Check for updates

RESEARCH ARTICLE

REVISED How many online workers are there in the world? A data-driven assessment

[version 4; peer review: 4 approved]

Otto Kässi[1,2], Vili Lehdonvirta[3,4], Fabian Stephany[3]

[1]ETLA Economic Research, Helsinki, Finland
[2]University of Turku, Turku Centre for Labour Studies (TCLS), Turku, Finland
[3]Oxford Internet Institute, University of Oxford, Oxford, Oxfordshire, OX1 3JS, UK
[4]Alan Turing Institute, London, UK

v4 **First published:** 19 May 2021, **1**:53
https://doi.org/10.12688/openreseurope.13639.1
Second version: 01 Sep 2021, **1**:53
https://doi.org/10.12688/openreseurope.13639.2
Third version: 07 Oct 2021, **1**:53
https://doi.org/10.12688/openreseurope.13639.3
Latest published: 15 Oct 2021, **1**:53
https://doi.org/10.12688/openreseurope.13639.4

Abstract

An unknown number of people around the world are earning income by working through online labour platforms such as Upwork and Amazon Mechanical Turk. We combine data collected from various sources to build a data-driven assessment of the number of such online workers (also known as online freelancers) globally. Our headline estimate is that there are 163 million freelancer profiles registered on online labour platforms globally. Approximately 14 million of them have obtained work through the platform at least once, and 3.3 million have completed at least 10 projects or earned at least $1000. These numbers suggest a substantial growth from 2015 in registered worker accounts, but much less growth in amount of work completed by workers. Our results indicate that online freelancing represents a non-trivial segment of labour today, but one that is spread thinly across countries and sectors.

Keywords

Platform Economy, Online Labour Markets, Gig Work

This article is included in the Economics and Business gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ✓

	1	2	3	4
version 4 (revision) 15 Oct 2021				
version 3 (revision) 07 Oct 2021			✓ view	
version 2 (revision) 01 Sep 2021		✓ view	? view	
version 1 19 May 2021	✓ view	? view	? view	✓ view

1. **Paola Tubaro**, Université Paris-Saclay, Paris, France
2. **Annarosa Pesole**, Institute for Prospective Technological Studies (IPTS), Seville, Spain
3. **Richard Heeks**, University of Manchester, Manchester, UK
4. **Valeria Cirillo**, University of Bari Aldo Moro, Bari, Italy

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Horizon 2020 gateway.

This article is included in the European Research Council (ERC) gateway.

This article is included in the The Future of Work collection.

Corresponding author: Otto Kässi (otto.kassi@etla.fi)

Author roles: Kässi O: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation; **Lehdonvirta V**: Conceptualization, Funding Acquisition, Project Administration, Supervision; **Stephany F**: Project Administration, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No [639652]), (project iLABOUR) and (grant agreement No [875255]), (project GIGSTATS).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Kässi O *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kässi O, Lehdonvirta V and Stephany F. **How many online workers are there in the world? A data-driven assessment [version 4; peer review: 4 approved]** Open Research Europe 2021, **1**:53 https://doi.org/10.12688/openreseurope.13639.4

First published: 19 May 2021, **1**:53 https://doi.org/10.12688/openreseurope.13639.1

REVISED **Amendments from Version 3**

We have corrected the summary worker head counts (now 14 / 3.3 million active / significantly active worker) in the Abstract. An additional affiliation has been listed for Otto Kässi: University of Turku, Turku Centre for Labour Studies (TCLS), Turku, Finland.

Any further responses from the reviewers can be found at the end of the article

Introduction

Development in digital communication technologies has made transacting work remotely far easier and more economical. At the forefront of this phenomenon are so-called online labour platforms, also known as online outsourcing, crowdwork, or online gig platforms. They allow workers to serve multiple clients at varying hours remotely from their homes or co-working spaces instead of working full-time for a single employer. In this short paper, we refer to the phenomenon as online freelancing, though the employment status of platform-based work is in some cases contested.

Current economic statistics are not well suited to measuring the online freelance economy, in terms of both capturing its full extent as well as distinguishing its impact from other activities (Abraham *et al.*, 2017). Kässi & Lehdonvirta (2018) give several reasons for this: the standard definition of employment is someone who has done at least one hour in the tracking period. Since online work is often a source of supplementary income (Farrell & Greig, 2016), labour force surveys do not capture it. Moreover, many online workers might not report their earnings to tax agencies, especially if their earnings are small. Tax non-compliance might be particularly prevalent among online workers living in lower-income countries with weaker tax enforcement. In most cases platform companies are not considered employers and thus are not required to report the income earned by the workers (Ogembo & Lehdonvirta, 2020).

Lehdonvirta *et al.* (2019), Kuek *et al.* (2015), Horton *et al.* (2018), Braesemann *et al.* (2018) and Melia (2020), among others, have argued that digital jobs can facilitate virtual migration, or bring jobs to people instead of forcing workers to migrate to where the jobs are. This, in itself, could be a powerful mechanism for development as a large share of the global digital labour force resides in developing countries and social distancing counter-measures against the coronavirus disease 2019 (COVID-19) pandemic bolster the importance of remote online freelancing (Stephany *et al.*, 2020). The argument has been challenged, among others, in Anwar & Graham (2019), Graham & Anwar (2019), Casilli (2017), and Berg & Johnston (2019) because digital workers lack formal labour protection and are easily exploited by their employers. We assert that this debate lacks hard data. Since a large share of this activity happens under the radar of national statistical agencies, policymakers and researchers have limited possibilities for assessing the extent and impact of digital labour markets on workers.

The objective of this paper is simple. To assess the global significance of online labour platforms as a source of income, we produce an estimate of the total number of online freelancers globally and document the uncertainties related to the calculation. To facilitate replication and follow-up research, we use publicly available data and make our assumptions explicit.

There are three existing data analyses with a similar goal to ours. Kuek *et al.* (2015) and Codagnone *et al.* (2016) used a combination of expert interviews and data disclosed by online labour platforms to estimate the numbers of platform workers globally. Heeks (2017) used estimates from these papers in conjunction with survey estimates to make inferences about the geographic distribution of online work.

Calculations based on expert interviews are useful, but their sources and methods lack transparency and are difficult to repeat regularly in a way that would produce comparable statistics over time. Kässi & Lehdonvirta (2018) took a different approach by estimating the growth rates and geographic distribution of online freelancing by observing vacancies posted on selected English language platforms, but they were not able to count the absolute number of workers filling these tasks.

In addition to international mapping exercises, there have been several national surveys that have assessed the local relevance of digital platform work. These include, among others, Pesole *et al.* (2018) and Huws *et al.* (2017) who both concentrate on selected European countries. Unfortunately, many populous countries that supply large shares of online labour have not completed such surveys. Moreover, many surveys fail to distinguish local platform work, and activities such as e-commerce and house rental from remote platform work. Another, related, but distinct approach is used in Le Ludec *et al.* (2020), and Difallah *et al.* (2018). These papers used a capture-recapture model inspired by ecological sciences to infer the size of microworker populations in France, and number of workers on a single platform, respectively.

The work presented in this paper is thus to our knowledge the first to use a fully quantitative and transparent approach to estimating the absolute number of online workers globally. Beyond providing a headline number, the more general contribution of this paper is that we outline the relevant quantities a researcher needs to know when trying to understand how many online workers there are worldwide.

Methods

We started by attempting to create, as far as possible, a complete census of all online labour platforms of non-trivial size. We used three main sources of information to compile a list of platform names. First, we analysed a publicly available database of the crowd-sourced company information platform, Crunchbase, especially its 'freelance' and 'crowdsourcing' categories. Our second data source for platform names is a cross-regional survey collected in Wood *et al.* (2019a). Finally,

we supplemented our list with information found through Google searches concerning Spanish, Latin American, Russian, and Chinese online freelancing platforms[1].

We limited our attention to platforms where the transaction is fully digital; that is, the work is delivered and paid remotely over the internet. Local gig economy platforms such as ride-hailing apps and food delivery platforms are thus excluded. Distinguishing the online freelance economy from the local gig economy is important because, among other reasons, the transnational nature of the market means that it has different potential implications to global service trade and development. We ended up with a list of 351 online freelancing platforms. To best of our understanding, these 351 platforms constitute nearly the full extent of online freelancing platforms at the end of 2020. The full list of 351 platforms, including information on worker count, sources, and type of platform work offered, can be found in this supplementary Table A.

We then used public data sources to obtain three measures of worker numbers for each platform: number or registered worker profiles, number of profiles of registered workers who have ever worked, and number of profiles of registered workers who have worked significantly, i.e., who completed at least 10 projects, or earned at least $1000, when available. We collected these numbers through a combination of media mentions, literature review, and platforms' search functionalities.

