform and of the requesters.

Second, workers' compensation is limited. In some circumstances, workers are ready to work for free as they find some intrinsic motivation to do the job (pleasure to do a fun task, satisfaction to take part in a social/humanitarian/scientific project, willingness to be recognized by peers or the need to develop human and social capital). However, voluntary work also raises a number of issues: how to recruit enough volunteers? How to ensure work quality? How to mobilize them for a specific task within a specific time frame? Most of the time, workers are compensated by means of task-based payments. Here also, a huge number of issues emerge regarding this compensation mechanism. Task-based payment often leads to very low income, which is typically under the official minimum wage of the worker's country as the workers cannot negotiate their compensation. In addition, workers do not have the usual work benefits they would typically have in traditional work such as luncheon vouchers and health insurance, for instance. Finally, workers have no social security as they do not get paid if they are unable to work for any reasons (disease, lack of skills, less job offers, etc.).

Third, job platform offer a great opportunity for people who cannot work in a physical workplace to work anyway online. However, it remains, at the same time, a place where discrimination is currently maintained and perpetuated.

4 Requirements for the Future of Work

Workers' **privacy** should be at the heart of the future of work. Platforms should provide different privacy options and be transparent about what worker data is exposed and to whom. Access control for requesters should also be supported, for instance by having workers sign non-disclosure agreements for sensitive work. Finally, requesters should be transparent about what the work is for, and how the work outcome will be utilized and who owns what.

Compensation for workers should be fair and regulated by law. More training for workers should be provided, as this is mutually beneficial for workers and requesters and leads to a higher worker retention rate. Transparent work contracts that clearly indicate conditions of work should also be available. Base payments and bonuses should be provided for paid work as well as work protection and insurance for volunteers. Workers should have the freedom to choose among different compensation types (either paid, free or platform credits, for instance).

Finally, **job allocation** should be transparent and fair. Work platforms should explain to workers why or why not they have been allocated certain jobs. The job allocation algorithms should be open and explainable by design. Access control on sensitive attributes that might bias the job allocation process should be supported. Auditing mechanisms to ensure compliance with fair, transparent and explainable job allocation should be in place. Regulations to guarantee compliance with fairness requirements should be implemented.

The Universal Declaration of Human Rights (UDHR) (www.un.org/en/universal-declaration-human-rights/) points to the inherent dignity of human beings as the foundation for freedom, justice and peace in the world. It would be important to revisit the UDHR to develop fundamental principles of consistent and fair crowdsourcing practices for the future.

5 Impact on Data Engineering Research

5.1 Research on Privacy and Access Control

In terms of privacy and access control, required changes for FoW platforms will trigger advances in research on cybersecurity including topics on privacy management and access control. This will include designing novel and

usable privacy control mechanisms for all involved actors where workers can decide how much information to disclose to the platforms at different points in time and for different jobs, requesters can have data confidentiality guarantees in place where confidential data and business processes can be safely managed even if exposed to external on-demand workers. For example, NDA workers will be trusted not to disclose confidential information they may encounter while completing jobs with well defined legal implications in case of non compliance.

Together with advances in the computational field, we envision a catch up of the regulatory framework around the future job market where international law research will need to deal with challenges of conflicting regulations across national boundaries for cases in which the different actors involved are covered by different legal systems. From an ethical point of view, research practices will need to adapt to make sure experimenters disclose information to participating subjects on how they are being involved in an experiment and for which purpose their data is being used (e.g., informed consent).

In terms of **compensation**, we envision a more structured approach where researchers will be required to follow standard guidelines on how to reward participating subjects. This may include standards for monetary rewards (e.g., official price list per task type) as well as standards on how to manage volunteer participation (e.g., researchers being required to follow up with participants to disseminate the results of their research conducted thanks to volunteer contributions).

5.2 Research on Fairness

Fairness is an ethical concept which is addressed by numerous fields and perspectives. Roughly defined, fairness is the idea that an individual should obtain what it deserves. In work platforms, the issue of fairness mainly concerns the compensation system and job allocation process.

