Expectation vs. Reality – Pitfalls in Working with Crowdfunded Data Collection Exemplified by a Case Study on the CrowdSignals Dataset

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

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. This might be countered by providing a community dataset. 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 backing the data collection campaign with our findings when analyzing the dataset. We highlight shortcomings of the dataset for our intended research purposes and discuss how future crowdsourced data collection campaigns might be improved.

Author Keywords

Crowdfunded Data Collection; Community Dataset; Experience Sampling; Ecological Momentary Assessment; Failures

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 [11]. 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, also known as 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 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 [1], 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* [3]. 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 Community Dataset

The dataset consists of two different kinds of data: (1) mobile data gathered from smartphone and smartwatch sensors and (2) ground truth labels provided by participants via survey responses.

Sensor data consists of information about geo-location, social factors, system and networking, user-device-interaction and motion [4]. It was gathered by accessing common Android system APIs.

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 [12]). In addition, participants were free to provide labels voluntarily at any time.

We received datasets 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 age of about 37 years. Their ethic backgrounds, martial statuses, physical exercise level, and health level were manifold. The dataset seems representative for common smartphone users.

The final dataset consisted of "data gathered from 30 participants over a timeframe of 30 days" with "more than 150GB of data containing 1000 interval labels and over 3000 lock-screen survey responses" [2]. 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.

Our Contribution

Researches from different communities (IoT, DataScience, UbiComp, Sensors, Networks/Systems [2]) backed this project on Indiegogo [6]. We supported the *CrowdSignals* campaign by buying the “Guarantee Your Label" package – which introduces an additional ground truth label to be assessed via surveys. In agreement with the organizers, we decided to assess the label *interruptibility*, i.e. “the quality of being interruptible” [14]. 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 [8, 10]. 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.

Our Expectations

As mentioned before, we bought the "Guarantee Your Label" package to receive the full dataset plus labels indicating the user’s interruptibility. We wanted to use the dataset:

- To correlate interruptibility with other ground truth labels [2], mentioned as follows, and context-information inferred from smartphone sensors to specify conditions under which users feel more or less disturbed by notifications
  - Current place + transportation mode to get there
  - Mood and physical well-being
  - Phone position
  - Sedentary activity
- To save time and money by letting others conduct the user study to collect the data we need
- To support the idea of a community dataset that allows comparative and reproducible research

To allow the correlation analysis we intended to run, we required a large set of data labeled for interruptibility and literally at the same time further labels for well-being or place type, for example. This means that 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.

The Reality

Now, let us “zoom into” the dataset and highlight some issues we identified when we started to look into the dataset.

1. The dataset is very sparse for all labels (see Figure 1), but especially for interruptibility labels (see Figure 2). 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 was answered at the same time (see Table 1).
3. Due to the random selection of survey items, the share of data points per labels is unbalanced (see Figure 1). The problem intensifies as the chance of being selected as one of the two to four survey items decreased 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.

