LiveLabs: Building In-Situ Mobile Sensing & Behavioural Experimentation TestBeds

Kasthuri Jayarajah, Rajesh Krishna Balan, Meera Radhakrishnan, Archan Misra, and Youngki Lee
School of Information Systems, Singapore Management University
{kasthurij.2014,meeralakshm.2014}@phdis.smu.edu.sg, {rajesh,youngkilee,archanm}@smu.edu.sg

ABSTRACT
In this paper, we present LiveLabs, a first-of-its-kind testbed that is deployed across a university campus, convention centre, and resort island and collects real-time attributes such as location, group context etc., from hundreds of opt-in participants. These venues, data, and participants are then made available for running rich human-centric behavioural experiments that could test new mobile sensing infrastructure, applications, analytics, or more social-science type hypotheses that influence and then observe actual user behaviour. We share case studies of how researchers from around the world have and are using LiveLabs, and our experiences and lessons learned from building, maintaining, and expanding LiveLabs over the last three years.

Keywords
Mobile Sensing; Experimentation; Testbed

1. INTRODUCTION
A key desire for many mobile computing researchers is the opportunity to test their hypotheses, systems, and techniques on real users in real environments in realistic ways. However, making this happen is hard: in particular, you need to recruit participants, secure an environment where such participants will exhibit natural behavior, and finally, put in all the instrumentation necessary to accurately trigger your experiment as well as to collect the data needed to evaluate the effectiveness of the experiment. All of these steps are hard, painful, and sometimes impossible to do.

For example, recruiting participants can take many man-months of work to manage recruitment calls, screen the responses for suitability, engage with suitable candidates to collect demographics data, and to keep them interested and informed about the experiments, and finally, follow up with payment and other paperwork, where necessary. Next, securing a suitable test environment can also take many man-months as the venue owner needs to be convinced that running experiments in their venue is acceptable. For example, even at universities, this may involve long negotiations with the facilities management staff as well as the institutional review board. Finally, instrumenting the environment is strenuous as some types of useful experiment data, such as where the participants went or who they were with, is hard to capture without in-environment sensors. As such, experimenters often resort to instrumenting a mobile application and asking the participants to download that app and run it. However, this limits the types of data that can be collected and also the participant pool – as the app may only work on Android for example.

In this paper, we describe our multi-year effort to provide an in-situ real-time behavioural experimentation testbed, called LiveLabs, that can be used by researchers to run rich human-centric mobile sensing & behavioural trials in realistic settings. In particular, LiveLabs provides three testbeds at: our university (SMU), a convention centre (Suntec), and a resort island (Resort1). These provide real-time attributes about participants at each venue, such as location and group membership, and additionally, at SMU, a large and diverse fully opted-in and IRB-approved participant pool that can receive, as mobile notifications, and react to real-time in-situ experimental stimulus. LiveLabs can also be used at all 3 venues, SMU, Suntec, Resort, to test novel mobile sensing applications, infrastructure, analytics, and other research ideas and prototypes.

LiveLabs has been actively developed for the last 3 years\(^2\) by over 90 people, and leverages many deep technology components such as server-side location tracking [26, 60], dynamic group [50] and queue detection [42], energy efficient mobile data collection software [7], and a very usable and intuitive UI. In the rest of this paper, we briefly describe the main LiveLabs components and showcase its integrated use for various behavioural studies.

Building LiveLabs required this effort level as we had to develop, implement, harden, and maintain many different novel pieces of technologies that could work with minimal maintenance across Android, iOS, laptops, etc. We also had to secure the necessary funding, hire the right people, convince our venue partners to deploy LiveLabs, and most importantly, convince participants and researchers to join and use LiveLabs. This last point needs to be stressed as forming a stable test and user population was much harder than we thought. Indeed, to overcome these challenges, we had to change our lab structure – from a standard research lab into more of a startup with separate roles and personnel for research, production, maintenance, client management, administration, and participant recruitment. For example, technology developed by our researchers is never integrated into LiveLabs directly; instead it is handed over to the production team to re-implement correctly – i.e. hardened and made maintenance-friendly.

LiveLabs has been described previously [8, 35], however, this is the first presentation of the full end-to-end system along with

---

\(^1\)Venue intentionally anonymized

\(^2\)Through generous support by Singapore’s National Research Foundation http://www.nrf.gov.sg/
detailed usage results. Specifically, we focus on how LiveLabs allows researchers to run innovative in-situ real-time human-centric behavioural studies. We will describe the experiment types that LiveLabs supports and show how this improves on the state-of-the-art. We then present numerous case studies showing how LiveLabs has been successfully used for various experiments. The main contributions that we make in this paper are as follows:

- We present, to our best knowledge, a first-of-its-kind, available-for-use testbed specifically tailored towards in-situ, real-time mobile sensing & human-centric behavioural experiments.
- We describe how LiveLabs uses well-established experimentation methodologies that allow both social science and technology researchers to test new and innovative multi-disciplinary ideas in mobile sensing, real-time analytics, behavioural studies, marketing methods, and other related domains.
- We present the various LiveLabs building blocks, showing how deep technology components, such as location sensing, couple with a rich, yet intuitive, UI to provide a complete end-to-end experimentation platform.
- Finally, we present case studies showing how LiveLabs has been used to investigate various research hypotheses by many different groups of researchers. LiveLabs has been operational since January 2015 at SMU and since June 2015 at both Suntec and Resort. As of May 2016, ≈ 4,000 undergraduates have signed up as LiveLabs participants at SMU; at Resort, ≈ 27,000 members of the public have downloaded the resort’s publicly available (Android and iOS) smartphone application that incorporates our technology.

LiveLabs is available for use, for free, by any researcher.

2. DESIGNING LIVELABS

Designing and conducting unbiased behavioural experiments is a highly complicated process that involves recruiting the right participant pool to running minimally biased experiments. For minimum bias, subjects should be placed in natural environments where nothing is artificial – i.e., they are using their own devices and interacting in normal ways with their friends and others around them. In this setup, you then want to inject a stimulus and observe how the subject pool behaves (or how some system behaves if you are testing a technology component). It is this kind of natural test environment that LiveLabs was built to provide. A natural environment allows for running truly innovative experiments that go far beyond the constraints of a standard user study where the experiments are usually not realistic, e.g., “shopping” simulations done inside a lab, and the subjects are biased as they are aware that they are part of an experiment, e.g., the experimenter is present.

2.1 Design Goals

The two main design goals of LiveLabs were:

1. **Rewarding & Safe for Participants:** LiveLabs requires the active participation of either students or members of the public to be successful. Thus, it is important that being a part of the testbed is seen as both useful and safe to all the participants. We describe our privacy mechanisms in Section 2.2 and our attempts to make LiveLabs useful in Section 6.3.

2. **Useful for Technology & Behavioural Researchers:** The second major goal of LiveLabs is to make it useful for multiple types of experiments. For example, LiveLabs should support systems-type experiments that require installing and testing new sensors or forms of data collection, new sensing algorithms, or types of real-time analytics. In addition, LiveLabs should also support behavioural-type experiments where the participants are provided some kind of stimulus, via an in-app notification for example, followed by observation of their reaction to that stimulus.

To support both types of experiments, LiveLabs must thus be flexible to support multiple experiment types, easy to modify to add new functionality, and easy to use for experiments even by non computer-scientists. We describe in Section 3 the main components and implementation of LiveLabs that allows us to satisfy these design criteria.

2.2 Privacy Considerations

A key LiveLabs design goal is to achieve a useful tradeoff between maintaining user privacy while still supporting useful context-aware experiments. For example, LiveLabs could average the locations of all users in a large area and store just one location for all of them. However, experiments that require identifying the locations of specific people such as “Send a coupon only to people at Table 23” cannot be run if the averaging has removed the required fidelity.