In terms of fairness, an interdisciplinary approach will be required to develop novel methods to assess and quantify algorithmic fairness in job allocation practices. For example, looking at bias trade-offs between fully-algorithmic vs human-in-the-loop job allocation approaches where algorithmic bias could be different from implicit bias in humans. It would also be necessary to be cognizant of selection bias, where the data inputs to the algorithms are not representative of a population and could result in conclusions that would favor certain groups over others. Further, it would be necessary to look out for unintended promotion of biases where feedback loops perpetuate bias in results.

By compensation, we refer to what is obtained by the worker in exchange for their work. In some cases, workers are willing to work without financial compensation in the extent that they find some intrinsic motivation to do the job, for instance, pleasure to do a funny task, satisfaction to take part to a social, humanitarian or scientific project, development of human and social capital. Most of the time, workers are compensated by a task-based payment, which raises a huge number of fairness concerns. One of the most prominent issue is the low income associated to task-based payments. As the workers cannot negotiate their compensation, they are often paid under the official minimum wage of their country. In addition, workers cannot benefit of usual advantages granted to traditional workers as health insurance, for example. Lastly, workers lacks of job security since if they are unable work for any reasons (illness, lack of skills, no job offer), they would not get paid.

As the unfairness of compensation is the most visible abuse of current paid crowdsourcing, we advocate for the development of strong regulation through national and international law. From an ethical point of view, we propose to implement for each kind of task, a base payment estimated by automatic calculations or negotiated by stakeholders. This minimum payment could then be supplemented by bonuses based on work performance. In addition, we claim that the main compensation type (i.e. financial compensation) is severely limited as compared to the wide range of compensation types potentially adopted. We therefore suggest to extended the kind of compensation also including training, bonuses, credit for insurance or social protection, or any other convenience item on the platform. Workers could hence choose the kind of compensation they prefer in accordance with theses goals or actual situation. As far as voluntary work is concerned, we recommend the mandatory integration of an insurance scheme which could prevent the occurrence of mental health issues caused by the job (e.g., due to

online content moderation tasks).

5.3 Research on Algorithmic Transparency

This will also result in higher levels of algorithmic transparency for job allocation where job allocation decisions should be easy to explain independently of whether they have been made by humans, algorithms, or by a combination of those. The other side of the coin is to develop methods to assess algorithmic unfairness [17]. While it may be easy for a human to determine what is fair or unfair based on fundamental ethical principles, understanding the level of fairness or unfairness is important in order to make decisions. Quite obviously there would be several situations where one would have to take the most fair or least unfair decision. This also calls for an additional research direction - **methods to fix unfair decisions**. In cases in which job allocation or compensation decisions have been judged to be unfair by a compliance verification step, processes should be in place to specify how to best address the situation, e.g., by means of additional rewards for workers or novel/better job opportunities. We also envision novel methods to make job allocation distribution (i.e., the long tail effect where few workers complete most of the available jobs) and time spent on jobs more transparent to workers and external actors like compliance agents. For example, visual analytics dashboards that communicate to workers how much time they spent and how much money they have earned on a platform with warnings on risks for addiction or unfair payments.

While, to-date, the literature has discussed the human-in-the-loop, soon, researchers will need to be studying society-in-the-loop (SITL) [12] methods. Innovations need the wisdom of the crowd, in fact, the collaboration of the crowd with algorithms is expected to be the future [11]. Rahwan [12] points out that, to move from human-in-the-loop to SITL, would be necessary in the case to have mechanisms for negotiating the values of various stakeholders, bringing in the social contract. This results in a whole new level of complexity, since often, aspects of the social contract are implicit rather than explicit and are embedded in social norms bringing in issues of ethics. Further, there is the issue of how to resolve trade-offs between security and privacy, or various aspects of fairness while at the same time considering unbiased inputs to algorithms through algorithmic regulations.

An interesting evolution in the context of FoW is the use of **smart contracts** which are developed to design and manage virtually contracts. They are used to establish contract between different participants and the rules of this contract are ensured by methods currently based on block chains. The ability to express a formal contract that could be transparent for the different participants and be ensured by some clear and automatic mechanisms would simply and clarify the interactions between workers and requesters.

6 Impact on Work Platforms

As ethical issues of FoW lies heavily on governance mechanisms, work platform will be impacted at several levels. Depending on the evolution of law, platform transformations will be voluntary or mandatory.