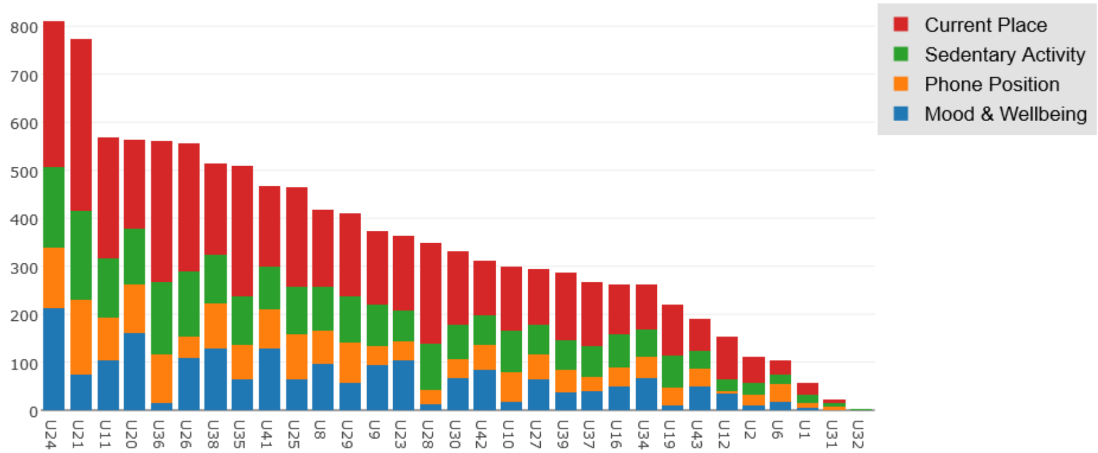

Figure 1: Overview of the number of survey responses for each label and for each user (U).

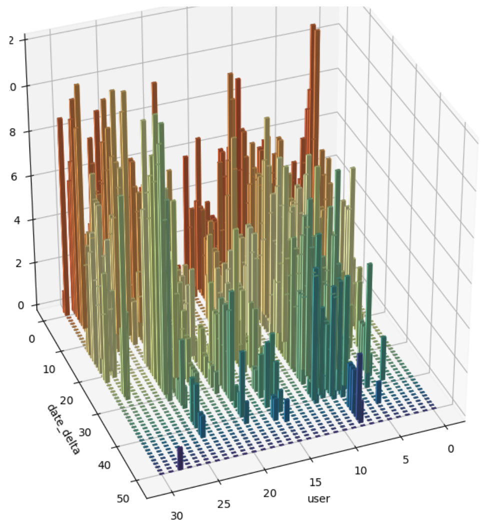

Figure 2: Overview of the number of interruptibility survey responses (z axis) per user (x axis) over time (y axis).

Survey item	# parallel interruptibility	Median time to next answered interruptibility
Current Place	17	63 minutes
Sedentary Activity	6	187 minutes
Phone Position	2	287 minutes
Mood & Wellbeing	2	119 minutes

Table 1: 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.

Feedback From Backers and Supporters

To have a broader impression of the dataset we contacted supporters and backers [5]. 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 did not yet have time to use the dataset. Two parties looked into the dataset, one of them already published their results [13].

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 expectations. We suggest to have rather multiple datasets, aligned with each application area than trying to find the jack of all trades dataset.

The backers we asked had different expectations. For some of them the dataset met their expectations:

> “We were looking for a set of diverse real data series, which is what Crowdsignals.io delivered.”

However, some of them identified issues with the dataset, especially concerning timestamps and ground truth labels:

> “[The dataset] came with shifted timestamps. We had to pre-process data to correct timestamps. The ground truth about activities is missing or incorrect. This makes the data useless for supervised machine learning.”

> “What we expected are related to data streams of these ubiquitous sensors (mobile phones, smart watch/wristband and other embedded sensors) which contain many environmental contexts that can be derived/labeled from the users (close to real time).”
>
> “[A] common problem that we nearly met is related to inconsistency in data labeling [...], which is inherently difficult for real-time activity recognition in our research.”
>
> “[...] a substantial process for data cleaning may be necessary for the purpose of our work.”

There are also researchers for which both applies:

> “Fortunately, it met our expectations in terms of the activity that the user is performing to a certain degree.”
>
> “However, we were expecting more environmental contexts (labels) or situation label (e.g. crowdedness).”

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

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 [7]. It is possible that different participants interpret the study task such as labeling activities differently which might lead to inconsistent data. It is also an open issue how to engage participants for a longer time period, avoiding a drop in data quality (cf. Figure 1 and 2). 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.

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 worse (i.e. sparse) the dataset gets. To avoid this, it might be worth having multiple smaller user studies which focus on one extra label each instead of having one super big user study trying to cover it all. This will lead to a larger sample size, but the extra money might can be used for that.

Second, despite a monetary incentivation, participants' the response rates were rather low. Probably, shorter surveys, i.e. less labels, that are faster to fill in would lead to a better user experience. In addition, gamifications methods might enhance the user motivation and keep the compliance high even after some weeks.

