The researcher turk in action: experiences from the LLTC4J project

Steffen Herbold
steffen.herbold@tu-clausthal.de
TU Clausthal
Clausthal-Zellerfeld, Germany

Alexander Trautsch
atrautsch@cs.uni-goettingen.de
University of Goettingen
Göttingen, Germany

Benjamin Ledel
benjamin.ledel@tu-clausthal.de
TU Clausthal
Clausthal-Zellerfeld, Germany

ABSTRACT

The researcher turk is a concept to conduct large-scale research projects through crowd working. Within this paper, we describe our experience with the first application of this concept within the Line Labeling Tangled Commits for Java (LLTC4J) project.

CCS CONCEPTS

• **General and reference**;

KEYWORDS

researcher turk, recruitment, registered reports

ACM Reference Format:
Steffen Herbold, Alexander Trautsch, and Benjamin Ledel. 2018. The researcher turk in action: experiences from the LLTC4J project. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The concept of the researcher turk was proposed by Herbold [1] as a means to recruit participants to solve complex research tasks. The goal of the researcher turk was to solve the problem of recruiting highly skilled individuals, that cannot be easily found on crowd working platforms like the MTurk.[1] The researcher turk proposes to utilize open science principles to solve this problem: the *principal investigators (PIs)* register a study protocol and include requirements for participation and authorship within the protocol. Interested researchers can join this study and help to conduct the research project as *participants*. The study protocol serves as "contract" between the PIs and participants, i.e., clarifies the minimal requirements for joining a project, for rewards (e.g., authorship, acknowledgements), and the duties of the participants and PIs (e.g., technical work, writing, reviewing drafts).

Within this paper, we report on our first experience with the researcher turk. We registered a study on the Line Labeling of Tangled Commits for Java (LLTC4J) [2]. We completed this study after successfully recruiting 45 participants,[2] through which we achieved a scale which would not have been possible without involving such a large group [3]. In the following, we briefly summarize the relevant parts of the study, how we recruited participants, as well as the potential and challenges regarding recruitment of the researcher turk.

2 STUDY SUMMARY

For the study, the PIs provided a website through which registered participants could label textual differences in bug fixing commits, to mark which lines contributed to the bug fix, and which lines were tangled changes [4]. The protocol contained a detailed description of the labels, data, and analysis, which we omit here because they are not relevant for this experience report. Moreover, the protocol defined two criteria for eligibility to participate in the study: either (at least) an under-graduate degree majoring in computer science or a closely related subject or at least one year of programming experience in Java. Participating researchers were offered to become authors of the resulting manuscript, if they contributed by labeling at least 200 commits and if they contribute to the writing by reviewing the draft of the manuscript, including potential revision. Additionally, there was a requirement that should ensure that participants (and all produced data) may be dropped from the study, if the labels have a poor agreement with other participants, indicating that there would be a lack of required expertise or even malicious mislabeling.

We recruited participants from May 2020 until October 2020. Due to the Covid-19 pandemic, recruitment was conducted online. We regularly shared the call for participation on Twitter throughout the whole recruitment period. We asked participants who already registered to share the call, which served to amplify our signal and led to a clear increase in participants. We also shared our call for participation in Facebook groups related to software engineering research, which attracted more participants. During conferences, we used the public chat channels of the conference (e.g., Slack) to advertise for our study. Moreover, we had presentations at the MSR 2020 and the ICSE 2020 about the registration [2] and the researcher turk [1]. Overall, the online recruitment worked fairly well and 79 researchers registered for participation in our study. All media we used (virtual conferences, Twitter, Facebook) helped us to attract more participants. Of the 79 researcher, 45 fulfilled the requirements and became authors. 15 participants dropped out without labeling any data, 19 participants labeled only few commits. More details on the recruitment and the labeling can be found in our study [3].

