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

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

productive use of workers/volunteers in important domains such as Wikipedia, data annotation for machine 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 questionswhile 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 JournalThe 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.