Hence, we decided to anonymise all collected data without performing any kind of aggregation or reduction that preserves privacy at the expense of data fidelity. We created a unified framework (Section 3), where all data is processed by an anonymisation sub-component that 1-way hashes all personally identifiable information. The raw data is then removed and only anonymised data is retained. In addition, during participant sign up, we explicitly state how we collect and anonymise all data including demographics information and regular uploads of sensor and app usage data from their smartphones as well as location data from infrastructure sensors. The entire LiveLabs participant data and consent handling procedures have been approved by both the university’s Institutional Review Board (IRB), and legal department who deemed our procedures to be in compliance with Singapore’s Personal Data Protection Act data privacy regulations.

However, even with these procedures in place, it is possible for a malicious hacker, with access to all the LiveLabs databases, to identify specific individuals as we use a consistent hashing scheme. We are leaving, as future work for both ourselves and for any interested systems privacy researchers, the development of better privacy schemes that can protect against more attacks, and still allow LiveLabs to provide the level of context-driven personalized experimentation that it currently supports.

3. IMPLEMENTING LIVELABS

To achieve the goals in Section 2, LiveLabs needs to overcome the following technical challenges. First, sensor data has to be collected without imposing unacceptable privacy or energy concerns for the participants. Next, accurate and useful contextual triggers have to be provided from the data collected (e.g., accurate activity detection from inertial sensing data from participant smartphones). Then, a powerful yet easy to use interface to describe experiments that LiveLabs can run has to be designed. This is particularly important, and hard, as a large number of experimenters are not computer scientists and have limited experience and patience with complicated IT systems. Finally, LiveLabs needs to provide detailed experiment results that can be used by experimenters to understand and improve their experiments. This is surprisingly non-trivial as 3http://is.gd/livelabs to test a sandboxed version for yourself 4https://www.pdpc.gov.sg/legislation-and-guidelines/overview
running experiments in real-world spaces introduces errors, e.g., the computed location of a participant can be wrong, that need to be accounted for in the experiment reports. We describe how we solve these challenges in the rest of this section.

3.1 Architecture Overview

The LiveLabs architecture is shown in Figure 1. The process flow is as follows: (i) Participants sign up for LiveLabs and install the LiveLabs Service app on their personal iOS or Android phones. (ii) The service apps running on each phone along with the environmental sensors, Wi-Fi in particular, regularly upload raw data to the LiveLabs data collection service. The data is anonymised and then processed to generate real-time contextual information about the participants. (iii) In parallel, experimenters provide experiments using the LiveLabs experiment specification user interface. (iv) For each approved experiment, the LiveLabs experiment engine checks when the experiment should run based on the dates and times provided by the experimenter. When the experiment needs to run, the experiment engine contacts the analytics engine to retrieve a list of participants that match the specified contextual triggers. (v) The experiment engine then sends the treatment, as a notification, SMS, or email message to a specific LiveLabs app, to each matched participant. (vi) The experiment engine monitors what each participant did after a treatment was sent to them. (vii) The anonymised responses are sent to the experimenter for analysis and improvements. To achieve this entire logical flow, LiveLabs builds on multiple technical components, which we describe next.

<table>
<thead>
<tr>
<th>Content</th>
<th>Android</th>
<th>iOS</th>
<th>Information Recorded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility Event</td>
<td>Yes</td>
<td>No</td>
<td>Accessibility Event, App Name, Package Name, Event Class</td>
</tr>
<tr>
<td>Battery</td>
<td>Yes</td>
<td>Yes</td>
<td>Battery level</td>
</tr>
<tr>
<td>Calendar</td>
<td>Yes</td>
<td>Yes</td>
<td>Title, Notes, Location, Dates</td>
</tr>
<tr>
<td>Call Logs</td>
<td>Yes</td>
<td>Yes</td>
<td>Last Dialed Number</td>
</tr>
<tr>
<td>Contact</td>
<td>Yes</td>
<td>Yes</td>
<td>FirstName/ Middle/ LastName</td>
</tr>
<tr>
<td>Location</td>
<td>Yes</td>
<td>Yes</td>
<td>x, y Coordinates</td>
</tr>
<tr>
<td>Program State</td>
<td>Yes</td>
<td>Yes</td>
<td>State</td>
</tr>
<tr>
<td>System Settings</td>
<td>Yes</td>
<td>Yes</td>
<td>App Version, Language, Carrier, Phone Name, Signal Strength</td>
</tr>
<tr>
<td>System Log</td>
<td>Yes</td>
<td>Yes</td>
<td>ASLMessageID, Level, PID, Facility, Sender, Message</td>
</tr>
<tr>
<td>WiFi Trace</td>
<td>Yes</td>
<td>Yes</td>
<td>Action, BSSID, RSSI</td>
</tr>
<tr>
<td>Bluetooth</td>
<td>Yes</td>
<td>No</td>
<td>State</td>
</tr>
<tr>
<td>External Media</td>
<td>Yes</td>
<td>No</td>
<td>Media state</td>
</tr>
<tr>
<td>Installed Apps</td>
<td>Yes</td>
<td>No</td>
<td>Package Name</td>
</tr>
<tr>
<td>Network State</td>
<td>Yes</td>
<td>No</td>
<td>Data Source, Data Type</td>
</tr>
<tr>
<td>Profile State</td>
<td>Yes</td>
<td>No</td>
<td>Possible Actions</td>
</tr>
<tr>
<td>Cell Tower</td>
<td>Yes</td>
<td>No</td>
<td>Cell ID (CID), Location Area Identity (LAC), Primary Scrambling Code (PSC)</td>
</tr>
<tr>
<td>Sensor Data</td>
<td>Yes</td>
<td>No</td>
<td>Accelerometer, Gyroscope, Light, Pressure, Rotation</td>
</tr>
</tbody>
</table>

The collectible data changes with each OS version update, e.g. iOS 6 vs 7. This table lists the maximum data we can collect from each OS.

Table 1: Contextual Data Collected: Android vs. iOS

3.2 Collecting Data

Collecting Data From Participants: A key LiveLabs component is the LiveLabs Service app for both Android 4+ and iOS 6+. As stated earlier, we do not provide any devices to the participants – instead participants install the LiveLabs app on their primary smartphone. This decision was made to reduce the testbed bias as we strongly believed that you could only observe natural behaviour when participants were using their own devices and data plans.

The data collected by the LiveLabs Service app depends on the OS and the current data collection policy. For example, we are able to collect more data on Android versus iOS – See Table 1 for details. However, a key lesson learned early on was that participants were extremely sensitive to any extra smartphone power drain or heat generation caused by our data collection routines and that they would quickly uninstall the LiveLabs app, leave the pool, and also speak negatively about LiveLabs to their friends.

Hence, we extensively measured the power drain of various sensors both individually and when they were used together (key results are presented in Balan et al. [7]), and created energy profiles for various sensors and sampling frequencies. We then added policy support into the LiveLabs service app to allow the data collection policy to be dynamically updated, e.g. which sensors to collect data from, at what sampling rates, for how long, etc. All data collected is compressed, and stored as files on the participants phone with quotas determining how much data to store before overwriting. The data is uploaded to the LiveLabs Analytics Engine whenever the phone is using Wi-Fi to eliminate any cellular data charges. After uploading, the data is cleared from the participants phone.

To achieve the highest participant satisfaction, the default policy is to introduce no additional power drain by collecting no data at all. However, based on experiment requirements, we can and do turn on phone sensors, e.g. inertial sensors to detect current activity, for a few hours, at most, per day.

