Challenges and Lessons from Working with Data Collected by Crowdfunding in the Wild

Anja Exler
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
exler@teco.edu

Andrea Schankin
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
schankin@teco.edu

Matthias Budde
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
budde@teco.edu

Till Riedel
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
riedel@teco.edu

Erik Pescara
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
pescara@teco.edu

Michael Beigl
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
michael@teco.edu

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore
ACM 978-1-4503-5966-5/18/10.
https://doi.org/10.1145/3267305.3267545

Abstract

The rise of the smartphone opens up new possibilities for researchers to observe users in everyday life situations. Researchers from diverse disciplines use in-field studies to gain new insights into user behavior and experiences. However, the collected datasets are mostly not available to the public and thus results are neither falsifiable nor reproducible. Community datasets attempt to counter this problem, i.a. by sharing the cost of the data collection. One example is the crowdfunded campaign *CrowdSignals*. In this paper, we report on our experiences in doing research with crowdfunded data, drawing on the example of this dataset. By “zooming into” specific aspects of the data, we juxtapose the expectations we had when co-funding the data collection with our findings when analyzing the dataset. We highlight shortcomings and benefits of crowdfunded datasets, draw lessons and discuss how future crowdsourced data collection campaigns might be improved.

Author Keywords

Crowdfunded Data Collection; Community Dataset; Experience Sampling; Ecological Momentary Assessment; Activity Recognition; Data Quality; Challenges; Lessons

ACM Classification Keywords

H.3.5 [Online Information Services]: Data Sharing

Introduction

In science, we often face the issue of not being able to compare results from different methods due to unavailability of the dataset. There is a need for public datasets that are large enough to allow statistical analyses on the one side and that are usable for a broad range of researchers on the other side. Some datasets for machine learning purposes are already available, e.g. in the UCI Machine Learning Repository [6]. However, to our best knowledge, there is no dataset of rich smartphone data connected to user behavior data. Experience sampling during everyday life activities in natural environments, a.k.a. ecological momentary assessment (EMA), is a common means to gather such data. That is, users receive smartphone notifications, prompting them to answer usually short self-report surveys about their current daily experiences such as well-being or activities.

To have a representative and reliable dataset that allows conducting statistical analyses, it is necessary to draw a large and representative sample and to gather data over an appropriate amount of time. This is cost and time consuming as it includes, e.g., app creation and distribution, participant acquisition, supervision and compensation. It is tempting to bring the community together, collect money and delegate this task. AlgoSnap, an enterprise focusing on data-driven research and intelligent algorithms, dared to undertake a first attempt to collect a community dataset: they initiated and organized the crowdfunded campaign *CrowdSignals*[1]. Their objective was to collect a large dataset consisting of labeled mobile and sensor data collected via smartphone and smartwatches. We backed this great idea by financially supporting the campaign. In this paper we present our experiences with this dataset. We discuss pitfalls and present ideas on how future crowdfunded data assessment campaigns might avoid them.

The CrowdSignals.io Community Dataset

The *CrowdSignals* dataset consists of two different kinds of data: (1) mobile sensor data gathered from smartphone and smartwatch sensors and (2) ground truth labels provided by participants via survey responses. Sensor data was gathered over a period of three weeks and consists of information[2] about geo-location, social factors, system and networking, user-device-interaction and motion.

Survey responses provide ground truth about user demographics, place labels, contact labels, activity intervals, and situational information such as well-being. They were assessed using ecological momentary assessment (EMA) and lock-screen surveys (similar to [7]). In addition, participants were free to provide labels voluntarily at any time.

We received data from 31 participants, 11 of them female. Each of them owned a different Android smartphone. 11 participants were enrolled as a student. The participants' age ranged from 18 to 69 with an average of about 37 years. Their ethic backgrounds, marital statuses, physical exercise level, and health level were manifold. The dataset can be assumed to be reasonably representative for many different study applications.

The final dataset consisted of more than 150GB of data containing 1000 interval labels and over 3000 lock-screen survey responses" [1]. This large number of data seems promising for analyses of correlations among smartphone features, among ground truth labels, and between smartphone features and ground truth labels.

User-Requested Labels

To fit the needs of researches from different communities (IoT, DataScience, UbiComp, Sensors, Networks/Systems [1]), the *CrowdSignals* campaign offered a so-called

[1] http://crowdsignals.io/

[2] http://crowdsignals.io/#data

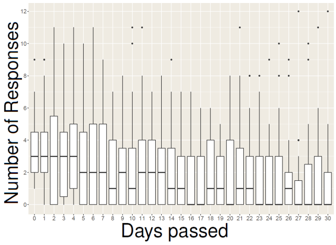

Figure 2: Overview of the number of interruptibility survey responses (y axis) over time (x axis).