We were able to observe the number of registered workers in 162 of 351 cases. The distribution of this variable is plotted in Figure 1. The sum of registered workers across all the 162 platforms is 140,000,000[2]. Most of the platforms have fewer than a million registered workers. Three outliers have particularly large numbers of registered workers: freelancer.com (31 million workers), epwk.com (23 million registered workers), and zbj.com (23 million registered workers).

[1]We used the search terms "online freelancing platform", "online labour platform", and "online gigwork" (google-translated to the corresponding languages).

[2] We round all numbers to two significant digits throughout this paper.

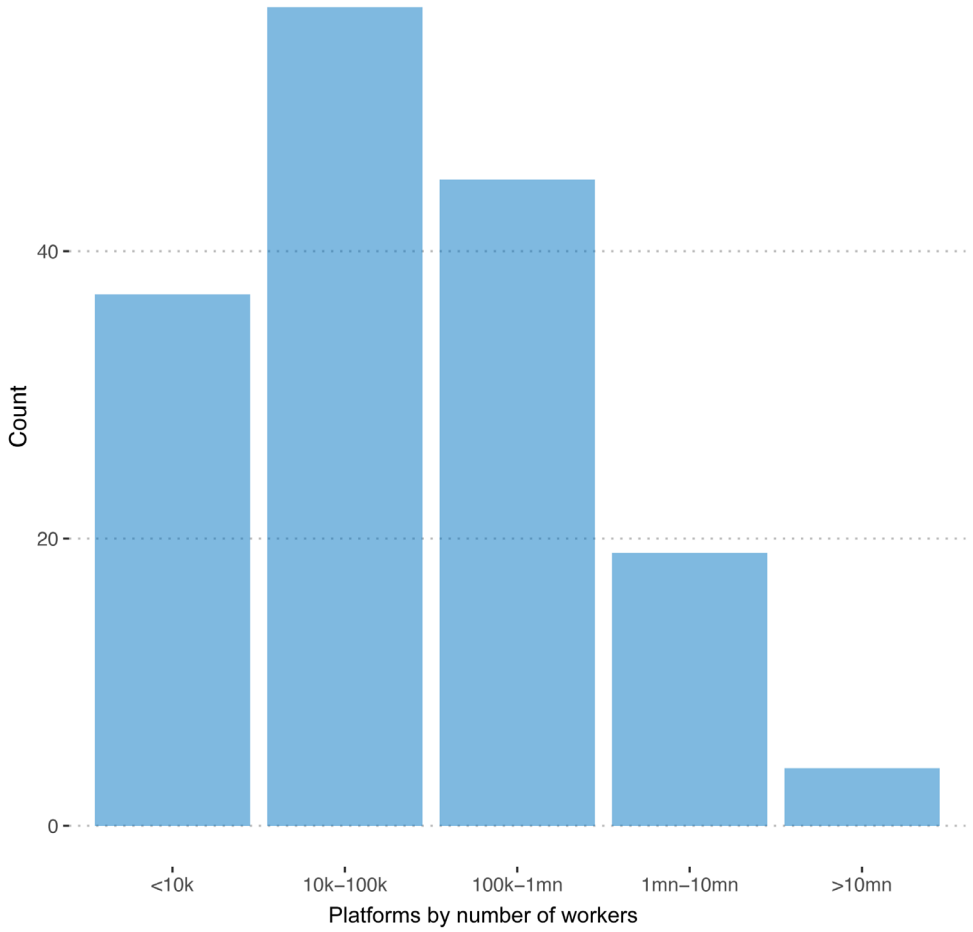


Figure 1. Number of freelancers. On 162 out of 351 registered platforms, we are able to observe the number of registered freelancers.

We were able to observe the number of workers who had ever worked on the platform for only 7 platforms. For 6 platforms we were able to observe the number of workers who had worked significantly, i.e., who completed at least 10 projects, or earned at least $1000[3]. Fortunately, this information is available for some of the largest platforms, such as Freelancer, as well as for some smaller ones.

For those platforms for which we could not obtain numbers from public sources, we instead imputed the quantities as described in the next subsection.

Predicting number of registered workers

Previous research has used various rules-of-thumb methods for estimating numbers of workers registered on online labour platforms. For example, Kuek *et al.* (2015) assumed that the top three firms form 50 % of the entire online freelancing market. Using this assumption, by obtaining data on the top three platforms only, they generalised their findings to the market as a whole.

We instead adopted a data-driven approach to predict the number of workers for the platforms for which this information was not available. We collected a list of publicly available predictive features and trained a machine learning model to predict the number of workers for the platforms where this information was not available.

We use the following predictive features, all measuring different aspects of website popularity, in building the model:

- **Alexa rank.** Alexa is a web traffic analysis company whose data is frequently used to compare the popularity of different websites. We used the most recent Alexa rank as reported by the siterankdata.com analytics tool (accessed 2020-09-29). If the Alexa rank is not reported for a given site, we have inputted the maximum in the data as the rank.
- **Estimate for monthly unique users.** Estimated number of monthly unique users as reported by the analytics tool siterankdata.com (accessed 2020-09-29). If the estimate is not available, we have inputted a zero.
- **Median of daily Google Trends index values between 2019-09-01 and 2020-09-01.** Downloaded from Google Trends using the site URL (e.g. 'upwork.com') as the search term. If values are not available, we have inputted a zero[4].
- **Sum of daily Google Trends index values between 2019-09-01 and 2020-09-01.** Downloaded from Google Trends using the site URL (e.g. 'upwork.com') as the search term. If values are not available, we have inputted a zero.

The data are summarised in Table 1. Google Trends produces time series data, which for the purposes of this exercise we needed to summarise as a single number. Since there is no one theoretically correct way of doing this, we used two different summarisations: the sum and median of the time series' values, and relied on a machine learning algorithm to choose relevant weights for both features. All features were standardised by subtracting the mean and dividing by the standard deviation and logarithmised before entering them into the model.

Table 1 also summarises our main dependent variables of interest: number of registered workers, and shares of active workers and workers, who have worker significantly. As is evident from the summary statistics, both of the share of active workers and workers, who have worked significantly variables are highly right-skewed with means much larger than the median values.

Our training data consists of 159 observations[5]. We decided to use 128 observations (80%) as our validation set and 31 (20%) as our training set. After some experimentation, we chose an XGBoost (Chen & Guestrin, 2016) model with a Poisson objective. The main upside of the Poisson assumption is that it naturally limits the dependent variable to positive values without any additional adjustments. After an extensive grid search along the model's hyperparameters, we chose the model that minimised the Root Mean Square Error (RMSE) of the prediction in the validation data. The performance of the best performing model is reported in Table 2. We see that the RMSE of the predictions is relatively large at 577,000, which reflects the fact there is non-negligible uncertainty in the prediction. As we argue below, despite the large RMSE, the prediction is informative of the number of workers registered on platforms.

Figure 2 summarises the performance of the predictive model graphically. As is expected, the correlation between observed and predicted numbers of registered workers is exactly 1. A more honest test for the performance of the predictive model is its performance in the validation data. Here, we find that the correlation between training and test data is more moderate (0.27), but still clearly positive.

Adding up the predicted numbers of registered workers across the platforms yields a total of 23,000,000 workers. Adding to this the number of directly observed workers (140,000,000) yields 163,000,000 workers, which is our point estimate for the number of registered workers across the global online freelance economy. Figure 3 plots the distribution of observed and predicted numbers of workers across platforms. We see that the platforms for which we predict the numbers of registered workers are predominantly on the smaller end.

[3] Only this set of platforms visibly provide the relevant information.

[4] Google Trends is a product that summarises search volume by time. For more details, see: https://trends.google.com/trends (accessed 2020-03-27). A feature of the Google Trends tool is that it is only possible to compare five keywords for each search. To work around this issue, we always pull data for 'upwork.com' and four other platform names, and continuously normalise the search terms as $\text{GoogleTrend}_s/\text{GoogleTrend}_{\text{Upwork}}$. This way, all index values are relative to Upwork.

[5] We excluded the three Chinese platforms, 680.com, epwk.com and zbj.com, from the training data because Google and Alexa probably do not cover Chinese web traffic due to the Chinese internet censorship. Moreover, excluding the three web sites resulted in a much better performance of the machine learning models.

