Three is a crowd? Our experience of testing large-scale social software in a usability lab

Dana McKay
University of Melbourne
Parkville, VIC 3010, Australia
dmckay1@student.unimelb.edu.au

Kagonya Awori
Microsoft Research Centre for Social NUI
University of Melbourne
sawori@student.unimelb.edu.au

Hasan Shahid Ferdous
Microsoft Research Centre for Social NUI
University of Melbourne
hferdous@student.unimelb.edu.au

ABSTRACT
'In the wild' testing has been the cornerstone of HCI in past attempts to create large scale social software, such as conference software. Conversely mobile software is frequently tested in a lab environment, thus banishing typical context of use. In this paper we present our attempt at merging the two approaches for conference social software. We tested in the lab, but attempted to replicate some of the social context of field-based testing. We report our learnings and propose future research for this type of hybrid testing.

Author Keywords
User testing, mobile software, social software

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
There is a long-standing debate in the mobile HCI literature as to whether better results are obtained in lab-based or field-based user testing. Early work suggests that field-based testing reveals problems not seen in lab testing, though these were typically hardware usability issues, rather than software (Waterson et al. 2002). A second comparison found the converse, that there is little advantage in field-based testing despite the seemingly obvious drawbacks to lab-based testing (Kjeldskov et al. 2004). A more recent comparison showed similar results—that there is little difference between lab- and field-based testing (Sun et al. 2013). This contrasts sharply with a non-mobile comparison of remote and lab-based user testing (Greifeneder 2011), which shows that remote user testing gets results that are at odds with lab-based testing. Whatever the advantages and disadvantages, lab-based testing is more common than field-based testing even for context-aware mobile systems, probably due to cost and difficulty gathering field-based data (Kjeldskov et al. 2003; Eshet et al. 2014).

None of the examples listed above, however address software that is specifically designed to improve in-person socialisation, such as conference software. The literature has long recognised the need to develop software to support conference attendees in getting the full benefits of conferences (McCarthy et al. 2004; Ross et al. 2011; Hoffelder 2013), particularly those who are from a non-English speaking background or who are new to academia (McCarthy et al. 2004). These systems often rely at least partially on mobile technologies, usually some kind of smart badge, for example (Cox et al. 2003). Testing of conference systems is near-invariably conducted in the field, perhaps because the social aspects of a conference are challenging to replicate in the lab.

OzCHI '15 , December 07 - 10 2015, Melbourne, VIC, Australia
ACM 978-1-4503-3673-4/15/12... $15.00
http://dx.doi.org/10.1145/2838739.2838827

Relying on a conference to test conference-based software is a risky strategy. Conferences typically do not occur more than annually, and small usability problems can easily derail testing. Similarly, this approach is likely to be expensive, particularly if it requires bespoke hardware, and it limits the opportunity to conduct the kind of detailed observations that make for successful fieldwork (Kjeldskov et al. 2003). Relying on conferences to test also does not allow for the rapid prototyping and development called for in good user centred design practice (Rogers et al. 2007).

Being able to test mobile social software such as conference systems in a lab environment in a realistic way is likely to result in improved field-testing; but is it even possible? What would 'realistic' lab-based testing of conference software look like? How can large, complex social settings be replicated quickly and cheaply? In this paper we present our experience of attempting exactly this: quickly and cheaply testing a conference prototype in a lab-based environment, but attempting to test in a way that replicates a large, multifaceted social context.

The remainder of this paper is divided up as follows: First, we describe our prototype; then we discuss our approach to testing. Finally, we reflect on the efficacy of our approach and lessons learned for the future.

OUR PROTOTYPE
While our key interest in this paper is addressing the challenges of generating social context in lab-based testing, it is impossible to describe our approach to testing without first examining our test system. We first describe our concept, then the components of our system, then finally the implementation we used in testing.

Concept
Our prototype was conceptualised and created using a traditional user-centred design approach. First, we

reviewed the literature on conference systems to establish what had worked (or not) in the past. Next, we conducted a requirements gathering exercise interviewing a number of academics at various stages in their career about both their experiences of conferences and their requirements of software for conference socialising. The literature and our requirements gathering exercise identified four key problems faced by academic conference attendees:

1. Identifying someone known only by name
2. Locating attendees with similar interests
3. Finding an interesting conversation partner (whom attendee has never met before), and breaking the ice with them
4. Determining when to interrupt an ongoing conversation, and conversely managing interruptions.

To address these needs, we designed a prototype that had two core parts: a technology-enabled conference badge; and an ambient display coupled with a smart room layout. The intended users of our prototype are conference attendees that are new to a conference or unfamiliar with other attendees, those that would like to converse with particular conference attendees, and those who need support in socialising or networking with academics in their fields of interest.