Collecting Data From The Environment: In addition to phone data, LiveLabs also collects data from the environment. In particular, we collect Wi-Fi signal strengths of every connected device directly from the Wi-Fi infrastructure.

Data Anonymiser: All data collected, from either phones or the environment, is processed by our anonymisation service before it enters any LiveLabs data store. This service uses a consistent 1-way hashing function to hash any personally identifiable information in the data, such as MAC addresses, email ids, user ids, names and numbers. In addition, any data that was collected unintentionally, e.g. data that is not required by any existing data store, is discarded. The raw data that enters the anonymisation service is discarded and only the hashed data is stored.

3.3 LiveLabs Context Services

Another key component of LiveLabs is the analytics engine that processes the data collected from smartphones and environments to generate the contextual information used to trigger experiments and to understand their effectiveness. This engine uses contextual data generated by numerous pieces of prior work built by the LiveLabs research team over the last few years. We briefly describe the key prior work below:

Location Service: A key LiveLabs contextual trigger is the current location of all participants. The challenge was to obtain these locations with both high accuracy and high recall, both indoors and outdoors. To achieve both goals, we decided to collect raw signal strength data directly from the commercial Cisco or Aruba Wi-Fi controllers used at each venue. This is because collecting data from phones consumes energy about which our participants are sensitive to. Also, as many devices are either not part of LiveLabs or run OSs that don’t allow client-side Wi-Fi sniffing (e.g., iOS), this results in data sparsity. In contrast, collecting data from the infrastructure allowed us to track every device that connects to the Wi-Fi infrastructure – even those not part of LiveLabs. Note: every device that connects to the Wi-Fi infrastructure at each venue has agreed to this tracking as part of the sign-on agreement. In addition, for devices that are not part of LiveLabs, the system tracks them as hashed entities with no additional knowledge about them.

We use fingerprint maps and a RADAR-inspired [5] solution to convert the raw data from the controllers into precise locations of
every device at each venue. In the process, we discovered unique challenges of doing this type of server-side localization and refined our system to overcome these challenges. Nairan et al. [60] describes our solutions to these challenges and provides detailed performance results for our system. Overall, our location system has been operational at SMU since August, 2013 and at both Santec and Resort since June 2015. It tracks, as of May 2016, ≈19,000 unique devices per day across all venues. At SMU, location results are updated every 1-2 minutes with a 6-8 meter accuracy.

**Group Detection Service:** We leveraged the fact that our location system tracks every device, whether they are LiveLabs participants or not, to build a dynamic group membership system called GruMon [50]. This system combines the location traces of every device detected in the environment with probabilistic analysis to identify devices that are moving together with intent – i.e., they are interacting with each other and thus in the same group. This is particularly hard in crowded places such as our university campus where tens or hundreds of people can be physically co-located in the same area, such as the campus food court, even though they are not interacting with each other. We have deployed GruMon at SMU since August 2014, and the other venues since June 2015, and provided the ability to select participants based on the size of the groups they are in.

Group membership is important as our recent results [22] showed that the behaviour of a group, in terms of how long they spend in places and where they are likely to go next, is very different when compared to the behaviour of individuals. In addition, different sized groups, e.g. small 2 person groups versus large 7 person groups, behave quite differently from each other as well. Hence, we believe, with corroboration from our psychology and marketing peers, that controlling for group membership is important when trying to understand the reactions of individuals to any kind of stimulus. As far as we know, LiveLabs is the only testbed that allows researchers to dynamically target different sized groups in an easy and non-invasive manner.

**Queue Detection Service:** In crowded places, such as SMU, queues can form dynamically at ATM machines, food stalls etc. To detect these dynamic queues, we built QueueVadis [42] that uses sensor data collected from participants’ phones to detect the wait and service times of queues. We then combine the data from multiple phones to create an overall view of the current wait and service times of all current queues across the campus. QueueVadis employs a two-tier classification approach for detecting individual-level queuing episodes using the locomotive signature of short bursts of shuffling forward between periods of standing. We showed, via numerous experiments in both Singapore and Japan, that it produces accurate values irrespective of queue shapes and even if participants join or leave the queues prematurely. We also showed that only a fraction of the people standing in a queue need to generate the data for it to work. Currently, we use QueueVadis only for special experiments as the need to actively collect data from participants’ phones prevent it from being an always-on service.

**Activity Recognition Service:** Finally, we have also developed a number of activity detection solutions that can detect what the participant is doing from the inertial sensor data collected from their phones and, more recently, smartwatches. For example, A3R [59] that detects standing, walking, sitting, and other basic activities as well as Annapurna [52, 51] which detects eating activities. Similar to QueueVadis above, we do not offer these activity detectors as a standard service, and use them only on a case-by-case basis, as they require active data collection from the participants’ phones.

### 3.4 LiveLabs Experiment Engine

The final component of LiveLabs is the experiment engine. This component receives experiments from experimenters through a GUI-driven entry process, validates, schedules, and selects the right participants for any given experiment, executes the experiment, and provides the experimenter with data that can be used to validate the efficacy of the experiment.

**Experiment Specification (UI):** Figure 2 shows part of the experimenter UI. This UI portion allows experimenters to select the contextual triggers, e.g., the precise location in the top left area and the group membership criteria in the top right, for their experiment. Overall, the UI was developed and refined over many months with various experimenters to be easy to use, yet powerful enough to specify all experiment details, such as the number of control versus treatment participants and the actual intervention sent when the experiment runs. The LiveLabs UI can be accessed at [http://is.gd/livelabs](http://is.gd/livelabs) and sample experiments can be created in this sandbox environment.

**Experiment Validation:** Currently, all new experiments entered through the non-sandbox UI portal must be approved by a LiveLabs
director, who will receive email reminders, before the experiment can be executed. This is to ensure that every experiment is non-damaging to our participants, e.g., the experiment does not spam participants with junk messages or try to steal their personal data, as a bad user experience will affect our participant retention rate.

**Participant Selection:** Once an experiment has been approved, it gets scheduled according to the date and time selected by the experimenter. When the scheduled date and time occur, the participant selection component will query the analytics engine to ask for a list of participants who currently match the contextual triggers specified by the experiment. The analytics engine continuously queries its up-to-date databases and returns a list of all participants that match the triggers. The selection component then partitions this list into control and treatment groups based on the experiment specifications. If the size of any group is smaller than the minimum required size, the experiment is delayed until the appropriate minimum sizes are reached. If the sizes are sufficient, the selection component sends the list of treatment participants to the notification service along with the content to send as specified by the experiment. The experiment continues until either enough participants, as specified by the maximum required size criteria, have received the treatment or the experiment time window has expired.

**LiveLabs Notification Service:** This component uses existing push notification services provided by Google (GCM [18]) and Apple (APNS [4]) to send the experiment content to the selected participants’ phones. On the participants phone, the messages are routed either to the LiveLabs service app which then displays them as a notification or to another LiveLabs app as shown in Figure 3. The last time each LiveLabs app on a participants phone connected to a LiveLabs service is recorded in specific databases and used to check if specific participants can receive rich messages.

**Experiment Logger:** The experiment logger activates every time an experiment is executed. It retrieves and stores the location and context updates by every participant in that experiment, for both treatment and control groups, for a configurable period.

**Collecting User Feedback:** When an experiment is sent to a participant, the experiment logger could decide to turn on specific data collectors, e.g., collect inertial data to generate activity information, on that participants phone. If so, an updated policy is sent to the phone. In addition, the experimenter can also specify a post-experiment survey, to collect qualitative feedback, that will be sent to each participant, after the experiment concludes, as a link in a notification message.