Table 1. **Summary statistics (observed data).**

		N	Mean	Median	Min	Max
siterankdata.com						
	Alexa Rank	339	282,182	258,266	346	986,677
	Estimated monthly unique users	339	4,697,358	367,782	95,241	289,444,543
Google trends						
	Sum of index values	351	0.02	0.01	0.00	1.40
	Median of index values	351	0.02	0.00	0.00	1.59
Number of Workers						
	Total	162	869,173	50,000	80	31,464,473
	<10k	37	3,237	2,700	80	8,000
	10k–100k	57	33,648	25,187	10,000	90,237
	100k–1mn	45	321,628	259,508	100,000	856,327
	1mn–10mn	19	1,891,763	1,480,933	1,000,000	7,000,000
	>10mn	4	22,087,877	22,443,518	12,000,000	31,464,473
Share of active workers						
	Worked at least once	7	8.6%	1.3%	0.2%	36.8%
	With at least 10 completed projects or at least $1000 earned	6	2.0%	0.3%	0.1%	10.0%

Table 2. **Estimation results.** While observing 140 million workers, we predict there to be an amount of 23 million unobserved freelancers worldwide. Notes: For row (b), the error band is calculated as bootstrapped 95% confidence interval; for rows (c), and (d) the error bands are calculated as minimum and maximum values in the data; for row (e) the error band is calculated as the 2.5th and 97.5th percentile point of the distribution of platforms mentioned; for row (f) the error band is calculated as +/- 1.96 * std. error. All numbers are rounded to two significant digits. See text for details.

		Estimate	Error band
	Number of workers		
(a)	registered (observed)	140,000,000	
(b)	registered (predicted)	23,000,000	[12,000,000, 65,000,000]
(c)	with at least one project completed	8.6%	[2.0%, 36.8%]
(d)	with at least 10 completed projects or at least $1000 earned	2.0%	[0.1%, 10.0%]
(e)	Average of multi-homing platforms	1.83	[1, 4]
(f)	Proportion of workers sharing accounts	0.21	[0.18, 0.24]

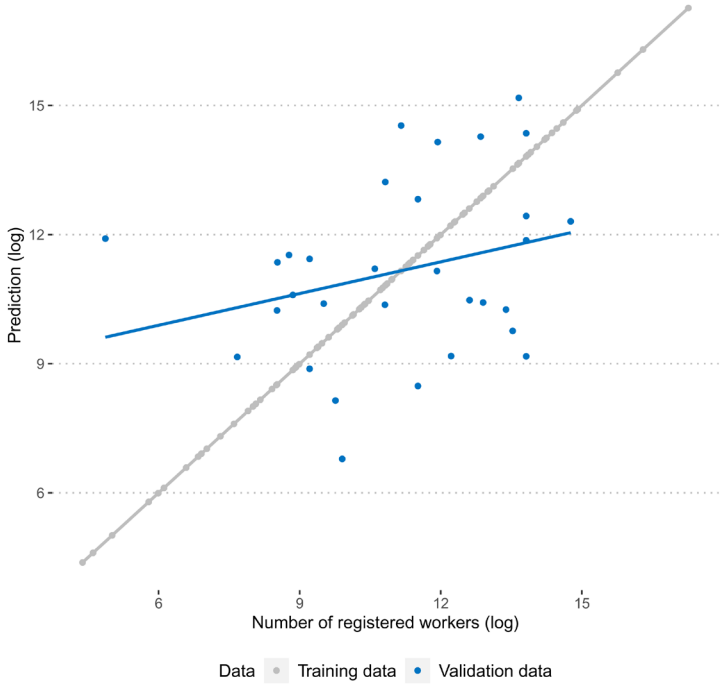


Figure 2. Correlation between predicted and observed number of workers in training and validation data.

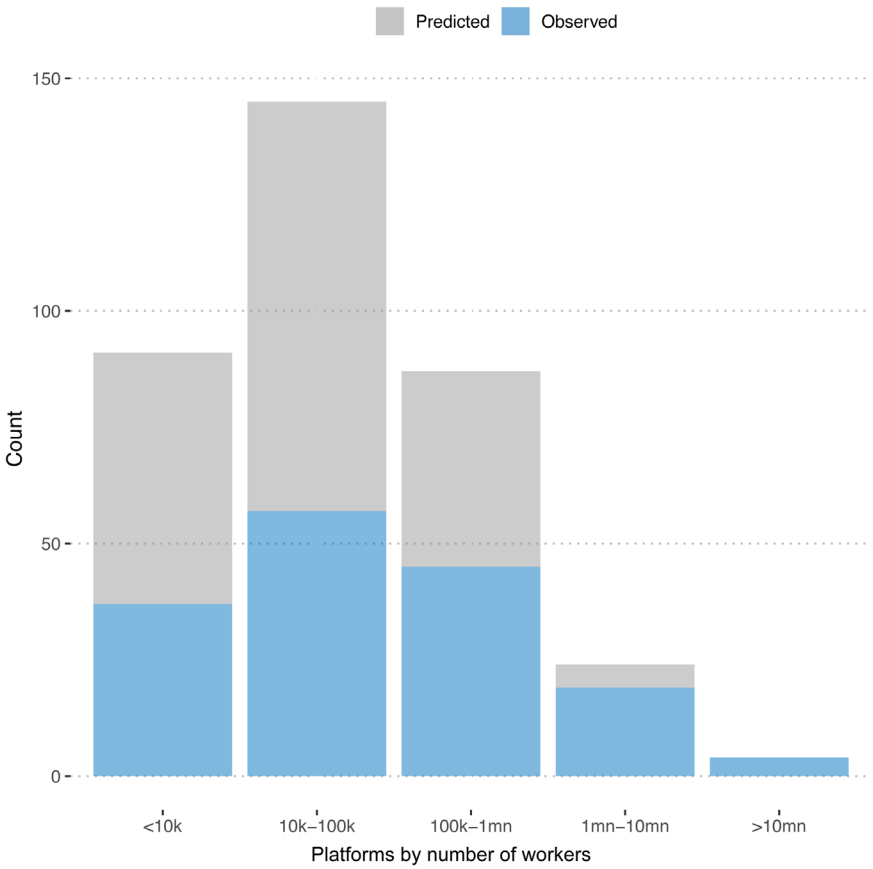


Figure 3. Number of freelancers. Distribution of registered workers for predicted and observed platforms.

To indicate the uncertainty related to the prediction, we have also estimated a 95% prediction interval for the numbers of registered workers by bootstrapping.

Inferring the number workers who have worked

On most online labour platforms, the number of registered users might not represent the number of online workers who are actually active. To capture this, we follow a similar approach to above and generalise from the known population. On average, only 8.6% of registered users have ever worked. These findings imply that for a large majority of the platform workforce the platform is a source of occasional additional income rather than the main income source. Moreover, the strikingly small share of workers, who have worked significantly, indicates that there is vast oversupply of workers on the platforms we observe.

We note, that only a handful of platforms reveal this information publicly, the sample sizes for these estimates remain very small. Thus, instead of calculating a formal confidence interval for these estimates, we use the minimum and maximum values of the samples as our error band estimate in sensitivity analyses.

Multi-homing

Multi-homing, or the practice of agents being affiliated with more than one platform, can lead to double-counting of workers. If a worker is active or registered on more than one platform, they will be counted more than once in our data.

There are no measures for double-counting available through public data sources. Fortunately, questions about multi-homing have been asked in several surveys administered by the International Labour Organisation (ILO, 2021), and by Wood *et al.* (2019a). These surveys asked active freelancers to list how many platforms they worked on. Across the surveys, 48% of the respondents mentioned that they worked exclusively on a single platform. On average, the survey respondents were active on 1.83 platforms. Thus, we can further adjust down the number of active worker profiles to account for multi-homing by dividing the number of active workers by 1.83.

However, our results on multi-homing could be challenged because our results on multi-homing are from a non-representative convenience sample. We note that our numbers align well with those reported in Le Ludec *et al.* (2020).

Multi-working: multiple workers using a single account

Qualitative evidence discussed in detail in Lehdonvirta *et al.* (2015), Wood *et al.* (2019b) and Melia (2020) suggests that in some cases several workers might be working under a single freelancer account (multi-working). To the best of our knowledge, there are no systematic studies on this phenomenon.

The surveys discussed in ILO reports and in Wood *et al.* (2019b) asked the following three questions from workers:

"*Over the last 7 days, I have hired workers in my local area to do online work that I got from a client*", "*Over the last 7 days. I have hired family or friends to do online work that I got from a client*", or "*Have you ever participated in digital platform work using a login / account / profile that belongs to someone else or that is shared by multiple people?"*

Overall, 21% of respondents across the surveys answered yes to one of these questions. If we further assume that an account is shared between a maximum of two workers, we can adjust our numbers for multiple workers working under a single account by multiplying the number of workers by 1.21.