Ensuring privacy and increasing access control means providing all users (worker and requesters) with the ability to modulate their privacy parameters. They must be able to decide what kind of information can be disclosed and to whom. For instance, the user account may include a section called “privacy parameters” where users have access to information related to them (personal information, work history, log files, performance measure, and so on) and for each users could switch a button to express their willingness to disclose it. Since users’ behaviours are tracked and monitored, they need to know what kind of data is being collected, for what reasons, how is is treated, where it is stored and when it gets deleted. Regarding the platform, theses needs implies to provide users a space where they can get informed (e.g., a section in the user account). Information relating to the treatment of behaviour’s data could also be included in the general policy which should be read and accepted by any users.

As fairness issues lie on the algorithms used by platforms to select workers, assign jobs, and evaluate outputs,

platforms should be accountable for them. Many mechanisms to restrict the use of highly-unfair algorithms are available: disclosing them to the public, legal authorities or third parties, compliance to law, and frequent auditing. Thanks to the future developments in algorithms auditing research [14], Fow should rely on sophisticated auditing techniques and tools. Platform development would be impacted since they will have to be more rigorous when designing algorithms, integrating fairness while building algorithms, and accept frequent auditing processes. Depending on the development of law and scientific research on fairness measures and scales, compliance to standards may be formally required. Support to requesters and workers is another driver of fairness. Platforms are in the best position to do so. Today, some platforms already offer their support for designing tasks, defining the reward, and controlling quality. However, this service is rare and most of the time it does not consider fairness. In addition, platforms often take the requesters side in case of conflict with workers as they are seen as the customer. In the Fow, platforms should become more active and concerned about requesters needs as well as about the quality of requester-worker relationships. They can, for example, develop algorithms in order to model the utility function of the workers, be moderating and listening to both sides during conflict resolution, helping to fix performance criteria for requesters, and provide opportunities for improvement to workers. A greater monitoring of the platform activities may lead to avoiding mistakes, abuses, and unfair decisions.

Fow platforms should integrate a sophisticated compensation system. First, the different kinds of compensation can be expanded to include training, bonuses, insurance, and others. Training appears as a compensation type that may be beneficial to all stakeholders: For the platform, it is a good way to obtain better crowd retention rates; requesters then would have access to highly skilled workers; and workers could earn better wages. Insurance could prevent workers being harmed by the work itself, especially in the case of voluntary jobs. Taxes may be charged to platforms, in order to fund compliance activities and worker insurance. With this system, platforms could then offer the opportunity to workers to decide how they want to be compensated. To go further, they could design a recommendation system which does a matchmaking between work/workers/kind of compensations. At the second level, when focusing on the monetary compensation, FoW will include specific support mechanism for setting rewards to prevent unfairness (e.g., checking the average wage balance across genders). Helping requesters to fix a reward which matches the difficulty and the length of the task would be a good way to ensure that workers will receive a decent payment. Guidelines for worker compensation can be also be made available by platforms. Ideally, fair Fow platform will deny jobs with a very low monetary compensation. However, as requesters are the main customers of these platforms, it is more realistic to envision a negotiation process unless strict regulations are put in place.

As discussed in this study, FoW platforms have to face complex challenges as they are designed to serve different purposes: their main role is to match workers and requesters but they also provide payments to the workers, guarantee the security of workers and requesters, provide the technical support for running the tasks. If the matchmaking challenge has got a lot of attention by the research community, the other issues are often forgotten even though they have important ethical aspects. The current mix of the different roles in work platforms creates high complexity in managing access control over the data as well as challenges related to fairness and explainability. We believe that a clearer distinction of the different roles in platforms and possibly a better separation of them could lead to more ethical work processes.

7 Conclusions

FoW will require overhauling the design and engineering of online job platforms to enable the collection, storage, retrieval, analysis, and mining of a wide array of human data across different types of technology-driven work. A fair bit of engineering and testing will be needed to ensure the development of scalable and portable platforms and the integration of multi-stakeholder goals in efficient and effective ways.

We call the data engineering community to consider working on the upcoming FoW research challenges as data will be the key enabler of the FoW. While doing this, we urge the community to also consider the

ethical dimension of innovative solutions to make sure that aspects of privacy, accountability, fairness, and transparency/explainability are embedded by design in such data-driven solutions for FoW platforms.

References

[1] G. Adda, J. Mariani, L. Besacier, and H. Gelas. *Crowdsourcing for Speech: Economic, Legal and Ethical analysis*. PhD thesis, LIG lab, 2014.