**Context Uncertainty Handler:** A hidden LiveLabs bias is that the trigger events are derived from noisy sensor data. For example, the location system has an average 6 to 8 meter error with all the other analytics services having their own specific errors. Thus the participants chosen by LiveLabs can include people who do not satisfy the experiment criteria. Quantifying this error, which can never be eliminated as every sensing system will have errors, is an active area of research. Currently, we developed models that combine the errors generated by the low level sensors into probabilities that can be used by experimenters to understand and improve the accuracy of their experiments. For example, we use the notion of overlap regions to determine the probability that a chosen participant pool was actually located where the system thought they were [37]. In addition, we use models to suggest whether the minimum participant numbers need to be increased to reduce the uncertainty caused by the underlying sensing systems [36]. For example, choosing only 2 participants will lead to a large probabilistic selection error if room-level location accuracy is required.

**Report Generator:** Finally, the experiment engine will generate a detailed post-experiment spreadsheet containing information such as the timestamp and location where every participant received the experiment, along with the timestamp and location where they saw any notification, clicked any link, opened any LiveLabs app, and activated any LiveLabs app features. Example reports can be seen at the LiveLabs sandbox portal at http://is.gd/livelabs under “Experiment List”.

### 3.5 Current LiveLabs Testbeds & Status

Table 2 shows an overview of the three testbeds where LiveLabs is deployed. These testbeds are **SMU** – our entire university campus, **Resort** – a popular outdoor resort island in Singapore, and **Suntec** – a large and popular convention centre. At all three venues, we have deployed our infrastructure sensors, that track device movements using Wi-Fi, along with the real-time analytics solutions. At both SMU and Resort, we also engage with participants who install our mobile apps, built by LiveLabs for SMU and by enhancing the existing resort-specific app for Resort. We then provide researchers and marketing personnel with the portals and ability to run behavioural trials on these participants. Currently, Suntec does not have any participants; however that will change when the conference app (Section 4.1.2) is deployed for MobiSys 2016 and future conferences.

Most of the LiveLabs back-end systems are built using PHP with some components built using Java and JSP. PostgreSQL is used for most databases along with MongoDB to store unprocessed location data. We also built and maintain two campus-specific apps to help participant retention – an app to provide promotional content, and another to provide event information. We are currently building a third app that will be used during MobiSys 2016.
LiveLabs is maintained by a team of full-time professional developers who ensure that all core systems and apps are always working and available. In addition, we have a separate team of outreach and administration personnel who manage, maintain, and increase the participant pool. Separately, our research staff, Ph.D. students, postdocs, and faculty, design, prototype, and test the next set of LiveLabs improvements such as new sensing modules and better uncertainty handling. Once these new features are stable and deemed of value to experimenters, they are handed over to the professional developers to integrate into the main system as a core feature. Overall, about 90 people, including students, faculty, developers, post-docs, interns, admin staff, have helped to build LiveLabs.

3.6 What Can LiveLabs Be Used For?

LiveLabs can be used to test the following: (1) Mobile Sensing—hardware or software sensors that can be easily installed/deployed and used to monitor some aspect of human behaviour, (2) Mobile Analytics—algorithms, policies, and analytics that work on human-generated location and contextual data, (3) Mobile Applications—any Android or iOS application that does not require root access, and finally, (4) Behavioural Trials—experiments that interact with humans in real environments in real-time using contextual triggers and phone-based notifications. We provide examples of the system experiments in Section 4 and the behavioural experiments in Section 5.

3.7 Limitations of LiveLabs

LiveLabs was designed and optimised for experiments that can be triggered on a participant’s smartphone – either through an app that the participant willingly downloads or an SMS. This works well for experiments that require the participant to see a coupon or a message, run an app, or answer a survey etc.

A limitation of LiveLabs is that it does not, yet, integrate with social media data such as Facebook or Twitter. We are investigating ways to integrate social media feeds as the context they describe does not necessarily match the physical environment that LiveLabs operates in. Related to this, LiveLabs cannot be used to trigger experiments that require data that either LiveLabs does not collect (e.g., mental well-being status of a user), or cannot infer accurately (e.g., centimeter-level locations).

Another limitation is that the LiveLabs contextual data feeds may be limited by practical considerations such as participants uploading minimal data or by the accuracy and latency of the Wi-Fi data collected from the commercial Wi-Fi controllers. Also, experiments that require rich media, such as promotion coupons with images, can only be sent to participants who have voluntarily installed the appropriate LiveLabs app that can receive the rich content. By default, the LiveLabs service app can only receive plain content.

Finally, LiveLabs does not facilitate responsive surveying where experimenters immediately react and adapt the experiment. In addition, unlike lab settings, LiveLabs does not isolate participants – their response times are thus not guaranteed and may be influenced by non-experiment real-world stimulus.

4. RUNNING EXPERIMENTS

We now describe how LiveLabs has enabled research in Systems and the Social Sciences.

4.1 Using LiveLabs For Systems Research

A key use case for LiveLabs is to help test new technology. In fact, the context services that we described in Section 3.3 were all incubated as part of LiveLabs. We are actively improving LiveLabs with extensions to our environmental sensing by adding BLE-based fine-grained localization[46], context sensing capabilities with occupancy detection[39], emotion sensing [21], GPU-based vision sensing [20], privacy violations detection [25], improved power management for OLED displays [55, 54], support for wearable devices that can be used to detect eating [51], shopping [45] etc. Also, the ability to perform additional experiment types with support for social network-based experiments [23], trajectory-based recommendation systems [24], and audio log-based experiments. Finally, we have hosted interns from PhoneLab [38], KAIST, Wisconsin, and PEC, and have started the tech transfer process to setup a similar testbed at UMASS Amherst. We welcome and offer other researchers the chance to test their technologies in our campus. We now present two upcoming systems experiments that could only be done on a testbed like LiveLabs.

4.1.1 Real-Time Contextually-Aware Public Displays

One of our venues, Suntec, uses large public displays, as shown in Figure 4, throughout their premises. As part of a joint study with professors and a Ph.D. student from Lancaster University, we plan to investigate how to use these displays to engage more effectively with customers using the premises. One key idea is to use contextual information collected from our real-time sensors to automatically display relevant information to customers passing by the displays; for example showing the location of a meeting to a business user or providing information about shopping promotions to a customer heading towards the mall. In addition, we plan to add interactive sensors, such as the Kinect [33], to some screens to allow customers to interact with the screen in useful ways. Finally, we plan to identify techniques to partition the screen space of a really large screen, such as the one shown in Figure 4, between multiple users located in the same proximity.

4.1.2 Gamification of Conference Programmes

As part of the MobiSys 2016 conference [1], which will be held at Suntec, LiveLabs, together with researchers from Microsoft Re-
search Redmond, are exploring the use of a mobile iOS and Android app to improve the interactive quality of the conference through gamification and other social networking features. The key features of the app include smart friend finding, using the LiveLabs location service, contextually-triggered games, notifications, and other experiments using a modified version of the LiveLabs experiment engine, continuous engagement through creative games, points, leaderboards, and badges using technology built for the SMU apps, and social interaction during talks and other events using a technology called Microsoft Research Embedded Social Software Development Kit Service [34] from MSR Redmond. Figure 5 shows how the current app prototype 3 looks like; the reception from our initial testers has been very positive. We will report on the success of the app at MobiSys in a future article and we hope to use this conference app at other academic conferences, such as MOBICOM and SIGCOMM, as well as trade shows and consumer events run at Suntec.

4.2 In-Situ Behavioural Studies

Creating a natural experiment environment for running behavioural studies, without cutting corners or introducing additional bias, is extremely challenging. In this section, we review how conventional experiments are usually performed and then we describe how LiveLabs improves on various aspects of the experimentation process. We also discuss how we reduce the experimentation bias as much as possible. In Section 5, we provide case studies of experiments that we have run with colleagues from around the world cutting across various disciplines such as marketing, psychology and computer science.