Results

This section combines the individual parameter estimates discussed above into a single number. Moreover, we provide data-driven error bands for the parameters underlying our estimate.

According to these numbers, our point estimates suggest that there are 163 million registered worker profiles online freelancing platforms. Of them, roughly 14 million have ever worked, and 3.3 million have had worked significantly (our definition applies to workers who have had total earnings of at least $1000 or who have at least 10 completed projects). Further adjusting for multi-homing, these numbers reduce to 7.7 million and 1.8 million, respectively. Finally, adjusting for possible multi-working increases these numbers to 9.3 million and 2.2 million, respectively[6].

Nonetheless, there is considerable uncertainty in these estimates. Given the relatively large error bands, our estimates suggest that there could at most be as many as 205 million registered worker profiles, 75 million workers who have ever worked through an online labour platform, and 21 million workers, who have worked significantly[7].

Discussion and conclusions

This paper used a combination of data sources to produce an estimate for the number of online workers in the online freelancing economy.

According to our headline estimates, we estimate that there are 163 million registered workers on online labour platforms, and 8.6% of them have ever worked through a platform. These numbers point to a stark growth if compared with the 2015 estimates by Kuek & colleagues (2015), whose corresponding numbers were 50 million and 10%.

The differences in these estimates are not only due to methodological differences. Kuek and colleagues assumed that Upwork, Freelancer, and Zbj form half of the total market. Using their assumptions with our data, our estimate would have been 130 million freelancers in 2020.

Instead, new platforms with a large reach have emerged between 2014 and 2020. Moreover, it could be the case

[6] 7.7 million * multi-working factor (1.21) = 9.3 million and 1.8 million * multi-working factor (1.21) = 2.2 million.

[7] These numbers result from the upper benchmarks of our confidence intervals, i.e. upper prediction benchmark, i.e., ever worked: 205 million * 36.8% = 75 million, worked significantly: 205 * 10% = 21 million.

that there are now more, and more geographically or professionally specialised online labour platforms than when Kuek & colleagues (2015) conducted their study.

Compared to most recent previous estimates (Kuek *et al.*, 2015), our figures support the narrative that online work is growing rapidly (Chan & Wang, 2018; Huws *et al.*, 2017). Nonetheless, the fact that only a small minority have completed any projects, let alone a substantial number of projects suggest that digital platform work is a viable way to make a living only to a small minority of registered workers.

We stress that our estimates come with fairly big error bands. However, even the lower end of our estimates suggest that the online freelancing economy has grown. For instance, the Online Labour Index (Kässi & Lehdonvirta, 2018) indicates yearly growth rates of over 10%. Upwork, one of the larger online labour platforms, has reported almost 20% year-on-year growth rates in gross freelancer revenues.

We believe that our approach is transparent and our methodological choices sound. There are a few sources of error that could bias our estimates downwards that we cannot tackle. In particular, we want to highlight two error sources.

First, the estimates for the shares of active workers, multi-homing, and multi-working come from small opportunistic samples from limited countries. It is very possible that multi-homing and account sharing practices vary considerably by country and platform.

Second, quantitative evidence on the extent and nature of working on shared accounts (multi-working) is particularly slim. For more reliable evidence on these, we would need representative surveys of freelancers working on the major platforms. Fortunately, since only a handful of platforms cover most of the market, a survey that covers only the major platforms should give us a good understanding of the total market.

More broadly, platform mediated remote work is just one facet of computer-mediated labour. Other facets, such as platform-mediated place-based work, i.e., the local gig economy of ride hailing and delivery services, remote work for overseas clients, and business process outsourcing, can have their own specific impacts on economic development, work and labour market statistics. We hope to see more research on developing better measures for these phenomena as well.

Data availability

Figshare: How many online workers? https://doi.org/10.6084/m9.figshare.14472915.v3 (Kässi *et al.,* 2021)

This project contains the following underlying data:

- */data*, platform data
- */model_and_vis*, code for model and visualisations
- */graphs*, paper figures

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

References

Abraham K, Haltiwanger J, Sandusky K, *et al.*: **Measuring the gig economy: Current knowledge and open issues**. *Measuring and Accounting for Innovation in the 21st Century.* 2017.
Reference Source

Anwar MA, Graham M: **Hidden transcripts of the gig economy: labour agency and the new art of resistance among African gig workers.** *Environment and Planning A: Economy and Space.* 2020; **52**(7): 1269–1291.
Publisher Full Text

Berg J, Johnston H: **Too Good to Be True? A Comment on Hall and Krueger's Analysis of the Labor Market for Uber's Driver-Partners.** *ILR Review.* 2019; **72**(1): 39–68.
Publisher Full Text

Braesemann F, Lehdonvirta V, Kässi O: **ICTs and the Urban-Rural Divide: Can Online Labour Platforms Bridge the Gap?** *Information, Communication & Society*. 2018.
Publisher Full Text

Casilli AA: **Digital Labor Studies Go Global: Toward a Digital Decolonial Turn.** *Int J Commun Syst.* 2017; **11**: 3934–3954.
Reference Source

Chan J, Wang J: **Hiring preferences in online labor markets: Evidence of a female hiring bias.** *Manage Sci.* 2018; **64**(7): 2973–3468.
Publisher Full Text

Chen T, Guestrin C: **Xgboost: A scalable tree boosting system**. In *Proceedings of the 22nd ACM SIGKKD International Conference on Knowledge Discovery and data mining.* ACM. 2016; 785–794.
Publisher Full Text

Codagnone C, Abadie F, Biagi F: **The future of work in the 'sharing economy'**. *Market efficiency and equitable opportunities or unfair precarisation*. 2016.
Publisher Full Text

Difallah D, Filatova E, Ipeirotis P: **Demographics and dynamics of mechanical turk workers**. In *Proceedings of the eleventh ACM international conference on web search and data mining.* 2018; 135–143.
Publisher Full Text

Farrell D, Greig F: **Paychecks, paydays, and the online platform economy.** In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association.* National Tax Association. 2016; **109**: 1–40.
Reference Source

Graham M, Anwar MA: **The global gig economy: Towards a planetary labour market?** *First Monday.* 2019; **24**(4).
Publisher Full Text

Heeks R: **Decent work and the digital gig economy: a developing country perspective on employment impacts and standards in online outsourcing, crowdwork, etc**. Development Informatics Working Paper no 71. 2017.
Publisher Full Text

Horton J, Kerr WR, Stanton C: **3. Digital Labor Markets and Global Talent Flows**. University of Chicago Press, 2018; 71–108.
Reference Source

Huws U, Spencer N, Syrdal DS, *et al.*: **Work in the European gig economy: Research results from the UK, Sweden, Germany, Austria, the Netherlands, Switzerland and Italy**. 2017.
Reference Source

ILO: **World Employment and Social Outlook 2021: The role of digital labour platforms in transforming the world of work**. Geneva, 2021.
Reference Source

Kässi O, Lehdonvirta V: **Online labour index: Measuring the online gig economy for policy and research.** *Technol Forecasting Social Change.* 2018; **137**: 241–248.
Publisher Full Text

Kässi O, Stephany F, Lehdonvirta V: **How Many Online Workers?** *figshare.* Dataset. 2021.
http://www.doi.org/10.6084/m9.figshare.14472915.v3

Kuek SC, Paradi-Guilford C, Fayomi T, *et al.*: **The global opportunity in online outsourcing**. 2015.
Reference Source

Le Ludec C, Tubaro P, Casilli AA: **Counting 'micro-workers': societal and methodological challenges around new forms of labour.** *Work Organ Labour Glob.* 2020; **14**(1): 67–82.
Publisher Full Text

Lehdonvirta V, Hjorth I, Graham M, *et al.*: **Online labour markets and the persistence of personal networks: evidence from workers in Southeast Asia**. *Paper presented at American Sociological Association Annual Meeting 2015, Chicago, 22–25 August.* 2015.
Reference Source

Lehdonvirta V, Kässi O, Hjorth I, *et al.*: **The global platform economy: A new offshoring institution enabling emerging-economy microproviders.** *J Manage.* 2019; **45**(2): 567–599.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Melia E: **African jobs in the digital era: Export options with a focus on online labour.** (No. 3/2020). Discussion Paper, 2020.
Publisher Full Text

Ogembo D, Lehdonvirta V: **Taxing Earnings from the Platform Economy: An EU Digital Single Window for Income Data?** *British Tax Review.* 2020.
Reference Source

Pesole A, Brancati U, Fernández-Macías E, *et al.*: **Platform workers in Europe**. Luxembourg: Publications Office of the European Union, 2018.
Reference Source