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

[2] Together with the three PIs this meant 48 researchers total.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

3 POTENTIAL OF THE RESEARCHER TURK

In general, we are happy with our first researcher turk study. The pre-registration served as quality assurance early in the process, both for the study, but also for the recruitment criteria. The open recruitment enabled any qualified researcher to join, which is good from an equity point of view: no connections, recommendations, or similar are required to join the project: you like it, you join. This enabled us to get a globally distributed project team with authors from 17 different countries from five continent.[3] We also achieved a scale that would not have been possible otherwise. The internal peer-review with so many authors was great for shaping the manuscript, because we involved so many perspectives. This involvement also leads to a networking effect, obviously for the PIs, but we are also aware of new collaborations between participants.

Based on this experience, we believe that the researcher turk can also be valuable in the future. While we demonstrated the potential for manual validation, we believe other types of research can benefit as well. Surveys could be scaled up this way. Participants could be required to recruit subjects for a user study from their network. This could help to generalize studies beyond local settings, by involving not only more people in the study, but also by globally distributing the study through the recruitment of participating researchers. Of course, reality is more complex than this, especially sampling issues are harder to control for with such an approach. Still, Ralph et al. [6] used a similar approach (without pre-registration) to scale out their global pandemic survey with great success.

Benchmarks could also become community projects. This is already done by as part of challenges (e.g., [5]). These challenges are currently done with a "bring your own tool" design. An alternative would be to pre-register a detailed benchmark protocol, including which techniques should be compared to each other. Participants could register and get techniques assigned, which they need to implement for the benchmark.

4 CHALLENGES AND ISSUES

The recruitment must be designed in a way that is ethical: authorship should not be tossed around without meaningful contributions. For example, just sharing a link within a community should not earn authorship. The underlying question, how much is enough to earn authorship is too complex to be solved in such a short paper. Instead, we believe that the review of the pre-registration should be used to critically evaluate this. Hence, it should be part of the reviewer's duties to comment on whether they believe the amount of work required for participation is sufficient for authorship. We note that the researcher turk also requires that the requirements are compatible with authorship guidelines [1]. However, these guidelines use undefined terms like "significant contribution."[4]

Another challenge was that our requirements for authorship overlooked one aspect: we did not exclude the reviewers of the pre-registration. If they could become authors later on, there would be a conflict of interest with the reviews of the protocol. The core problem here is that this cannot be checked by the PIs, because they do not know the reviewers. Since 2021, the submission guidelines of the registered reports at MSR and ICSME were extended to cover this case, and prohibit former reviewers and their directly supervised students to become authors of the same study. What happens if the students are not aware that their supervisors reviewed the pre-registration and they register regardless, is an unresolved issue. Keeping them is a potential conflict of interest. Dropping them after this is discovered after submission of the manuscript is potentially unfair towards the student and may also unblind a reviewer. Conflicts are also a problem for reviews: the large group of authors mean a large group of current (and future) conflicts, which restricts the pool of potential reviewers.

We also received complaints regarding advertising for participation based on the promise of authorship, especially in combination with stating the protocol is pre-registered with an in-principal acceptance (IPA) at a certain journal. While we agree that this is unconventional, we argue that this happens everywhere, when people get together to do research: you work together, you publish together. Instead, we would argue that the researcher turk makes unwarranted authorship harder. The minimal requirements for everybody to become an author are actually peer reviewed, while the actual contributions of individual authors is usually opaque for "normal" research collaborations. We also do not believe that authorship as "payment" makes this unethical for the same reason: it is the same as with every collaboration, except that the recruitment is different, based on a public and open offer.

Beyond recruitment, running a project is this manner also comes with challenges. For example, malicious participants could falsify data. We countered this by requiring multiple participants to label each data point. In other settings, this approach is not possibly. E.g., consider a distributed survey, where each participants collects their own data, which is later aggregated. Usually, such studies are made with a group of already known people, in which trust is already established. This trust may be missing in a research turk project and is hard to replace. This could possibly be countered by more stringent requirements, e.g., an academic track record of some sort or required references from personally known researchers. However, this would run counter the the inclusiveness that we were able to achieve.