User	Mean (± SD)	Min	Max
1	0.10(±0.40)	0	2
2	0.35(±0.75)	0	3
6	0.84(±1.49)	0	5
8	4.58(±3.27)	0	11
9	1.65(±2.44)	0	9
10	0.19(±0.65)	0	3
11	3.74(±3.04)	0	11
12	1.52(±1.84)	0	6
16	1.77(±1.54)	0	5
19	0.35(±0.75)	0	3
20	4.32(±2.36)	0	8
21	3.32(±1.80)	0	7
23	3.48(±1.57)	1	6
24	5.55(±1.69)	3	9
25	2.06(±1.48)	0	6
26	1.74(±3.02)	0	10
27	2.61(±2.17)	0	8
28	0.97(±1.20)	0	4
29	2.61(±1.33)	1	7
30	1.68(±1.74)	0	6
31	0.06(±0.25)	0	1
32	0.03(±0.18)	0	1
34	1.42(±1.20)	0	4
35	2.71(±2.15)	0	10
36	0.71(±1.30)	0	5
37	2.90(±1.33)	1	6
38	4.68(±1.97)	0	8
39	1.39(±1.28)	0	4
41	4.32(±4.87)	0	12
42	3.39(±4.03)	0	11
43	1.97(±2.76)	0	8

Table 1: Average, minimum, and maximum number of interruptibility survey responses per user per day.

“Guarantee Your Label” package, which allowed introducing an additional ground truth label to the data collection.

In agreement with the organizers, we decided to assess the label *interruptibility*, a factor that was considered in related work before [9]. In the context of human computer interaction this might be interpreted as a probability with which it is acceptable for the user to be interrupted by the computing device during their current task [4, 5]. We agreed that *Interruptibility* shall be assessed by six daily EMA surveys and as part of the lock-screen responses with a probability of 50% to be one of the two to four questions to be displayed.

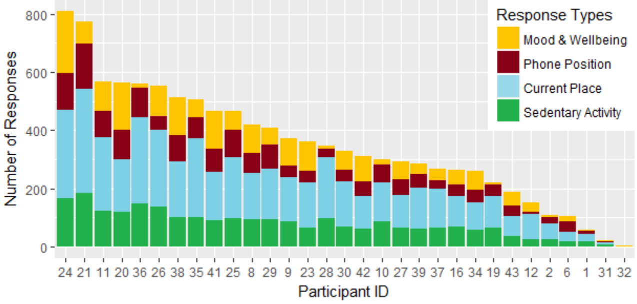

Figure 1: Overview of the number of survey responses for each label and for each participant.

Looking into the Data

When analyzing the dataset, we found several characteristics that might influence the data analysis:

1. The dataset is very sparse for all labels (see Figure 1), but especially for interruptibility labels (see Figure 2 and Table 1). For 8 users, we have less than 30 data points, for one participant only 1 overall.
2. The survey items were selected randomly so that there are very few instances in which a label for interruptibility and for another survey item were given at the same time (see Table 2).
3. Due to the random selection of survey items, the share of data points per labels is unbalanced (see Figure 1) and intensifies with each additional label.
4. The response rates were not tracked. We know how many survey items were answered, but not how many survey prompts were sent out. However, it is visible that the engagement varied among users and decreased over time (see Figure 2).
5. The dataset is missing synchronized timestamps which makes it difficult to analyze for correlations.

Feedback From Backers and Supporters

To have a broader impression of the dataset we contacted supporters and backers [2]. Seeking qualitative feedback, we asked the following questions:

- Did you use the final *CrowdSignals* dataset for your research (yet)? What kind of research are you doing with it (or planning to do)? (e.g., activity recognition, position detection, well-being correlation analysis, ...)
- What did you expect from the dataset when backing? Did the final dataset meet these expectations? If not, what was different than expected?

Eight out of 35 contacted people responded to our request. Two of them were supporters with no intention to use the dataset. Six persons actually backed the project. Four of them had not yet have time to use the dataset. Two parties looked into the dataset, one of them already published their results [8].