Stephany F, Dunn M, Sawyer S, *et al.*: **Distancing Bonus Or Downscaling Loss? The Changing Livelihood of Us Online Workers in Times of COVID-19.** *Tijdschr Econ Soc Geogr.* 2020; **111**(3): 561–573.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wood AJ, Graham M, Lehdonvirta V, *et al.*: **Good gig, bad gig: autonomy and algorithmic control in the global gig economy.** *Work Employ Soc.* 2019a; **33**(1): 56–75.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wood AJ, Graham M, Lehdonvirta V, *et al.*: **Networked but Commodified: The (Dis)Embeddedness of Digital Labour in the Gig Economy.** *Sociol.* 2019b; **53**(5): 931–950.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔ ✔

Version 3

Reviewer Report 07 October 2021

https://doi.org/10.21956/openreseurope.15305.r27768

© **2021 Heeks R.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Richard Heeks**
[1] Global Development Institute, University of Manchester, Manchester, UK
[2] Global Development Institute, University of Manchester, Manchester, UK

Issues identified in the previous review have all been addressed.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digital economy, digital development

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 24 September 2021

https://doi.org/10.21956/openreseurope.15178.r27512

© **2021 Pesole A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Annarosa Pesole**
[1] European Commission - Joint Research Centre, Institute for Prospective Technological Studies (IPTS), Seville, Spain
[2] European Commission - Joint Research Centre, Institute for Prospective Technological Studies (IPTS), Seville, Spain

[3] European Commission - Joint Research Centre, Institute for Prospective Technological Studies (IPTS), Seville, Spain

Dear authors, thanks for your efforts.

Here some comments:

"Figure 2 summarises the performance of the predictive model graphically. As is expected, the correlation between observed and predicted numbers of registered workers is exactly 1."

Indeed, I would have been surprised if correlation was different from 1. What I wanted to suggest in my previous review, clearly I didn't express it well, it was (if possible) to use the number of registered users of (some or one) the platform for which authors also know the number of workers, impute the number of workers in the selected platform to zero and then check if the estimate number was close to the observed one. Perhaps such a test cannot be performed as I am imagining that it will be impossible to isolate one single platform and also that changing the validation set will modify the algorithmic process. I still believe that the paper could benefit from some more practical intuitions about strength and limitations of the methodology.

Some footnotes are misplaced

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Labour economics, digital labour platforms.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 02 September 2021

https://doi.org/10.21956/openreseurope.15178.r27514

© 2021 Heeks R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Richard Heeks
[1] Global Development Institute, University of Manchester, Manchester, UK
[2] Global Development Institute, University of Manchester, Manchester, UK
[3] Global Development Institute, University of Manchester, Manchester, UK

The paper has been improved but there continue to be errors:

- Despite the author response statement to have amended this ("We now clarified that we refer to "those who completed at least 10 projects OR earned at least US$1,000 in the last 30 days""), there is still an ongoing lack of clarity about the third measure of workers. In the abstract and in the first and second appearance of this criterion in the text in the Methods

section and also in Table 1 and Table 2 still lacks the ".. in the last 30 days" addition; then in the 'Inferring the number workers who have worked' section the measure is listed as "have worked over the past 30 days"; then in the results it's back to "total earnings of over $1000 or over 10 completed projects". This is inconsistent generally and also specifically ('at least 10' is not the same as 'over 10', ditto with the $ figure).

- More generally, refer to earlier comments about the third measure e.g. incorrect assumption of equivalence to full-time work.
- Methods section, should be ".. each platform: number of registered worker profiles .."
- Footnote 5 is misnumbered in the text.
- Heading should be "Inferring the number of workers who have worked".
- Results: how do you get from 9.3m and 2.2m to 75m and 21m. Explain and justify.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digital economy, digital development

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 04 Oct 2021

Fabian Stephany

Dear reviewer, thank you for your time and consideration in reading our revised manuscript and in suggesting further improvements.

We have made the following edits based on your valuable suggestions:

The third group of workers - after registered workers and workers, who have worked at least once - is now called workers, who have worked significantly. This means that these workers have, at some point, gained a significant part of their income from working on the platform. We avoid the terminology of "full-time" workers, as we can not say with certainty that these workers are currently engaged in "full-time" acitivities on the platform. Workers in this third category have either completed AT LEAST 10 projects OR have AT LEAST earned 1000 USD in the past. This definition DOES NOT include any reference to a 30 day timespan. The text has been adapted with regard to this definition.

Our Results Section concludes "...that there could at most be as many as 205 million registered worker profiles, 75 million workers who have ever worked through an online labour platform, and 21 million workers, who have worked significantly". These numbers result from multiplying the upper benchmark of the number of registered workers (205

million) with the respective upper confidence intervals for workers who have ever worked (36.8%) and workers, who have worked significantly (10%).

Spelling and footnote setting errors have been corrected.

Thanks again for your contributions and time. Best, The authors

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 16 July 2021

https://doi.org/10.21956/openreseurope.14710.r26940

© **2021 Cirillo V.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Valeria Cirillo
[1] Department of Political Science, University of Bari Aldo Moro, Bari, Italy
[2] Department of Political Science, University of Bari Aldo Moro, Bari, Italy
[3] Department of Political Science, University of Bari Aldo Moro, Bari, Italy
[4] Department of Political Science, University of Bari Aldo Moro, Bari, Italy

The paper deals with an interesting and crucial topic that is the quantification of platform workers over the world. Although platforms rely on an increasing number of workers (as aggregate statistics have shown with respect to food delivery sector during the Covid crisis), a punctual estimation of the size of platform economy is still missing. Therefore, combining data from several sources, the authors follow a two-step strategy to provide a first assessment of the number of online workers globally. In this regard, the paper is ambitious and represents one of the first attempt to provide an absolute value of individuals offering their workforce through an online digital platform. Furthermore, while many contributions focus on gig workers – that is the fraction of workers working for platforms operating in local labour markets and therefore offering tangible services in exchange of money -, the authors decide to map the large population of online freelancers often neglected by official statistics.

The paper is short and well-focused. It offers new insights on the quantification of platform economy by adopting a data-driven approach to predict the number of workers for the platforms for which this information is not available. The novelty of the approach is to collect a list of publicly available predictive features and to train a machine learning model to predict the number of workers for the platforms where this information is not available. This is a very interesting application of machine learning algorithms which deserve to be detailed in the manuscript.

The two-step strategy is crucial for the development of the paper and should be better described. The first step entails the compilation of a complete census of all online labour platforms by relying on: (i) Crunchbase; (ii) cross-regional survey in Wood et al. (2019); (iii) Google search of specific verbal locutions. Can we know which platforms have been selected? I think it is interesting to focus on online freelance economy, however the reader would like to acquire more info about these 351 platforms included. Can these platforms be grouped according to the prevailing type of activity? I suggest emphasizing this part of the research showing the main outputs.

Furthermore, the authors should provide more details on the so-called known population, that is the fraction of platforms for which information of workers is available. If platforms are grouped according to the prevailing type of activity (for example, platforms specialized in high-skilled consultancies, middle-skill or low-skill, all type of skills), some descriptive features can be traced.

There are few typos in the text that need to be corrected.

Overall, I think this is a nice paper, very useful, that can pave the way to further research aiming to quantify the size of platform work. Related to this point, few sentences in the conclusion can be added to illustrate next steps in the research.

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Labour Market and Technology; Inequalities; Occupational dynamics; Wages

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 25 Aug 2021

Fabian Stephany

Dear reviewer,
Thank your very much for your time and effort - the revised version of this work has benefited significantly from your comments. In summary, based on the suggestions made by all reviewers, we have corrected some calculation errors, made further clarifications on the rationale and procedure of our big data approach, and provided additional graphical evidence for the goodness of our estimations. In addition to the supplementary material of the initial version, we now provide a summary table listing all 351 platforms that are included in our analysis together with information on the type of online work performed on the platform, the number of observed (experienced) workers, and details on where this information has been retrieved.

In the following, we would like to respond to your individual comments:

"The two-step strategy is crucial for the development of the paper and should be better described. The first step entails the compilation of a complete census of all online labour platforms by relying on: (i) Crunchbase; (ii) cross-regional survey in Wood et al. (2019); (iii) Google search of specific verbal locutions. Can we know which platforms have been selected? I think it is interesting to focus on online freelance economy, however the reader would like to acquire more info about these 351 platforms included. Can these platforms be grouped according to the prevailing type of activity? I suggest emphasizing this part of the research showing the main outputs."
This is an important suggestion. A new table, which is provided in the data repository linked to the paper, now provides this information on all 351 platforms.