PIs of research turk projects also need to find a balance between the different input they receive from participants, e.g., as part of the peer-review. The PIs need to be open for advice and integrate improvements. The difficulty here is that with the number of opinions about study design and writing naturally increases with the number of participants: imagine getting getting 45 reviews from a journal with the request to prepare a major revision. There are naturally opposing views on issues that need to be mediated, especially since these are not suggestions from a third party (journal reviewer), but from co-authors, who need to agree to everything that is submitted.

5 CONCLUSION

The researcher turk has a great potential for scaling up research projects. However, there are several ethical challenges involved, especially due to conflicts of interests which may arise through rewarding participation with authorship.

[3] South America and Antarctica are missing.
[4] https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/ethical-requirements/#authorship

REFERENCES

[1] Steffen Herbold. 2020. With Registered Reports towards Large Scale Data Curation. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results* (Seoul, South Korea) *(ICSE-NIER '20)*. Association for Computing Machinery, New York, NY, USA, 93–96. https://doi.org/10.1145/3377816.3381721

[2] Steffen Herbold, Alexander Trautsch, and Benjamin Ledel. 2020. Large-Scale Manual Validation of Bugfixing Changes. https://doi.org/10.17605/OSF.IO/ACNWK

[3] Steffen Herbold, Alexander Trautsch, Benjamin Ledel, Alireza Aghamohammadi, Taher Ahmed Ghaleb, Kuljit Kaur Chahal, Tim Bossenmaier, Bhaveet Nagaria, Philip Makedonski, Matin Nili Ahmadabadi, Kristof Szabados, Helge Spieker, Matej Madeja, Nathaniel Hoy, Valentina Lenarduzzi, Shangwen Wang, Gema Rodríguez-Pérez, Ricardo Colomo-Palacios, Roberto Verdecchia, Paramvir Singh, Yihao Qin, Debasish Chakroborti, Willard Davis, Vijay Walunj, Hongjun Wu, Diego Marcilio, Omar Alam, Abdullah Aldaeej, Idan Amit, Burak Turhan, Simon Eismann, Anna-Katharina Wickert, Ivano Malavolta, Matus Sulir, Fatemeh Fard, Austin Z. Henley, Stratos Kourtzanidis, Eray Tuzun, Christoph Treude, Simin Maleki Shamasbi, Ivan Pashchenko, Marvin Wyrich, James Davis, Alexander Serebrenik, Ella Albrecht, Ethem Utku Aktas, Daniel Strüber, and Johannes Erbel. 2021. A Fine-grained Data Set and Analysis of Tangling in Bug Fixing Commits. arXiv:2011.06244 [cs.SE]

[4] Kim Herzig and Andreas Zeller. 2013. The Impact of Tangled Code Changes. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (San Francisco, CA, USA) *(MSR '13)*. IEEE Press, 121–130.

[5] Sebastiano Panichella, Alessio Gambi, Fiorella Zampetti, and Vincenzo Riccio. 2021. SBST Tool Competition 2021. In *2021 IEEE/ACM 14th International Workshop on Search-Based Software Testing (SBST)*. 20–27. https://doi.org/10.1109/SBST52555.2021.00011

[6] Paul Ralph, Sebastian Baltes, Gianisa Adisaputri, Richard Torkar, Vladimir Kovalenko, Marcos Kalinowski, Nicole Novielli, Shin Yoo, Xavier Devroey, Xin Tan, Minghui Zhou, Burak Turhan, Rashina Hoda, Hideaki Hata, Gregorio Robles, Amin Milani Fard, and Rana Alkadhi. 2020. Pandemic programming. *Empirical Software Engineering* 25, 6 (Sept. 2020), 4927–4961. https://doi.org/10.1007/s10664-020-09875-y