4.2.1 Existing Experimentation Process

Currently, running experiments requires the following common steps. First, experimenters design the experiment. In this stage, they decide the scale of the study, the target participant demographics and recruitment locations, experiment methodology, e.g. survey, focus group, shadowing etc., prepare the survey instruments and related material depending on the chosen methodology, set aside the necessary budget for incentives, prepare and clear IRB approvals, and perform any other necessary preliminary steps. Second, experimenters start recruiting participants. During this process, the experimenters usually contact every interested participant to explain the purpose and process of the experiments, and to collect any necessary demographics and personal details for controlling or analysing the study results, processing payments etc. Third, the experiments are conducted with the twin goals of achieving reliable/meaningful responses while reducing as much bias as possible. However, achieving both goals usually requires an experienced experimenter – indeed, many studies are re-done as either valid responses were not collected or unintended biases were introduced into the study. Finally, experimenters analyse the responses and other data to draw useful conclusions. Usually, their responses will be a mix of quantitative, e.g., precise response times, and qualitative, e.g., free form questions, data.

4.2.2 How LiveLabs Improves The Process

We now describe how LiveLabs improves on the conventional experimentation process.

Step 1: Experiment Design. The key LiveLabs innovation is allowing experimenters to easily use contextual inputs as experiment triggers. For example, in conventional studies, the control and treatment groups are usually manually determined and pre-assigned by the experimenters. With LiveLabs, treatment groups can be dynamically assigned, at no additional cost, using various contexts.

In addition, LiveLabs allows experiments that are almost impossible to do, without excessive bias, in a lab setting. For example, LiveLabs can support experiments such as “Send a survey to the participants only when they are with 3 or more people”. We list the full set of LiveLabs experiment conditions in Section 4.3.

Based on the experimenter requirements, LiveLabs currently supports the following experiment designs: Simple Control and Treatment: a single treatment along with the size of the treatment and control groups, Multiple Treatment: multiple treatments that are sent concurrently to distinct participant pools with the same contextual triggers), and Chained Conditions: a series of treatments that are sent to the same set of selected participants in sequence. In all cases, LiveLabs will select the control and treatment pools using the specified contextual triggers. The experiment continues to run until the minimum pool size is met or the experiment expires.

Step 2: Participant Recruitment. LiveLabs greatly simplifies this process for an experimenter by providing a large, demographically diverse pool of opt-in participants belonging to multiple majors, gender, etc... If the experiment requires data or permissions beyond what the participants have agreed to, LiveLabs will send a new agreement. In many cases, the participants have already provided sufficient consent as the experiments usually involve sending a notification to the phone or collecting raw sensor data.

LiveLabs uses a team of dedicated staff to actively recruit and retain participants. To sign up, participants must provide explicit consent, basic demographics information, and then install the main LiveLabs service app – this serves as both an app store for participants to obtain other LiveLabs apps as well as a conduit that allows LiveLabs to collect data from the participants’ smartphone and to send treatments to the participants. The participants are incentivised to sign up through bonuses, periodic promotions and, most importantly, compelling apps, e.g., a campus events app, that they could not obtain otherwise. In addition, we make it clear that their data is kept private and that this is benefiting research. We discuss the recruitment process in more detail, along with our challenges, in Section 6.3. An advantage of this centralised approach is that our large pool of participants don’t actually know what experiments will be sent to them or why. Thus, every participant’s reaction to notifications, survey requests, or other experimental content is quite natural.

Step 3: Conducting Experiments. After an experimenter submits an experiment, it is validated by senior LiveLabs staff. An experimenter is allowed to run experiments that send a simple notification (e.g., a link to a survey, an SMS, or a rich message like

3Videos available at http://is.gd/mobisys16app
sending a visually appealing campus store discount coupon to a LiveLabs promotions app) to a participant’s smartphone. Currently, we have developed rich apps that can handle promotions, events, and micro-tasks. The LiveLabs experiment engine (Section 3) will execute the experiment at the correct time, with a participant set that matches the experiment’s contextual triggers, and then send the experiment content to those participants.

Note: the above is the standard setup for conducting social-science type experiments where a stimulus is sent to the participant and the reactions observed. LiveLabs can also be used to test the efficacy of various sensing and/or other technologies. In this case, the technology will be deployed to participants, via an app, and LiveLabs will collect both the necessary data from the app and appropriate ground truth, using environmental sensors etc., to allow the technology to be evaluated.

Step 4: Results Analysis. A key benefit of LiveLabs is that it is able to collect rigorous post-intervention data through data collected from the participants’ phones and environmental sensing. As such, we can provide experimenters with detailed reports on the efficacy of their experiments. For example, we can tell them when and where a particular promotion was received by a participant, where the participant actually viewed the promotion, and where they actually clicked the redeem button, along with the time taken to transition between each phase. This type of precise data is very hard to obtain, without bias, in a lab setting, and almost impossible in a real-world environment.

Note: we never reveal personally identifiable information to any experimenter or allow them to interact with participants directly. For example, experimenters only receive details such as “Participant 1110010 (Gender: Male) from Economics Year 3 saw the promotion at 11.33 a.m. at Level 3, Building A”. This level of detail, after discussions with various experimenters, was deemed suitable as it provides sufficient details while hiding the identities of participants from non-malicious experimenters – we use strict admission control to detect and remove malicious experimenters. If an experimenter requires personally identifiable information, additional consent from each participant will be required.

Note: the replicate entries were derived from actual prior work [14, 16]. Extend and innovate were influenced / designed by our social science colleagues – based on the state of research in their respective fields.

Table 3: Experiment Use Cases and Context Triggers

<table>
<thead>
<tr>
<th>Context Triggers</th>
<th>Replicate</th>
<th>Extend</th>
<th>Innovate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Static Contexts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Participant Pool</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Dynamic Contexts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building-level</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Floor-level</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Room-level</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Location Transitions</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Time</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Group Status</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Queue Status</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Activities (e.g., walking)</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Interests (e.g., web history)</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

5. BEHAVIOURAL EXP. CASE STUDIES

In this section, we show how LiveLabs allows experimenters to replicate traditional experiments faster and with less effort, extend existing experiments in interesting ways, and finally, and most importantly, to innovate and create completely new and previously impossible-to-do high research-value experiments.

5.1 Replicating Existing Experiments

We replicated two classic psychology experiments with the help of a SMU psychology faculty and a Temple university marketing faculty. As stated in Table 3, Replicate type experiments do not depend on real-time context triggers, but they help establish the representativeness and sufficiency of the size of our participant pool.

Bystander Apathy [29]: We replicated a classic phenomena, where the presence of others inhibits helping behaviour, described in Garcia et al. [16]. Our results are completely consistent with both the original work as well as a follow up study [15] that showed that bystanders could either increase or decrease the helping behaviour depending on where the bystander’s attention was focused. To replicate this study, we sent a survey, similar to the original study [16], that asked participants what percentage of their annual earnings, after graduation, would they be willing to donate to charity along with questions regarding their group context. The survey questions and responses are available at http://tinyurl.com/bystandersurvey.

Foot-in-the-Door Technique [14]: We demonstrate through LiveLabs, that it is possible to replicate this type of classical study [14] which examines a highly influential, compliance technique where an individual agrees to a large request only because they were first asked and agreed to a modest request. To do this, we created a control group which was asked for a large favor, a donation amount to create literacy programs for blue collar workers in Singapore, and a treatment group that was asked of a moderate favor, to sign a petition asking for more literacy programs for workers in Singapore, followed by the same large favor. The survey questions and responses are available at http://tinyurl.com/fitdsurvey.