"Furthermore, the authors should provide more details on the so-called known population, that is the fraction of platforms for which information of workers is available. If platforms are grouped according to the prevailing type of activity (for example, platforms specialized in high-skilled consultancies, middle-skill or low-skill, all type of skills), some descriptive features can be traced."
This information is also provided in the new table in the data repository.

Competing Interests: No competing interests were disclosed.

Reviewer Report 21 June 2021

https://doi.org/10.21956/openreseurope.14710.r26937

© 2021 Heeks R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Richard Heeks
[1] Global Development Institute, University of Manchester, Manchester, UK
[2] Global Development Institute, University of Manchester, Manchester, UK
[3] Global Development Institute, University of Manchester, Manchester, UK
[4] Global Development Institute, University of Manchester, Manchester, UK

Gig work – short-term tasks mediated by digital platforms – is a topic of growing research, policy, and practice interest. However, its novelty, existence largely outside conventional systems of labour statistics, and the commercial confidentiality of the gig platforms themselves mean that we are often fumbling around in the dark to understand two key things: how big is this sector, and how fast is it growing (assuming that it is growing). Estimation is also hampered by definitional variance, with broader definitions coming close to those for all types of self-employment – for example, estimating that 36% of the US workforce are gig workers (McCue 2018).

This paper is therefore a very welcome addition given it attempts a better approach to estimation than we have seen previously; albeit restricted to online gig work and thus excluding those engaged with physical gig work such as seen for platforms such as Uber, Didi, Gojek, etc. The paper is relatively brief and well-focused: while situating its findings it does not drift into discussion of broader issues but keeps to the track of its core objective, to "produce an estimate of the total number of online freelancers globally".

The critique of prior estimates is valuable; for example, helping steer future research away from limitations of past work. The methods used are explicit and for the compilation of platforms and also for the observation and prediction, look robust enough. The researchers deal well and openly with the uncertainties of their approach, and provide some valuable pointers to key issues such as ratio of registered to active workers. Unfortunately, there seem to be some basic errors in figures indicated or calculated in the text of the current version; errors that need to be rectified.

It is valuable that the authors provide access to their dataset. There are a number of ways in which the paper could fairly quickly and easily provide additional value:

- I wonder if there is any logic in putting the list of platform names onto a wiki-like site that others could add to or at least comment on.

- Likewise, this could be derived from the dataset but it would be useful for readers to have a table of, say, the top 20 observed platforms by size – e.g. those in the 1m+ categories, and also an Appendix table with all 162 platforms, their type and worker nos. That could significantly increase the value and utilisation (including citation) of the paper.

- On p4, it would be helpful to name the seven and six platforms with more accessible data: researchers will find that useful.

Amendments to consider:

- p3: "data disclosed by online labour" – is that meant to be "data disclosed by online labour platforms"?

- p4: the label for Figure 1 (45 out of 151) doesn't seem to match the text – I suspect this may derive from an earlier draft of the paper.

- p5: what are the 'Number of Workers' figures in Table 1. These seem to relate to just the 162 platforms where data could be accessed. If so, it seems a little unusual to mix that data in with data from the quite different methodology of the machine learning estimates. Wouldn't the bottom half of Table 1 better be separated out and placed with the p4 material? Second, on p4 the authors state that Freelancer is one of the platforms covered by 'ever worked' and 'worked last month' – how does that fit with the max values for these

being a few tens of thousands?

- p6: maybe just explain where 159 appeared from – it is neither the difference between 162 and 351, nor the difference between 162 and (Table 1) 339.

- Figures 3 and 4: I read the text and looked at the figures several times but could not understand them. First, the x-axis – is 0.2 for example equivalent to 20% or 0.2%? Why not just use percentages. Second, what does the y-axis show: share of what? Is there also a bit of a mismatch between the discontinuous categories for the three columns, and then plotting a mean as for a continuous variable? If the red line on Figure 3 is meant to represent 5% how does that fit with it being roughly half way between "0.0" and "0.2". This could all be my ignorance but I suspect these two figures and their text explanation need reworking. It later looks like the statistic should be 11% not 5%.

- p4-p7: what is the third measure. On p4 it is those "who completed at least 10 projects, or earned at least $1000"; on p6 it is those who "have worked over the past 30 days"; and it is also referred to as being equivalent to having a full-time job. It looks like clarification is required of what the actual measure is. In addition, calling either measure as akin to a full-time job seems inappropriate unless the measure is actually "those who completed at least 10 projects or earned at least US$1,000 in the last 30 days".

- p7: "because they our results" should be "because our results".

- p7: unclear how you get from 163 million to 19 million and 5 million. p6 states: "On average, only 5% of registered users have ever worked, and 4% have worked over the past 30 days". 19 million is around 12% of 163 million; 5 million is around 3% of 163 million. Some revision or at least explanation is required. (Even if I try using the 1.83 divider and 1.21 multiplier, neither of these figures comes out and in any case and as stated on p8 these figures were not used here.) When you correct the figures on p6, I suggest you give to one decimal place, not zero decimal places.

- p8: again it looks like some basic maths has gone awry. If you multiply a number by 1.21, you do not get a smaller number!

- p8: first, it would be useful to show what the error bands are and how they were calculated. Second, why not give both the lower as well as the higher error band estimate. As with other calculations, the figures may need to be redone given the issues identified with the main calculations.

- p8: would be useful to just be quite explicit about the 'growing rapidly' point that you mean compared to the 2015 figures from Kuek et al.

- p9: you have exceeded your quota of use of the word 'nonetheless': suggest you substitute for something else in one or two places.

References

1. McCue TJ: 57 Million U.S. Workers Are Part Of The Gig Economy. *Forbes*. 2018. Reference Source

Is the work clearly and accurately presented and does it engage with the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digital economy, digital development

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Aug 2021

Fabian Stephany

Dear reviewer,
Thank you very much for your time and effort - the revised version of this work has benefited significantly from your comments. In summary, based on the suggestions made by all reviewers, we have corrected some calculation errors, made further clarifications on the rationale and procedure of our big data approach, and provided additional graphical evidence for the goodness of our estimations. In addition to the supplementary material of the initial version, we now provide a summary table listing all 351 platforms that are included in our analysis together with information on the type of online work performed on the platform, the number of observed (experienced) workers, and details on where this information has been retrieved.
In the following, we would like to respond to your individual comments:

"I wonder if there is any logic in putting the list of platform names onto a wiki-like site that others could add to or at least comment on.....Likewise, this could be derived from the dataset but it would be useful for readers to have a table of, say, the top 20 observed platforms by size – e.g. those in the 1m+ categories, and also an Appendix table with all 162 platforms, their type and

worker nos. That could significantly increase the value and utilisation (including citation) of the paper.....On p4, it would be helpful to name the seven and six platforms with more accessible data: researchers will find that useful."
A new table, which is provided in the data repository linked to the paper, now provides this information on all 351 platforms.

"p3: "data disclosed by online labour" – is that meant to be "data disclosed by online labour platforms"?"
Yes, this has now been corrected.

"p4: the label for Figure 1 (45 out of 151) doesn't seem to match the text – I suspect this may derive from an earlier draft of the paper."
Indeed, this mistake has now been corrected.

"p5: what are the 'Number of Workers' figures in Table 1. These seem to relate to just the 162 platforms where data could be accessed. If so, it seems a little unusual to mix that data in with data from the quite different methodology of the machine learning estimates. Wouldn't the bottom half of Table 1 better be separated out and placed with the p4 material? Second, on p4 the authors state that Freelancer is one of the platforms covered by 'ever worked' and 'worked last month' – how does that fit with the max values for these being a few tens of thousands?"
Thanks for addressing these relevant issues. Table 1 contains only information on observed and not on estimated data. This has now been made explicit in the text. Furthermore, the data on the row "ever worked" has been corrected accordingly - we show the share of workers in the last two rows, illustrating the skewness of the data. In addition, a new table, which is provided in the data repository linked to the paper, now provides this information on all 351 platforms.

"p6: maybe just explain where 159 appeared from – it is neither the difference between 162 and 351, nor the difference between 162 and (Table 1) 339."
The number 159 results from 162 (observed platforms) minus the three Chinese platforms 680.com, epwk.com and zbj.com, as we are sceptic that Google trends and Alexa are capturing site popularity in China well, and the Poisson XGBoost model performed much better if these three observations were excluded. This reasoning is now summarised in a footnote.