The conference badge

The conference badge is designed as a location-aware smart technology that supported a number of functions. These functions included setting up and cancelling meetings with other conference participants; locating people with similar interests and providing a conversation starter (an icebreaker); addressing questions to the conference at large; and identifying speakers (including those asking questions) during conference sessions.

The conference badge is reliant on a combination of technology and user factors. Conference participants must provide a profile prior to the conference. This profile includes their academic backgrounds, topics of interest and availability during the conference. Location awareness is required in order to identify nearby conversation partners. Finally, voice recognition and activation is one of the channels by which the user could interact with the conference badge.

We tested two functions in the lab: finding an interesting conversation partner, and setting up a time to meet a speaker. These are described in more detail below.

Finding an interesting conversation partner

This function was designed to be used during breaks, and in our interface was labelled 'I'm Bored'. The match algorithm we propose for *I'm Bored* is based on proximity (locating a nearby conference participant who had also indicated boredom), and shared interest (based on the interests attendees indicate in their profiles). Once a match is found, the respective participants' badges show them a photo of each other with name and interest information. Attendees can then accept or reject the match. When matched attendees are in speaking distance proximity, the badges buzz and light up saying they should meet. The *I'm Bored* feature was designed to alleviate the problems identified in our user research. Interruptibility is managed by matching only those who indicate they are also 'bored'. The challenge of not knowing other conference attendees is met by creating opportunities for users to meet and converse.

Set up a meeting with a speaker

This function is for use during conference presentations and sessions. Users can request meetings with the presenters or anyone speaking at the time (e.g. someone asking a question). The badge identifies speakers via their voices and locations. Each badge contains the user's conference schedule and availability, meaning users can only request meetings with other attendees at unallocated times. During talk sessions, attendees can request a meetings with speakers/presenters by scheduling a time and place to meet. The request will in turn set up an alert on the speaker's badge; speakers can then accept or reject meeting requests.

Smart room and ambient display

The smart room is designed with physically fixed spaces dedicated to topics of interest within a conference. Topics of interest are identified by aggregating the research interests of all conference participants. These spaces should each have tables and chairs to allow people to talk, and a large screen where on-topic questions can be posted from users' conference badges. The questions posted to the screen also serve as icebreakers for those gathered in the space.

A complementary ambient display should be located in a central place, and show where topics are clustered based on attendee interests. The ambient display is designed to make it more likely for attendees to find others with shared interests, simply by being in the right physical place.

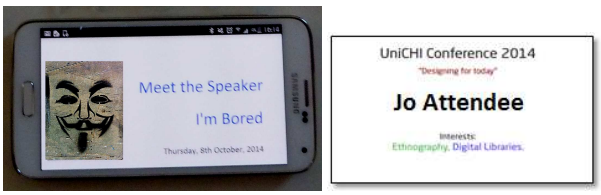

Figure 1: Prototype smart conference badge front (left) and back (right)

Implementation

Our prototype conference badge comprised mobile phones duct taped to standard conference badges on lanyards (see Figure 1). The interactive elements of the badge were implemented on the phone as high-fidelity prototype. This functionality was rapidly developed using user interface images with hardcoded information (e.g., profile picture, research interests, and schedule) for each participant. Interactive user interface elements were implemented using a free web based service[1]. Context aware alerts were simulated with text messages in a Wizard of Oz approach; participants were separated from the wizards by a one-way glass partition.

[1] Flinto, www.flinto.com

The prototype ambient display (see Figure 2(was projected onto a screen in the test room; only one participant was tracked at a time. We again used a Wizard of Oz approach for testing. The non-moving participants (who were actually images, rather than people—see the method described below) vibrated slightly on the screen to simulate small movements, and to generate visual interest.

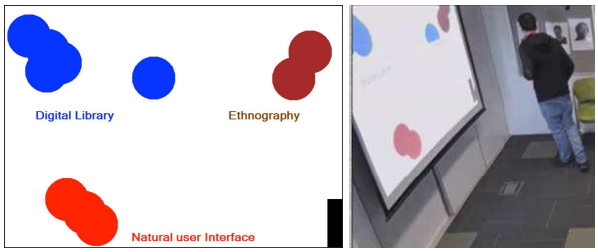

Figure 2: Ambient display, showing detail (left) and in-situ implementation (right)

TESTING AND APPROACH

While we recognise that some readers will find the details of our implementation interesting, it was a means to an end—we particularly wanted to see whether we could generate social context in lab-based testing. In this section we will describe our approach to testing.

Given that we were evaluating our test approach rather than our system, we opted for a pilot-study approach. We conducted our study with three participants, each performing three tasks. All the participants were graduate students in computer science; all had previously attended conferences. One participant was female, two male. Testing was conducted at a university usability lab.