Third, 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 [9] 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

Suggested improvements specific for the case study are:

- Larger dataset (more than 30 participants and more than 30 days)
- More information on the environment / surrounding
- Correct timestamping, synchronized surveys
- Correct labeling of data streams

Summary

In this paper, we presented our experiences with the *CrowdSignals* dataset, contrasted our expectations with the reality and share some lessons learned.

The dataset has some flaws who are partly caused by the nature of in-field user studies and partly by over-engaged projects aims. The dataset is pretty sparse for labels 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.

Due to the mentioned issues, correlation analysis among ground truth labels is difficult. However, other analyses appear promising, especially if linked to smartphone sensor data which is very rich in this dataset.

One benefit of the *CrowdSignals* dataset of our case study is that the sample of participants is manifold, even though it would be nice to have more than 30 people for a longer period of time.

For future crowdfunded in-field data assessment studies we suggest to:

- Focus on one thing at a time: rather several smaller studies with fewer ground truth labels and application areas but with more participants overall
- Apply gamification mechanisms to keep participants motivated, e.g. include rewards for regular survey feedback
- Rely on objective data where possible, e.g. use automatic place labeling instead of user-provided labels

- Run a test survey phase to check if everything fits and if participants understand the instructions correctly
- Draw a dataset sample from the test phase which can be discussed with the backers and compared to their expectations
- Do not hesitate to take time to talk to each other and to make sure you are speaking “the same language”

Acknowledgements

This research was partially funded by the German Federal Ministry of Education and Research (BMBF) within the Smart Data Innovation Lab (grant no. 01IS15035A).

In addition, we thank all backers and supporters of the campaign who responded to our quick survey.

REFERENCES

1. AlgoSnap Inc. 2016a. AlgoSnap. `http://algosnap.com/`. (2016). accessed on June 30th, 2017.
2. AlgoSnap Inc. 2016b. AlgoSnap - CrowdSignals.io Pilot Dataset Reference. published as part of the final dataset; not yet available online. (2016).
3. AlgoSnap Inc. 2016c. CrowdSignals.io: A Massive New Mobile Data Collection Campaign. `http://crowdsignals.io/`. (2016). accessed on June 30th, 2017.
4. AlgoSnap Inc. 2016d. CrowdSignals.io: A Massive New Mobile Data Collection Campaign – Data. `http://crowdsignals.io/#data`. (2016). accessed on June 30th, 2017.
5. AlgoSnap Inc. 2016e. CrowdSignals.io: A Massive New Mobile Data Collection Campaign – Experts. `http://crowdsignals.io/#experts`. (2016). accessed on June 30th, 2017.
6. AlgoSnap Inc. 2016f. CrowdSignals.io: Building a Community Dataset | Indiegogo. `https://www.indiegogo.com/projects/crowdsignals-io-building-a-community-dataset%2Dandroid-smartphone#/`. (2016). accessed on June 30th, 2017.
7. 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. *ACM Journal on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 3 (2017). (to appear).
8. Anja Exler, Marcel Braith, Andrea Schankin, and Michael Beigl. 2016. Preliminary investigations about interruptibility of smartphone users at specific place types. In *2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Adjunct Proceedings)*. ACM, 1590–1595.
9. Google Developers. 2016. Place Types | Google Places API | Google Developers. `https://developers.google.com/places/supported_types?hl=en`. (2016). accessed on June 30th, 2017.
10. Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. In *2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 897–908.

11. UCI. 2016. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml/index.php`. (2016). accessed on June 30th, 2017.

12. Rajan Vaish, Keith Wyngarden, Jingshu Chen, Brandon Cheung, and Michael S Bernstein. 2014. Twitch crowdsourcing: crowd contributions in short bursts of time. In *32nd annual ACM conference on Human factors in computing systems*. ACM, 3645–3654.

13. 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).

14. Wiktionary. 2016. interruptibility - Wiktionary. `https://en.wiktionary.org/wiki/interruptibility`. (2016). accessed on June 30th, 2017.