Backers intend(ed) to use the data for:

- Activity recognition, e.g., in public transportation
- Human movement analysis for public health, e.g., to build behavior models for everyday activities

- Indoor / outdoor location detection, e.g., for automatic emergency calling or for broadband speed adjustment
- Evaluation of existing and new algorithms for data series management and analytics, including data series indexing, data series similarity search, frequent pattern identification and outlier detection

We see that this is a broad field of application areas. It is tough for a dataset to fulfill these requirements.

Survey item	# occurrences of synchronous labels	Median time to next answered interruptibility item
Current Place	17	63 min
Sedentary Activity	6	187 min
Phone Position	2	287 min
Mood & Wellbeing	2	119 min

Table 2: Overview about how many survey questions were answered at the same time as an interruptibility question and about the time span between an answered survey question and interruptibility question.

The backers we asked had different expectations. For some of them the dataset met their expectations: they were "looking for a set of diverse real data series". However, some of them had identified issues with the dataset. They reported shifted timestamps that required additional data pre-processing. In addition, they mentioned that the ground truth about activities was partly missing or incorrect which inherently complicated machine learning and real-time activity recognition. There are also researchers for which both applies: they were satisfied with the activity part of the dataset, but would have desired more contexts (labels) or situation label such as “crowdedness”.

Overall, the usefulness of the dataset seems to depend on the research question. Some issues may arise for those who need correct and synchronized timestamps as well as correct data labeling.

Our Expectations

We primarily wanted to use the dataset to correlate interruptibility with other ground truth labels [1] as well as sensor measurements. Special interest lay on the current place and transportation mode, mood and physical well-being, smartphone position, and sedentary activity.

To allow the correlation analysis we intended to run, we required a large set of data labeled for interruptibility and, at the same time, for other labels such as well-being. That is, EMA or lock-screen surveys needed to ask for interruptibility as well as at least one other label. With at least 6 (EMA) but probably more (lock-screen surveys) prompts per day over a period of 30 days, we expected to end up with at least 180 interruptibility data points for each of the 30 participants and 5400 data points overall. We expected to have less of the remaining labels, but still enough to allow correlation analyses.

Discussion

General Issues When Conducting Field Studies

Indisputable, studies in the field under natural conditions provide a high ecological validity. But running those studies is always risky, in particular because the behavior of the participants can usually not be observed or even controlled, resulting in a low internal validity. For example, it remains unclear whether, when, or how often participants will respond to survey prompts, whether they will give true or socially desired responses, or how they will deal with any technical problems [3]. It is possible that different participants interpret the study task such as labeling activities differently which might lead to inconsistent data. Participants might even interpret the labels itself differently, e.g., when assigning place types to locations. It is also an open issue how to engage participants for a longer time period, avoiding a drop in data quality (cf. Figure 2 and 1) or sparse feedback in general (cf. Table 1). Also, participants know that they are part of an experiment and they may behave differently than usual. This raises the question how representative user responses will be for the behavior of interest.

- Thorough and correct timestamping
- Synchronized assessment of multiple survey items
- Tracking of response rates
- Correct labeling of data streams
- Mechanisms to motivate participants / keep them engaged
- Weigh focus against universality of expected dataset
- More information on the environment / surrounding?
- Larger dataset (more than 30 participants; more than 30 days)?

Table 3: Suggestions for future crowdsourcing projects.

Specific Issues For Crowdfunded User Studies

First of all, we are facing the extra label paradox: the more people buy an extra label (i.e., the more money is available), the sparser the dataset gets. This might be useful if your objective is explorative data analysis, i.e., you want a large dataset to explore correlations and draw hypotheses. However, if you already have research questions and hypotheses to be evaluated, a denser dataset with less labels might be better, i.e. having multiple smaller user studies which focus on one extra label each instead of having one big user study trying to cover it all.

Secondly, despite monetary incentivization, participants' response rates were rather low. Shorter surveys, i.e. less labels that are faster to fill in might lead to a better user experience and more responses. In addition, gamification methods might be employed to enhance user motivation and keep compliance high over longer periods of time.

Thirdly, timestamps were not synchronized and the start and end times of labels were not always unambiguous, which makes data cleaning necessary. This is partly an implementation issue that can be fixed. Though, it is also an issue that comes along with user-based data labeling: each user had a different understanding of the start and end times of an activity. Moreover, participants were allowed to define the end of an interval activity earlier if they felt the need to do so, e.g. to save battery power. Better user instructions and less battery drain, which goes together with assessment of less sensor data or a less frequent data acquisition, might improve the results.