The three tasks participants performed were:

1. Find someone interesting to talk to right now
2. Find people with shared interests
3. Set up a time to meet the speaker.

Data was collected using video recordings, mobile screen capture, handwritten notes and post-test interviews.

Testing was done in two phases. The first phase was the social testing phase that consisted of tasks 1 and 2 above. This phase replicated informal time at a conference, for example the coffee break. The second phase, comprising task 3 above, was the conference session testing phase. This phase replicated the formal settings of conferences, e.g. a paper presentation session.

The social testing phase was the more challenging of the two phases. Here, we had three participants and three testers—two of whom ran our Wizard of Oz prototype. The limited number of participants and testers made it difficult to replicate the typically busy environment of a conference, which would have been ideal for us to test the *I'm Bored* functionality (task 1) and the ambient display (task 2). To alleviate this difficulty we created 'paper participants'. These were images of researchers and their conference badges, printed out life-size and hung around the walls of the room (see Figure 3). Additionally, to enable the testing of the ambient display (task 2), these paper participants were clustered according to their interests. Participants were instructed to treat paper participants as though they were *real* conference attendees.

Figure 3: A paper participant

The social testing phase was done in three cycles, where each task was performed by only one participant at a time. Thus, each participant rotated through tasks 1 and 2 above and performed an additional control task. The control task was to approach the in room tester and strike up a conversation without the use of technology This task was to see how each participant would interact in a conference situation, and to give the 'spare' participant a task while the other two participants were interacting with the prototype. Each cycle took approximately 5-10 minutes.

In task 1 of the social testing phase, participants were required to find someone interesting to talk to. Once the participant selected *I'm Bored*, the prototype would display the profile of a conference attendee who had similar academic interests as them, and who was also available to socialise. The participant then walked around the room in order to visually identify the conference attendee that was displayed on their device. Once the participant was close to the attendee, the testers—who were behind a one way glass control room—sent an SMS to the participant's phone, mimicking the context aware functionality of the prototype.

In task 2 of the social testing phase, the participant was required to find a group of people in the room who had interests similar to theirs. This task relied on participants noticing the colour coded blobs of interest moving around on the ambient display, and that their blob moved when they did. They then needed to move to the area where there were other blobs of the same colour. The blob tracking and movement was controlled by the wizards in the control room.

The second phase of testing—conference session testing—was straightforward. We simply assigned all participants the task of scheduling a time to meet the speaker (task 3), and had one tester act as the speaker. To meet the speaker, the participant simply had to select *Meet the Speaker* on the prototype and schedule a time.

The social testing phase was carried out first, and then the conference testing phase. Between the phases, participants were allowed a break. This also gave time to rearrange the room from a large open space, to rows of chairs as is in a conference session. All paper participants were removed from the walls during the break.

LEARNINGS

In this section we discuss what we learned not about our prototype, but about testing social software in a lab environment; the key goal of our testing. Some elements of testing were more successful than others.

Failings of this approach

We faced a number of problems during testing, including physical problems with our prototype and technological problems affecting both our prototype and the screen capture software.

Our prototype was implemented quickly and cheaply by taping mobile phones to conference badges. This had its drawbacks, however: even with duct tape the phones came unstuck from the badges, and the rotation of the phones caused our app to crash or malfunction on several occasions.

Given that this testing was intended to be lightweight and inexpensive we elected to use readily available tools and software. We thus purchased mobile screen capture software. The software was effective during pilot testing but failed in situ. Any attempt to use this testing approach in practice would require reliable screen capture software.

Testing the ambient display failed for completely different reasons: in a small room with a limited number of (non paper) people it was simply not comprehensible. Given the large-scale nature of a typical conference ambient display (McCarthy et al. 2004; McDonald et al. 2008) and the reliance on multiple data points for the display in most cases, testing for these systems apparently requires the big data context in which they are designed to operate.

Successes of this approach

A number of elements of this approach worked surprisingly well. The conference setting testing was highly successful in identifying the influences of context, and participants engaged well with the paper participants in the social setting. Additionally, the prototype, while literally held together with duct tape, was a successful vehicle for testing the underlying concepts of our proposed system.

During the conference testing phase, we identified two approaches to meeting the speaker that were unexpected: one of our participants revealed in his post-test that he marked the speaker he was trying to meet with 'interesting' in an attempt to "kiss tail" before requesting an appointment; two participants discussed when they had scheduled to meet with the speaker, providing a potential ice breaker between conference attendees. These findings would not have emerged had we not tested in a contextual setting.

A large part of our testing was to determine whether (and how) our participants would interact with 'paper participants', making it possible to test a prototype that relies on a crowd, with a limited number of users. Our participants were not actors (an approach sometimes used in HCI (Newell et al. 2006)), they were academics. Nonetheless, they actively engaged with paper participants, giving us valuable information about the usability of our prototype.