5.2 Extending Experiments In New Ways

Many social science experiments involve priming and administering surveys to score participants along some behavioural or attitudinal trait such as psychometric scales etc. We show how LiveLabs has extended such techniques in new ways.

Priming Effects: This study, being run by an ESSEC business school marketing faculty, tries to understand how different types
of text, either deliberative or implemental [17], would subliminally affect where a participant would go next. We do this by sending a specific passage, written in either deliberative or implemental tone, to participants at specific locations on campus and then observing where they stayed after being subjected to the treatment. LiveLabs extends this experiment by allowing precise location triggers, e.g. “send only when they enter the library”, with non-invasive and accurate server-side monitoring of where the participants went to, e.g., did they pick a central or peripheral location in the library, after receiving and interacting with the text passage.

Promotion Framing: Another study run by the ESSEC faculty investigates how the framing of a promotion, e.g. offering an instant promotion versus one that is available after a delay, and the group context of a person impacts the redemption rate. Specifically, they hypothesize that the people who are alone are more likely to redeem the delayed promotion due to the effect of cognitive fluency. Further, the study also investigates whether the presence of strong or weak group ties when the participant receives the promotion is also a moderating factor in redeeming the offer. LiveLabs’s Group Detection Service enables this experiment by providing continuous monitoring of the participants’ group context, as well as quantified measures of tie strength through historical analysis of the participants’ co-location with others [23].

Opinion Leadership Study: A research team from Carnegie Mellon, Temple, and Emory universities is using LiveLabs to understand the role of leaders in groups – with the hypothesis that sending promotional content to the group leader, compared to other group members, will lead to better redemption rates. LiveLabs extended this study in the following ways: (1) the location traces collected by LiveLabs Location Service were used to generate models of who the leaders or followers for any group might be – validated by an opinion leadership scale survey, and (2) LiveLabs’s Group Detection Service allows those pre-identified leaders to be interacted with only when they are in groups. In particular, first part could have been done previously, but the second part is a new capability. Previously, they would have relied on self-reports of the identified group leaders to discover what they might have done or deployed shadowers to follow the leaders. These are now unnecessary as LiveLabs enables scalable, unobtrusive, and real time observations – allowing experimenters to use, with minimal effort, specific triggering conditions and to obtain accurate and non-invasive feedback on where and what the participants did.

Group-Aware Marketing: LiveLabs is being used by marketing faculty from Arizona State University to test group-aware marketing strategies where participants would receive group-aware, or generic, promotional content depending on whether they were in a group or alone. Their hypothesis is that sending group-aware promotions to people in a group results in higher redemption rates compared to generic promotions. This type of experiment is only possible, with minimal bias, using LiveLabs due to its dynamic location and group detection capabilities.

5.3 Innovating Experiment Designs

We now describe high-value experiments that could only be executed on LiveLabs.

Tracking The Spread Of Malware: One of our SMU computer science colleagues is using LiveLabs to track the spread of malware. In the first phase, they used LiveLabs to recruit participants for the study, via a survey. In the second phase, they plan to use LiveLabs Location and Group Detection Services to understand the physical context of participants at the time they receive “deceptive” messages that ask them to click on unknown web URLs – clicking on these URLs will trigger a malware that sends messages to friends in their phone contact lists. The new insights that LiveLabs provides is a deep understanding of the physical context of the participants at the time when they receive the malware message, clicked on or discarded the message, and their reaction when they learned that they had been deceived. Note: the experimenters will clearly explain to every participant about the deception and what they should have done as part of the debriefing process. The study was approved by the full IRB board. In addition, the experimenters had to convince the LiveLabs directors that the study was not going to harm the participants.

Personality Traits Study: Most personality studies in Psychology rely on self-reported data. However, self-reported data confounds situation selection with situation perception. Situation selection refers to the association between personality traits and objective situational attributes such as the number of people you are interacting with etc. Situation perception refers to the association between personality traits with subjective situational features such as their reaction to a red banner etc. Both variables are needed to understand how people experience a situation. However, self-reports make it impossible to distinguish selection from perception.

A SMU social psychology faculty is using LiveLabs to send surveys (≈ 400 surveys have been sent as of publication time) to known introverts and extroverts, identified a-priori through in-lab interviews, at various times and locations. The survey data, which provides the situation perception, is then combined with the LiveLabs Group Detection Service to accurately identify the situation selection, e.g. who is around them and who they interacted with, when they answered the survey. Thus LiveLabs provides a previously unrealisable ability to capture both the situation selection and perception simultaneously in a real world setting.

5.4 Success Metrics

In this section, we evaluate LiveLabs in terms of how it compares with traditional methods such as pencil-and-paper, web (e.g via Mechanical Turk), mail, and SMS surveys. We focus on two key survey success metrics – response rates and non-response bias [49]. In addition, we investigate the ability of LiveLabs to replicate results from conventional modes of experimentation.

Response Rate: In Table 4, we compare response rates previously reported for other modes of delivery against LiveLabs. We note that LiveLabs achieves comparable response rates, in addition to offering benefits such as real-time context-awareness. Further, we also note that the response rate is sensitive to the time period that
the surveys/promotions were sent out; we observe that the overall response rate was consistent at around 20-30% prior to the exam week, and dropped to 12-17% during the exam week.

Table 5 shows the response rate for each of the case study experiments – all the rates are comparable or higher than other approaches. In addition, participants respond to LiveLabs with more frequency than other phone notifications. To determine this, we analysed the phone data collected from 118 LiveLabs Android users, between Aug 2014 and Jan 2015, and determined that those users clicked on only 14.29% on average (sd = 17.7%) of the notifications they received whereas the average view rate observed with LiveLabs was much higher at 24.38%.

Figure 6 plots the CDF of the response rates for all those users. It shows that over 87% of the users responded to less than 24.38% of their notifications.

In Table 6, we summarize the response rates of experiments/trials run as part of the LiveLabs deployment at Resort. The testbed has been used to run many promotional campaigns, so far, across a wide demography comprising of all island visitors with the Resort-specific app installed. Interestingly, compared to SMU, the viewing rate is much higher at a lower redemption rate. Overall, LiveLabs achieves a better redemption rate (≈4.3%) compared to the SMS-based mobile coupon redemption study (3.2%) reported in Andrews et al. [3], for an Asian population similar to ours.

<table>
<thead>
<tr>
<th>Campaign</th>
<th>Subjects</th>
<th>Viewed Promo</th>
<th>Redeemed Promo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Bird Camps</td>
<td>3534</td>
<td>2457 (69.5%)</td>
<td>151 (6.15%)</td>
</tr>
<tr>
<td>Normal Camps</td>
<td>1265</td>
<td>863 (68.2%)</td>
<td>58 (6.72%)</td>
</tr>
</tbody>
</table>

Table 6: Summary Of Campaigns at Resort

Non-Response Bias: This is a critical survey flaw [49] where survey respondents are statistically different from the non-respondents in terms of demographics or attitudinal variables. To detect this bias, we separated the response and non-response rates by gender and school. We observed, similar to prior studies [49], that the women response rates were higher than men. However, and more importantly, we did not observe any significant difference in the demographics of respondents and non-respondents in terms of their school or year of study. Table 7 shows the response rate for all our studies broken down by our six schools. The percentages were not statistically different (using the Kolmogorov-Smirnov test), with $D = 0.33$ in all cases with $p$ values between 0.893 and 0.931.