"Figures 3 and 4: I read the text and looked at the figures several times but could not understand them. First, the x-axis – is 0.2 for example equivalent to 20% or 0.2%? Why not just use percentages. Second, what does the y-axis show: share of what? Is there also a bit of a mismatch between the discontinuous categories for the three columns, and then plotting a mean as for a continuous variable? If the red line on Figure 3 is meant to represent 5% how does that fit with it being roughly half way between "0.0" and "0.2". This could all be my ignorance but I suspect these two figures and their text explanation need reworking. It later looks like the statistic should be 11% not 5%."
This is a good suggestion. The numbers are now mentioned in the text and the graphs have been excluded. We show the share of workers in the last two rows of Table 1, illustrating the skewness of the data.

"p4-p7: what is the third measure. On p4 it is those "who completed at least 10 projects, or earned at least $1000"; on p6 it is those who "have worked over the past 30 days"; and it is also referred to as being equivalent to having a full-time job. It looks like clarification is required of what the actual measure is. In addition, calling either measure as akin to a full-time job seems inappropriate unless the measure is actually "those who completed at least 10 projects or earned at least US$1,000 in the last 30 days"."
We now clarified that we refer to "those who completed at least 10 projects OR earned at least US$1,000 in the last 30 days".
"
p7: "because they our results" should be "because our results"."
This has been corrected.

"p7: unclear how you get from 163 million to 19 million and 5 million. p6 states: "On average, only 5% of registered users have ever worked, and 4% have worked over the past 30 days". 19 million is around 12% of 163 million; 5 million is around 3% of 163 million. Some revision or at least explanation is required. (Even if I try using the 1.83 divider and 1.21 multiplier, neither of these figures comes out and in any case and as stated on p8 these figures were not used here.) When you correct the figures on p6, I suggest you give to one decimal place, not zero decimal places."
Thanks for pointing this out. Our calculations have been adjusted accordingly. We estimate: 163 million registered workers of which 11.6% (18.91 or 19 million) are active and 3% (4.89 or 5 million) completed at least 10 projects OR earned at least US$1,000 in the last 30 days.

"p8: again it looks like some basic maths has gone awry. If you multiply a number by 1.21, you do not get a smaller number!"
In addition, we adjusted estimates for multi-homing (average worker is active on 1.83 platforms) and multi-working (21% of all workers do this). The number adjust accordingly, for active workers: 19 / 1.81 * 1.21 = 12.7 million and for "full-time" workers: 5 / 1.81 * 1.21 = 3.34 million.

"p8: would be useful to just be quite explicit about the 'growing rapidly' point that you mean compared to the 2015 figures from Kuek et al."
We made this more explicit in the new version.

"p9: you have exceeded your quota of use of the word 'nonetheless': suggest you substitute for something else in one or two places."
Thanks, this has been adjusted.

Competing Interests: No competing interests were disclosed.

Reviewer Report 10 June 2021

https://doi.org/10.21956/openreseurope.14710.r26939

© 2021 Pesole A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Annarosa Pesole
[1] European Commission - Joint Research Centre, Institute for Prospective Technological Studies (IPTS), Seville, Spain
[2] European Commission - Joint Research Centre, Institute for Prospective Technological Studies (IPTS), Seville, Spain
[3] European Commission - Joint Research Centre, Institute for Prospective Technological Studies (IPTS), Seville, Spain
[4] European Commission - Joint Research Centre, Institute for Prospective Technological Studies (IPTS), Seville, Spain

Despite the growing interest in platform work, as of today, there is still little information about how spread the phenomenon is. This paper partially addresses this issue looking at the freelance activities mediated by digital platforms. The paper proposes an innovative data-driven methodology to assess the number of online freelancers building on a restricted number of platforms for which information are publicly available. The proposal for this new methodology stems from the current limitations of official labour statistics in depicting platform work. As correctly stated by the authors, the lack of a common definition together with the fragmented nature of this type of work make it very difficult for traditional statistical surveys to capture the prevalence of platform work.

The authors start from compiling a list of online freelancing platforms gathering a total of 351 platforms, out of which 162 reported information about registered users and only 7 about workers who actually completed at least one task. In order to estimate the number of registered workers for those platforms that do not report the information they implement a creative solution using public data on website popularity combining information from Google trends and alexa rank.

The idea is surely captivating and shows potential for the use of big data in traditional statistics. However, the high volume and volatility of stream data may bring noise accumulation and spurious correlation, creating issues in computational feasibility and algorithmic stability (Wang *et al*.2016)[1]. Perhaps these limitations should have been better addressed in the paper and also the strategy to cope with these potential issues explained in greater details Following this procedure, the authors estimated a number of 23 million registered workers. A potential test for the goodness of the estimate could have been to repeat the same methodology for the platforms they have information and check if the predicted numbers were close to the reported ones.

As they correctly point out, not all the users who register in the platform become workers (i.e. complete at least one task). From the platforms where information is available, they compute that on average only 5% of registered users have ever worked, however the estimated parameter for workers who completed one project is 11.6% (Table 2). A valuable information to help the readers understanding better it would have been to present a table with the real data for the 7 platforms with the full set of data, so to compare the magnitude and the real share of workers against the predicted one (i.e. if between the 7 platforms were to be included the 3 outliers, almost half of the

sample (77 million) should have been multiplied by 5%).

Multi-homing and multi-working are also taken into account highlighting the additional difficulties of correctly quantifying the number of workers in online freelancing platforms. In the result section, estimates adjusted for multi-homing and multi-working are reported, although the latter appear to be wrong as they should be 12 million and 3.2 million (as reported in page 7 “we can adjust our numbers for multiple workers working under a single account by multiplying the number of workers by 1.21.”, that is 10 by 1.21 and 2.7 by 1.21)

The paper could benefit from: 1) adding a table with detailed information on the 7 platforms for which it was possible to observe workers. In particular specifying also how the information was retrieved. 2) Testing the goodness of the methodology replicating the data-driven approach for the platforms reporting the number of registered workers and comparing the estimated and the observed numbers.

Despite data limitation and the need for more clarity in some methodological steps, the paper remains an interesting study bringing valuable insights on the debate about measurement. The novelty of the approach paves the way to a new stream of research stressing the importance and the challenges associated with the use of big data. I appreciate the efforts done by the authors and my overall comments are positive.

References

1. Wang C, Chen MH, Schifano E, Wu J, et al.: Statistical methods and computing for big data.*Stat Interface*. 2016; **9** (4): 399-414 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Labour economics, digital labour platforms.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Aug 2021

Fabian Stephany

Dear reviewer,

Thank you very much for your time and effort - the revised version of this work has benefited significantly from your comments. In summary, based on the suggestions made by all reviewers, we have corrected some calculation errors, made further clarifications on the rationale and procedure of our big data approach, and provided additional graphical evidence for the goodness of our estimations. In addition to the supplementary material of the initial version, we now provide a summary table listing all 351 platforms that are included in our analysis together with information on the type of online work performed on the platform, the number of observed (experienced) workers, and details on where this information has been retrieved. In the following, we would like to respond to your individual comments:

"The idea is surely captivating and shows potential for the use of big data in traditional statistics. However, the high volume and volatility of stream data may bring noise accumulation and spurious correlation, creating issues in computational feasibility and algorithmic stability (Wang et al.2016)1. Perhaps these limitations should have been better addressed in the paper and also the strategy to cope with these potential issues explained in greater detail. Following this procedure, the authors estimated a number of 23 million registered workers. A potential test for the goodness of the estimate could have been to repeat the same methodology for the platforms they have information and check if the predicted numbers were close to the reported ones."
There is a new Figure 2 and explanation, illustrating this test for the goodness of the estimate.

"As they correctly point out, not all the users who register in the platform become workers (i.e. complete at least one task). From the platforms where information is available, they compute that on average only 5% of registered users have ever worked, however the estimated parameter for workers who completed one project is 11.6% (Table 2). A valuable information to help the readers understanding better it would have been to present a table with the real data for the 7 platforms with the full set of data, so to compare the magnitude and the real share of workers against the predicted one (i.e. if between the 7 platforms were to be included the 3 outliers, almost half of the sample (77 million) should have been multiplied by 5%)." -> A new table, which is provided in the data repository linked to the paper, now provides information on all 351 platforms: Only one of the three mentioned outliers, *freelancer.com* with 31 million workers (50877 experienced workers), is contained in the sample of 7 observed platforms. *"Multi-homing and multi-working are also taken into account highlighting the additional difficulties of correctly quantifying the number of workers in online freelancing platforms. In the result section, estimates adjusted for multihoming and multi-working are reported, although the latter appear to be wrong as they should be 12 million and 3.2 million (as reported in page 7 "we can adjust our numbers for*

multiple workers working under a single account by multiplying the number of workers by 1.21.", that is 10 by 1.21 and 2.7 by 1.21)."
This mistake has now been corrected.

Competing Interests: No competing interests were disclosed.