Speaking of sensor data acquisition: sometimes, it is a good idea to let the device label data automatically to avoid user-dependent interpretation. This might be useful for place labels. For example, some participants might call a "McDonalds" a "restaurant", others a "meal take-away" or simply "fast food store". In this case, it might be worth to rely on automatically gathered data instead of user-provided ones if possible, e.g. using place types provided by the Google Places API[3] instead of user-provided labels to guarantee the same label for the same place.

Lessons for Crowdfunded Data Collection

Based on our own experiences we learned the following lessons:

- "Too many cooks spoil the broth": too many additional labels are a burden for the participant and decrease data quality: focus on as few labels at a time as possible
- "Sometimes less is more": focus on one specific use case instead of trying to create a jack of all trades dataset
- "Let the machine work for you": rely on objective, automatically gathered data wherever possible to counteract user-inflicted labeling flaws
- "No delegation without communication": if you want to go the easy way and let others do the data collection work, make sure to have a good communication and check regularly that you are talking about the same things
- "Take the time for a dry run": play the study through with a handful of participants to have a feeling for the data that you will get once you run the real user study and to see what might be missing
- "The show must go on": utilize gamification methods to keep the motivation of the participants high and reduce the churn rate

Suggested improvements specific for the case study are listed in Table 3.

[3] https://developers.google.com/places/supported_types?hl=en

Summary

In this paper, we presented our experiences with the *CrowdSignals* dataset, talked about our expectations and impressions of the dataset, and share some lessons learned.

The dataset is among the first of its kind. We wholeheartedly acknowledge the effort that was made to realize this campaign. Crowdsourced datasets have specific issues that are partly caused by the nature of in-field user studies, by over-engaged projects aims, or by subobtimal communication of expectations on the part of the future users. Considering our personal objective of label correlation analysis, the dataset is pretty sparse which is due to various factors such as randomized survey item selection, too many or too large surveys, and decreasing participant commitment over time – which might be interdependent. On the other hand, the dataset is rich in sensor data gathered from smartphones and smartwatches. One benefit of the *CrowdSignals* dataset is that the sample of participants is manifold. Though, some backers state that it would be nice to have more than 30 people for a longer period of time.

In the end the question is, how we as a research community can work together to collect good datasets. For which applications is crowdfunding a suitable solution? Do incentives (monetary or other) lead to higher user compliance, to more frequent and thorough labeling and, eventually, to better data quality? Would a longer user study with a higher number of participants, both possible with more funding, allow for a better data analysis, more insights, or more reliable results? Or should the focus rather be narrowed? We welcome a lively discussion during the workshop.

Acknowledgements

Partially funded by the German Federal Ministry of Education and Research (BMBF) within the *Smart Data Innovation Lab (SDIL)*, grant no. 01IS15035A. We thank all backers/supporters of the campaign who responded to our survey.

REFERENCES

1. AlgoSnap Inc. 2016a. AlgoSnap - CrowdSignals.io Pilot Dataset Reference. published as part of the final dataset; not yet available online. (2016).
2. AlgoSnap Inc. 2016b. CrowdSignals.io: A Massive New Mobile Data Collection Campaign – Experts. `http://crowdsignals.io/#experts`. (2016). accessed on June 30th, 2017.
3. Matthias Budde, Andrea Schankin, Julien Hoffmann, Marcel Danz, Till Riedel, and Michael Beigl. 2017. Participatory Sensing or Participatory Nonsense? – Mitigating the Effect of Human Error on Data Quality in Citizen Science. *IMWUT* 1, 3 (2017).
4. Anja Exler, Marcel Braith, Andrea Schankin, and Michael Beigl. 2016. Preliminary investigations about interruptibility of smartphone users at specific place types. In *Ubicomp'16 Adjunct*.
5. Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. In *Ubicomp'14*.
6. UCI. 2016. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml/index.php`. (2016). accessed on June 30th, 2017.
7. Rajan Vaish, Keith Wyngarden, Jingshu Chen, Brandon Cheung, and Michael S Bernstein. 2014. Twitch crowdsourcing: crowd contributions in short bursts of time. In *CHI'14*.
8. Megha Vij, Venkata MV Gunturi, and Vinayak Naik. 2017. Use of ECDF-based Features and Ensemble of Classifiers to Accurately Detect Mobility Activities of People using Accelerometers. (2017).
9. Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017. Predicting interruptibility for manual data collection: a cluster-based user model. In *MobileHCI'17*. ACM, 12.