Efficacy of Replicating Results: For the Bystander Apathy Study, our results, shown in Table 8, reinforce the original [29] and follow-up studies [15] in that individuals were less likely to help, donate in this case, when surrounded by people. However, if those people were not strangers, i.e., they were part of the same group, they were more likely to help.

Likewise, for the Foot-in-the-Door Study, similar to the original study [14], we observed a statistically significant increase in willingness to donate in treatment group individuals who agreed to the moderate favor, compared to those that did not. Table 9 presents these results with $t$ statistic = 2.8945, and $p$ value = 0.0111 for the difference in means test.

We do not provide results for the Extend or Innovate experiments as they are still ongoing and yet to be published. In addition, no baselines for comparison exist due to their originality.

6. KEY LESSONS LEARNED

We summarize here some of our main lessons learned as part of building and maintaining LiveLabs.

6.1 Deploying a Testbed Is Hard!

We initially envisioned [8, 35], deploying four separate testbeds, at SMU, Resort, an Airport, and a Mall, by the first half of 2015. However, as of May 2016, LiveLabs has only been deployed at SMU, Resort, and partially at a replacement venue Santee. A key reason was our gross underestimation of the effort, especially to agree on legal terms and generate convincing use cases, to deploy at commercial venues. Even at SMU, we learned that deploying at such unprecedented scales, e.g., a user pool in the thousands, is hard as supporting this large pool also meant supporting any software releases for month across heterogeneous devices and OSs. To address these challenges, we transformed from a standard research lab to a multi-faceted startup-like environment with dedicated staff and students handling research, software engineering, business development, participant recruitment, and administration separately.

6.2 Should Academics Be Doing This?

LiveLabs is supported by a large government research grant that allows us to hire the large full-time staff needed to build and support LiveLabs, and to offer it for free to other researchers. However, is this model sustainable, replicable elsewhere, and something academics should even try to do? Regarding sustainability, even though LiveLabs is currently fully funded, moving forward, we plan to offer membership tiers for organisations that run multiple experiments as well as charge a nominal per-experiment fee. These revenue streams will help pay for the full-time staff and any incentives needed to recruit participants.

The question of replicability and academic interest are inter-related as the interest drives the replication and vice versa to some extent. Over the last few years, numerous large multi-site testbeds such as PlanetLab [12], EmuLab [13] and GENI [10] have been proposed and successfully built. All of these testbeds were for core networking research use as that was of high interest to the research community. In all these cases, the testbeds were supported by both government grants, through the United States National Science Foundation (NSF), and by companies such as Intel for PlanetLab, Cisco for EmuLab etc. Hence, there are numerous examples of initiatives obtaining the resources needed to scale smaller testbeds into larger multi-site multi-country initiatives.

Our view is that academics should definitely strive to build compelling testbeds for various domains, such as mobile sensing, human interaction, energy solutions etc., as they will have a very large positive impact on the greater research community. However, creating testbeds is only partially about creating new technology solutions. Most of the time will be spent elsewhere – obtaining funding, hiring the right people, managing administrative requirements, recruiting, and retaining clients, experimenters, and participants, and building hardened solutions or prototypes for clients, participants etc. All of these additional tasks make building the testbed similar to running a startup — indeed we learned that we needed to convert our research lab effectively into a startup. Hence, building a testbed
Table 7: School-wise % of Respondents vs Non-respondents (Acc = Accountancy, Bus = Business, So.Sc = Social Sciences)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Garcia et al. [16]</th>
<th>People around in past 1 hr</th>
<th>With friends in past 1 hr</th>
<th>With friends currently</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 person, neutral controls</td>
<td>4.2 (1.6)</td>
<td>2.67 (1.03)</td>
<td>2.0 (0.63)</td>
<td>2.28 (1.01)</td>
</tr>
<tr>
<td>Group of 10</td>
<td>3.9 (1.5)</td>
<td>2.43 (1.01)</td>
<td>2.42 (1.00)</td>
<td>2.43 (1.03)</td>
</tr>
<tr>
<td>Group of 30</td>
<td>3.6 (1.3)</td>
<td>2.36 (1.08)</td>
<td>3.25 (1.71)</td>
<td>2.55 (1.21)</td>
</tr>
<tr>
<td>Trend</td>
<td></td>
<td>Inhibiting</td>
<td>Increasing</td>
<td>Increasing</td>
</tr>
</tbody>
</table>

Table 8: Comparison of Mean Charity Contributions of Garcia et al. [16] against LiveLabs Results

<table>
<thead>
<tr>
<th>Group</th>
<th>Amount Willing to Donate, mean (std.dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>3.655 (3.265)</td>
</tr>
<tr>
<td>Treatment (Yes to modest request)</td>
<td>4.276 (3.011)</td>
</tr>
<tr>
<td>Treatment (No to modest request)</td>
<td>1.5 (2.204)</td>
</tr>
</tbody>
</table>

6.3 The Challenge of Participant Retention

Maintaining a stable participant pool is our biggest challenge as using the standard approach of regular monetary incentives fails when the absolute dollar amount to keep incentivising 4,000+ participants is very large, and participants start dropping out rapidly when the incentives stop. We thus started providing unique functionality that only LiveLabs participants could enjoy – in this case, two campus-specific apps that we build and maintain for campus promotions and events tracking respectively. These apps have little research value but are key LiveLabs components that require a dedicated development team to provide support and regular new features to keep participant interest high.

We managed to build a reasonably large & stable participant pool, e.g. > 4,000 sign ups with ≈ 700 users actively using LiveLabs and running experiments, using these strategies: (1) provide a small monetary incentive ($10) to new participants, (2) run regular contests (e.g., lucky draws), during lull periods, to keep interest levels up, (3) refine and promote our exclusive applications that solve specific and compelling student needs, and (4) run experiments that offer interesting promotions/incentives to participants.

6.4 What Did We Do Wrong?

In this paper, we present, mostly, what we did right based on the final design. However, there were quite a few things that we did wrong – some quite obvious in hindsight. These include:

Believing that Phone Data Is Key: We initially assumed that most of our data would come from participants’ phones. However, we quickly learned that energy consumption was the dominant factor for participants staying in LiveLabs. In addition, the large variety of phones such as iOS, multiple versions of Android, meant that our data collection was not consistent even among our participant pool. Eventually, we re-focused on collecting most of our data from infrastructure sensors instead, e.g., Wi-Fi, BLE, interaction points such as displays. Indeed, all the behavioural experiments described in Section 5 use only data collected from environmental sensors as that was the only data sensor that was common across all participants. Currently, due to energy concerns, our current default setup is to turn off all data collection from phones unless needed – very different from how we initially envisioned LiveLabs operating.

Assuming that Others Would Use Our Cool Mobile Sensing Solutions: We also believed our colleagues would use our novel mobile sensing solutions, such as group and activity detection as experiment triggers. However, our non-computer-science colleagues did not know how to include these dynamic triggers into their current experiment methodologies. As such neither queueing nor activity detection was used by any of our case studies.

Our current approach is to provide familiar contextual triggers, such as location, time, demographics, interests etc., and slowly introduce new concepts, such as groups, queueing status, current activity etc., to them. We believe that our colleagues, with more LiveLabs experience, will integrate our novel new inputs to create new and innovative experiment designs. In addition, they will also tell us what new inputs they need to drive their research forward. This is already bearing fruits as one group of social researchers from Emory, Temple, and CMU, is already using our group detector, and suggesting improvements to it, to perform novel leader-follower type experiments (Section 5).

Hiring the Right Person is Crucial!: Finally, building a testbed is like running a startup where you need to hire talented people who are also innovative, adaptable, and flexible as their job scope / tasks can and will change often – sometimes daily. For example, our engineering staff frequently help out as advocates at participant recruitment fairs, and our admin staff also help to test engineering products. We learned to avoid hiring solely to fill specific work requirements as those requirements can and do change rapidly.