Reviewer Report 07 June 2021

https://doi.org/10.21956/openreseurope.14710.r26938

© 2021 Tubaro P. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Paola Tubaro
[1] Centre National de la Recherche Scientifique (CNRS), Université Paris-Saclay, Paris, France
[2] Centre National de la Recherche Scientifique (CNRS), Université Paris-Saclay, Paris, France
[3] Centre National de la Recherche Scientifique (CNRS), Université Paris-Saclay, Paris, France
[4] Centre National de la Recherche Scientifique (CNRS), Université Paris-Saclay, Paris, France

This is a really valuable paper on an important subject. As the authors say, official statistical instruments (like Labour Force Surveys) fail to capture platform-based activities that challenge established definitions of salaried employment. Particularly difficult to measure is the little visible and geographically spread activity of online workers whose activities are delivered and paid remotely over the internet. The authors endeavour to fill precisely this gap, and produce an estimate of the total number of online freelancers globally, using publicly available data gathered from internet sources. The work they propose is highly innovative and relevant, and I will highlight three main strengths.

First, while the authors (rightly) insist on the state-of-the-art machine learning techniques they leverage to estimate the number of workers on platforms that do not release this information, their work is also, to an extent, qualitative and mixed-methods. The very first step of their analysis, before any calculations, consisted in establishing a list of over 300 relevant online platforms, and gathering publicly available information about them (whether self-disclosed by the platforms or presented in other sources, e.g. the press or Wikipedia). This was done, basically, by hand, following a protocol that relies on an external, rich source (Crunchbase), a previous study of the same team, and a systematic online search. This should perhaps be recognized more explicitly, not only to give credit to the painstaking work of compiling the list, but also to discuss the issues that may have arisen – and that may add to the authors' discussion of the uncertainties related to the calculation. Surely the authors had to make exclusion/inclusion decisions especially in ambiguous cases, such as platforms offering both local gigs and remote tasks (like Bemyeye), or platforms operating rather as traditional BPO firms (like Playment). A discussion of these issues as is often done in qualitative research (such as decision by consensus in the team, or majority choice...), may be helpful.

Similarly, the authors must have had to make decisions on how to deal with heterogeneity of information released by platforms. As they notice, some platforms do not disclose any information at all, while others just provide number of registrations, not levels of engagement and activity. But even when they disclose information, platforms do not do so uniformly. Some give just orders of magnitude (such as Clickworker.com, 2200k workers in 2020) and update them very irregularly (such as Amazon, whose latest figures for Mechanical Turk date back to the early 2010s), others provide very precise numbers (as Microworkers.com which discloses its up-to-date numbers daily on its home page). Again, researchers have to make choices – take whatever platforms provide, approximations, or as precise figures as possible? And what happens when sources disagree?

These issues illustrate well the 'costs', so to speak, of relying on disparate online sources, which unlike surveys, do not allow relying on shared definitions and/or prior knowledge of sampling distributions. Yet this is the only possible solution when surveys cannot help, and has the merit of using publicly-available information that allows for check and replication. In the data economy that thrives through platforms, one part of the world of work escapes established definitions and misses the gaze of official statistics. This is why we need to devise creative ways of observing the less-and-less observable.

Another merit of the paper is to distinguish between registrations and actual levels of engagement. One may register on a platform for mere curiosity or to explore it as a journalist or as a researcher – as many of us indeed do. Additionally, some platforms such as Microworkers.com do not differentiate the two sides of the market, so that the total number of their registered users includes both workers and their clients/employers. Hence, the authors are right to count separately the users who worked at least once, and to further single out those who did at least 10 assignments or earned more than $1000. Indeed platform labour is not always an individual's main activity and may constitute just a side hustle, or an occasional buffer in periods of unemployment.

While this diversity of levels of engagement is widely recognized in the nascent literature on digital platform labour, it is not always operationalized in the same way. For example, Gray and Suri (2019, p. 104)[1] distinguish three groups, the 'experimentalists', 'regulars', and 'always-on', while Urzi-Brancati *et al.* (2020, p. 15)[2] propose a four-level classification with 'sporadic', 'marginal', 'secondary' and 'main' platform workers. Admittedly, a more homogeneous approach is out of reach because these studies differ in scope (as the latter, for example, also includes location-based platform workers) and types of available data. Nevertheless, it would be helpful to situate more explicitly the authors' classification with respect to these other attempts in the literature. Also, they might want to reconsider their terminology, especially the characterization of 10 assignments/ $1000 as a 'full time job' which seems a bit too far stretched.

Finally, I appreciate the authors' capacity to take into account user's concrete practices as reflected in multi-homing (simultaneous use of multiple platforms) and account sharing (more than one person working on the same account). The difficulty, as they rightly stress, is to accurately quantify them, because they have only been observed in non-representative samples. Perhaps it could be added that the question of the discrepancy between registration and real usage arises again with multi-homing, as a worker may have a different level of activity on each platform. Account-sharing also raises specific questions because it is formally forbidden on several platforms: in these cases, it may go under-reported, or perhaps it may just be limited to highly active and experienced users who have devised ways to conceal it. In passing, it is interesting to notice how smaller and even

qualitative studies were necessary to retrieve this information, although the design of the paper is essentially quantitative.

A final comment concerning the global estimate of 163M online workers: are these numbers large or small? The answer depends on the perspective taken and the goals – some will find this change huge, others will wonder why bother for such a tiny part of global production. Nevertheless, these estimates demonstrate that this population exists and needs attention – despite its limited visibility, fuzzy definition and contested boundaries. According to the authors, it has also grown rapidly. Future research will have to combine this result with further evidence on the geographical spread of this population and any effects of the COVID-19 pandemic that may have accelerated its growth.

A minor question:
In the Results section, it is said 'Further adjusting for multi-homing, these numbers reduce to 10 million and 2.7 million, respectively. Finally, adjusting for possible multi-working increases these numbers to 8.5 million and 2.3 million, respectively'. The last two figures do not seem right as numbers should increase rather than decrease here.

References

1. Gray ML, Suri S: Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. *Boston, MA: Houghton Mifflin Harcourt*. 2019.
2. Pesole A, Urzí Brancati MC, Fernández-Macías E, Joint Research Centre (European Commission): New evidence on platform workers in Europe: Results from the second COLLEEM survey. *Publications Office of the European Union*. 2020; **JRC 118570**. Publisher Full Text

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digital platform labour.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 25 Aug 2021

Fabian Stephany

Dear reviewer,

Thank you very much for your time and effort - the revised version of this work has benefited significantly from your comments. In summary, based on the suggestions made by all reviewers, we have corrected some calculation errors, made further clarifications on the rationale and procedure of our big data approach, and provided additional graphical evidence for the goodness of our estimations. In addition to the supplementary material of the initial version, we now provide a summary table listing all 351 platforms that are included in our analysis together with information on the type of online work performed on the platform, the number of observed (experienced) workers, and details on where this information has been retrieved. In the following, we would like to respond to your individual comments:

"The very first step of their analysis, before any calculations, consisted in establishing a list of over 300 relevant online platforms, and gathering publicly available information about them (whether self-disclosed by the platforms or presented in other sources, e.g. the press or Wikipedia). This was done, basically, by hand, following a protocol that relies on an external, rich source (Crunchbase), a previous study of the same team, and a systematic online search. This should perhaps be recognized more explicitly, not only to give credit to the painstaking work of compiling the list, but also to discuss the issues that may have arisen – and that may add to the authors' discussion of the uncertainties related to the calculation."
We have added a paragraph that recognises the platform selection process via Crunchbase in greater detail.

"Similarly, the authors must have had to make decisions on how to deal with heterogeneity of information released by platforms.... Again, researchers have to make choices – take whatever platforms provide, approximations, or as precise figures as possible? And what happens when sources disagree?" A new table, which is provided in the data repository linked to the paper, now provides this information on all 351 platforms. "Finally, I appreciate the authors' capacity to take into account user's concrete practices as reflected in multi-homing.... Perhaps it could be added that the question of the discrepancy between registration and real usage arises again with multi-homing, as a worker may have a different level of activity on each platform. Account-sharing also raises specific questions because it is formally forbidden on several platforms: in these cases, it may go under-reported, or perhaps it may just be limited to highly active and experienced users who have devised ways to conceal it."
This is an important limitation, which has now been acknowledged in the revised version. "In the Results section, it is said 'Further adjusting for multi-homing, these numbers reduce to 10 million and 2.7 million, respectively. Finally, adjusting for possible multi-working increases these numbers to 8.5 million and 2.3 million, respectively'. The last two figures do not seem right as numbers should increase rather than decrease here." This mistake has now been corrected.

Competing Interests: No competing interests were disclosed.