DISCUSSION AND CONCLUSIONS

Early system and concept testing is a core part of user centred design (Rogers et al. 2007). It is also apparent that context is (or may be) an important element in system testing of all kinds (Kjeldskov et al. 2004; Greifeneder 2011; Eshet et al. 2014); the concept of living labs, for example, is based entirely in the value of context (Bergvall-Kareborn et al. 2009). These two demands are often in competition: to test early requires relatively simple, inexpensive testing; in contrast contextual testing requires time and effort (Kjeldskov et al. 2003; Eshet et al. 2014). Given this conflict, there is a gap around lab-based social context testing. Conference systems, for example, are invariably tested in context (see for example (Cox et al. 2003; McDonald et al. 2008; Hoffelder 2013)). To our knowledge no-one has attempted lightweight lab-based contextual testing; we have begun to address this gap in the work presented here. We discovered that lab-based testing with a limited number of participants can afford interesting and valuable discoveries early on in system development, with some limitations. This type of testing, in our experience, was useful in individual and group interactions, but less useful for system mediated information (such as the ambient display). Why paper participants worked while the ambient display failed is unclear; perhaps it is because we are accustomed to person-based roleplay from childhood, but interpreting large scale data from a display is a skill acquired only later, and possibly not by everyone.

Clearly one experience does not make for a new testing method, and it would be interesting to attempt this approach in another social software domain. Our experience, though, is that this approach has potential for valuable insight; whether this is true in other domains remains future work.

ACKNOWLEDGMENTS

The Anonymous image on the conference badge in figure 1 is used under a Creative Commons By-Attribution Share-Alike license. The original is available at: https://www.flickr.com/photos/doctorow/12589013215/.

We thank our participants for their willingness to try something interesting with us.

We thank George Buchanan for allowing us to use his likeness in this paper.

Dana thanks her dear kitty Satchmo for their 16 together; he died the day this paper was submitted.

REFERENCES

Bergvall-Kareborn, B., Hoist, M. and Stahlbrost, A. (2009). Concept Design with a Living Lab Approach. System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on.

Cox, D., Kindratenko, V. and Pointer, D. IntelliBadgeTM: Towards Providing Location-Aware Value-Added Services at Academic Conferences. Proc. UbiComp 03, Springer Berlin Heidelberg. (2003), 264-280.

Eshet, E. and Bouwman, H. Addressing the Context of Use in Mobile Computing: a Survey on the State of the Practice. Interacting with Computers 27, 4, (2014).

Greifeneder, E. The Impact of Distraction in Natural Environments on User Experience Research Proc. TPDL 11, Springer. (2011), 308-315.

Hoffelder, N. Going to the Open Hardware Summit? Don't Forget Your E-ink Conference Badge, The Digital Reader (2013). Available at: http://the-digital-reader.com/2013/07/27/going-to-the-open-hardware-summit-dont-forget-your-e-ink-conference-badge/

Kjeldskov, J. and Graham, C. (2003). A Review of Mobile HCI Research Methods. Human-Computer Interaction with Mobile Devices and Services. L. Chittaro, Springer Berlin Heidelberg. 2795: 317-335.

Kjeldskov, J., Skov, M., Als, B. and Høegh, R. Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. Proc. Mobile HCI 04, Springer Berlin Heidelberg. (2004), 61-73.

McCarthy, J. F., McDonald, D. W., Soroczak, S., Nguyen, D. H. and Rashid, A. M. Augmenting the social space of an academic conference. Proc. CSCW 04, ACM. (2004), 39-48.

McDonald, D. W., McCarthy, J. F., Soroczak, S., Nguyen, D. H. and Rashid, A. M. Proactive displays: Supporting awareness in fluid social environments. ToCHI 14, 4, (2008) 16.

Newell, A. F., Morgan, M. E., Gregor, P. and Carmichael, A. Theatre as an intermediary between users and CHI designers. Proc. CHI 06, ACM. (2006), 111-116.

Rogers, Y., Sharp, H. and Preece, J. Interaction design: beyond human-computer interaction. Chichester, West Sussex, England, Wiley (2007).

Ross, C., Terras, M., Warwick, C. and Welsh, A. Enabled backchannel: Conference Twitter use by digital humanists. J Doc 67, 2, (2011) 214-237.

Sun, X. and May, A. A comparison of field-based and lab-based experiments to evaluate user experience of personalised mobile devices. Adv. in Hum.-Comp. Int. 2013, (2013) 2-2.

Waterson, S., Landay, J. A. and Matthews, T. In the lab and out in the wild: remote web usability testing for mobile devices. Proc. CHI 02, ACM. (2002), 796-797.