6.5 Future Work

In this section, we present our future work plan to address concerns raised by experimenters and to improve the overall testbed.

6.5.1 Experiment Scheduling & Fairness

Running concurrent experiments on LiveLabs raises an interesting scheduling problem as it may induce participant fatigue caused by receiving multiple treatments within a short period and this could deteriorate the quality of responses. While running three concurrent experiments, we observed that the participants chosen for the experiments had a 78% overlap and amongst those who responded there was a 58% overlap. In addition, we learnt that we also needed to maintain fairness – i.e. experiments with contradictory outcomes should not be run with overlapping participants within a short time period of each other. For example, scheduling an experiment which promoted soft drinks after one that promoted healthy...
living, with some overlapped participants. Currently, we interview all researchers to identify what they are testing for and what conditions would create bias. We then manually add the required exclusions to LiveLabs to maintain fairness.

To reduce the effect of fatigue and bias, we plan to integrate recent techniques [43, 31, 32] that predict the best times to interrupt participants as well as build mechanisms to allow experimenters to specify the amount of experiment overlap, if any, they can accept.

6.5.2 Uncertainty Handling

One of the questions we received from some of our colleagues was “What is the error in the results?” As discussed in Section 3.4, this is an active area of research. In particular, we are working with our colleagues, with machine learning and statistics expertise, to develop models to propagate the errors at every level of the system, in the form of error distributions, upwards where it is finally combined, using weighted distribution merging techniques, to create a final error distribution for the overall experiment. However, this is preliminary work that is promising but not usable yet.

7. RELATED WORK

Behavioural Experimentation: EmotionSense [44] used a rule-based logic inference engine which allowed social scientists to declare contexts, e.g., emotions, location and activity, relevant to their experiments. Social FMRI [2] combined rich smartphone data collection with intervention capabilities for a small family setting. More recently, Vivo [19] proposed to integrate smartphone-based crowdsensing with IoT to improve context analytics. LiveLabs is similar in vision but much larger in scale – in terms of venue provided, experiment types, and participant pool.

Mobile and Wireless Testbeds: PhoneLab [38] is the most similar to LiveLabs and provides custom Android smartphones to hundreds of university students to study and experiment with various smartphone technologies. SmartLab [28] built an open cloud of smartphones to conduct various mobile systems research. Wireless testbeds such as ORBIT [47] and WISEBED [11] allow testing of network protocols and technologies in laboratory settings, as opposed to simulations, whereas TFA [9] provided a city-scale mesh-networking testbed. The Santander testbed [48], is a large-scale IoT testbed geared towards smart city-type applications such as transport tracking and waste management. Mobilyzer [40] is a library for network measurements that can be seamlessly integrated with any app. LiveLabs differs from these testbeds by providing the opportunity to study completely natural human behaviour – i.e., the participants are using their own devices and are interacted with discretely – i.e. they don’t know why they received any particular notification. In addition, LiveLabs provides a platform that allow numerous research disciplines to innovate within their research areas. The panOULU WLAN testbed [41] is a city-scale, municipal wireless network, started with the goal of providing open and free internet to all. Although technically different to LiveLabs, it uses an academia-industry-government relationship model that is economically viable and plausible for testbeds such as LiveLabs.

Smartphone Sensing-based Studies: In Wang et al. [57, 58], the authors use smartphone-based sensing to derive mental well-being and other behavioural insights. NetSense [53] distributed instrumented smartphones to several hundred incoming freshman undergraduate students. Device Analyzer [56] is an Android App that collects a broad range of smartphone usage statistics, primarily as a means for personal analytics. This is one of the largest such deployments with over 33,000 contributors spread across the world. Smartphone-based Experience Sampling Methods [30] is also a popular method to query participants, in real-time, based on contextual triggers. LiveLabs differs in that it is a complete testbed that focuses on a broad range of experiments from a diverse set of experimenters. In particular, many of these previous works could be re-run on LiveLabs.

8. ACKNOWLEDGEMENTS

We sincerely thank our shepherd Geoffrey Challen for providing excellent iterative feedback that greatly improved our paper. We also thank the anonymous reviewers for their valuable feedback.

This research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IDM Futures Funding Initiative and partially by the Singapore Ministry of Education’s Academic Research Fund Tier 2 under research grant MOE2011-T2-1001. Kasthuri Jayarajah is supported by an A*STAR Graduate Scholarship. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies or Singapore Management University.

LiveLabs has been built by the efforts of multiple people over the years. The authors would like to humbly thank all current staff and students (as of May 1st 2016), namely Jonathan Wang, Sipei Huang, Kazae Quek, Luar Shu Hui (Admin), Francis Lim, Thannermalai Lakshmanan (Outreach), Ignatius Tan (PM), Tan Kiat Wee, Jeena Sebastian, Ritesh Kumar, Swetha Gottipati, John Lee, Tran Anh Quan, Le Hai Son, Aniket Pujara, Tieu Thi Thanh Chau, Prakhar Vishnoi, Huynh Quang Trinh, Arpita Chopra (Engineering), and Sougata Sen, Joseph Chan, Amit Sharma, Huynh Nguyen Loc, Huynh Nguyen Phan Sinh, Nguyen Huy Hoang Huy, Tran Huy Vu (Ph.D. Students), Vigneshwaran Subbaraju, Thivyans Kannappan, Nikita Jaiman, Randy Daratan, Bharat Dwivedi, Desmond Koh (LiveLabs associates), Jing Lingxiao, Jie Xiong (Faculty). In addition, we also humbly thank our former colleagues who helped build up LiveLabs. These include Narayan Ramasubbba (Faculty), Azeeem Khan, Tran Tu Duong, Lu Yu, Rijurekha Sen, Rajiya Laksami Lella, Xiaonan Guo (Postdocs), Yonne Mok, Daphanie Tay, Foo Zhi Yu (Admin), Walter Duno, Kenneth Fu (Outreach), Vivienne Tan (PM), Le Gia Hai, Nguyen Vu Nhat Minh, Harisankar Krishnasawamy, Vikash Ranjan, Khoi Quee Boon, Tadashi Okoshi, Sharib Anis, Nikunj Handa (Engineering), Kartik Murudalhiran, Nguyen Huynh (Ph.D. Students), Shrigruru Nayak, Kuldeep Kulshreshtha, (LiveLabs Associates), Chetna Vig, Joshua Tan, Kok Kiat Han, Nishant Saini, Rahul Majethia, Saharsh Tiwreval, Deepankar Tyagi, Abhinav Kumar, Varun Gandhi, Satayadip Chakraborti, Gomamat Prasanth, Ananda Krishna, Aryan, Praveen Kumar, Nakul Gulati, Changhun Lee, Yeo Guo Kuan, Clarence Koh, Zaman Lantra, Gihan Hettiarachchi, Amila Jayawickrama (Undergraduate Interns), and Anudipa Maiti, Nairan Zhang, Savina Singla (Ph.D. Interns).

9. CONCLUSION

We presented LiveLabs, a large testbed for conducting in-situ real-time mobile behavioural experiments. LiveLabs uses many deep technology components to create an environment where innovative and important insights into human behaviour can be deeply understood. We showed how researchers, from many different disciplines around the world are already using LiveLabs to generate those new insights. In addition, LiveLabs can also be used to test many different technology components – especially those in the mobile sensing, analytics, privacy, and applications domains. Finally, LiveLabs is free to use and open to any researcher – visit http://is.gd/livelabs for details.
10. REFERENCES


[26] Azeem Javed Khan, Vikash Ranjan, Trung-Tuan Luong,