[2] G. Adda, B. Sagot, K. Fort, and J. Mariani. Crowdsourcing for language resource development: Critical analysis of amazon mechanical turk overpowering use. In *5th Language and Technology Conference*, 2011.

[3] P. Brey. Disclosive computer ethics. *ACM Sigcas Computers and Society*, 30(4):10–16, 2000.

[4] X. Deng, K. Joshi, and R. D. Galliers. The duality of empowerment and marginalization in microtask crowdsourcing: Giving voice to the less powerful through value sensitive design. *Mis Quarterly*, 40(2):279–302, 2016.

[5] R. C. Ford, B. Richard, and M. P. Ciuchta. Crowdsourcing: A new way of employing non-employees? *Business Horizons*, 58(4):377–388, 2015.

[6] R. Gellman. Crowdsourcing, citizen science, and the law: legal issues affecting federal agencies. *Commons Lab, Woodrow Wilson International Center for Scholars*, 2015.

[7] M. A. Graber and A. Graber. Internet-based crowdsourcing and research ethics: the case for irb review. *Journal of medical ethics*, 39(2):115–118, 2013.

[8] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 449. ACM, 2018.

[9] C. G. Harris. Dirty deeds done dirt cheap: a darker side to crowdsourcing. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 1314–1317. IEEE, 2011.

[10] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318. ACM, 2013.

[11] T. W. Malone. How human-computer'superminds' are redefining the future of work. *MIT Sloan Management Review*, 59(4):34–41, 2018.

[12] I. Rahwan. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1):5–14, 2018.

[13] N. Salehi, L. C. Irani, M. S. Bernstein, A. Alkhatib, E. Ogbe, K. Milland, et al. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1621–1630. ACM, 2015.

[14] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22, 2014.

[15] G. D. Saxton, O. Oh, and R. Kishore. Rules of crowdsourcing: Models, issues, and systems of control. *Information Systems Management*, 30(1):2–20, 2013.

[16] F. A. Schmidt. The good, the bad and the ugly: Why crowdsourcing needs ethics. In *2013 International Conference on Cloud and Green Computing*, pages 531–535. IEEE, 2013.

[17] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248. ACM, 2018.

[18] V. Williamson. On the ethics of crowdsourced research. *PS: Political Science & Politics*, 49(1):77–81, 2016.

Data Engineering

It's FREE to join!

TCDE

tab.computer.org/tcde/

The Technical Committee on Data Engineering (TCDE) of the IEEE Computer Society is concerned with the role of data in the design, development, management and utilization of information systems.

- Data Management Systems and Modern Hardware/Software Platforms
- Data Models, Data Integration, Semantics and Data Quality
- Spatial, Temporal, Graph, Scientific, Statistical and Multimedia Databases
- Data Mining, Data Warehousing, and OLAP
- Big Data, Streams and Clouds
- Information Management, Distribution, Mobility, and the WWW
- Data Security, Privacy and Trust
- Performance, Experiments, and Analysis of Data Systems

The TCDE sponsors the International Conference on Data Engineering (ICDE). It publishes a quarterly newsletter, the Data Engineering Bulletin. If you are a member of the IEEE Computer Society, you may join the TCDE and receive copies of the Data Engineering Bulletin without cost. There are approximately 1000 members of the TCDE.

Join TCDE via Online or Fax

ONLINE: Follow the instructions on this page:

www.computer.org/portal/web/tandc/joinatc

FAX: Complete your details and fax this form to **+61-7-3365 3248**

Name _____

IEEE Member # _____

Mailing Address _____

_____

Country _____

Email _____

Phone _____

TCDE Mailing List

TCDE will occasionally email announcements, and other opportunities available for members. This mailing list will be used only for this purpose.

Membership Questions?

Xiaoyong Du
Key Laboratory of Data Engineering and Knowledge Engineering
Renmin University of China
Beijing 100872, China
duyong@ruc.edu.cn

TCDE Chair

Xiaofang Zhou
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, QLD 4072, Australia
zxf@uq.edu.au

IEEE Computer Society
10662 Los Vaqueros Circle
Los Alamitos, CA 90720-1314

Non-profit Org.
U.S. Postage
PAID
Los Alamitos, CA
Permit 1398