Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours

Jorge Goncalves[1], Denzil Ferreira[1], Simo Hosio[1], Yong Liu[1], Jakob Rogstadius[2], Hannu Kukka[1,3], Vassilis Kostakos[1]

[1]Department of Computer Science and Engineering, University of Oulu, Finland
[2]Madeira Interactive Technologies Institute, University of Madeira, Portugal
[3]HCI Institute, Carnegie Mellon University, USA
[1]firstname.lastname@ee.oulu.fi, [2]jakob.rogstadius@gmail.com

ABSTRACT

This study is the first attempt to investigate altruistic use of interactive public displays in natural usage settings as a crowdsourcing mechanism. We test a non-paid crowdsourcing service on public displays with eight different motivation settings and analyse users' behavioural patterns and crowdsourcing performance (e.g., accuracy, time spent, tasks completed). The results show that altruistic use, such as for crowdsourcing, is feasible on public displays, and through the controlled use of motivational design and validation check mechanisms, performance can be improved. The results shed insights on three research challenges in the field: i) how does crowdsourcing performance on public displays compare to that of online crowdsourcing, ii) how to improve the quality of feedback collected from public displays which tends to be noisy, and iii) identify users' behavioural patterns towards crowdsourcing on public displays in natural usage settings.

Author Keywords
Crowdsourcing; public displays; altruism; motivation.

ACM Classification Keywords
H5.m. Information Systems: Miscellaneous.

INTRODUCTION

Crowdsourcing has been adopted as an umbrella term to refer to the coordinated approach in which a computationally challenging task is broken down into several pieces. Those pieces are subsequently "solved" by humans (referred to as "workers") and the results are finally combined to construct an overall solution to the problem. People complete these for a number of reasons including payment, altruism or simply contribute without being aware (e.g. reCAPTCHA [30]). Crowdsourcing is highly effective for tasks that can be parallelized. The emergence of online crowdsourcing "markets" (such as Amazon's Mechanical Turk) make it convenient to pay for workers willing to solve such small problems, referred to as "tasks". Most importantly, the existence of these crowdsourcing markets makes it possible for researchers to conduct controlled experiments to investigate crowdsourcing itself and what affects task performance. However, online crowdsourcing markets do not always attract workers of desired background or skills. For instance, 80% of workers in Amazon Mechanical Turk are from the United States or India [15]. Therefore, it is a challenge as to how to recruit workers who speak a particular language or live in a given city [15].

Researchers have explored ways of crowdsourcing tasks onto mobile phones, thus pushing the tasks to the workers, anywhere and anytime. Mobile crowdsourcing has limitations of its own, such as the need for workers to own a compatible device, or having a convenient schedule of tasks for the workers. Also it often requires a sign-up procedure, and more crucially, mobile devices are battery and resource constrained. Here we explore a passive approach of crowdsourcing tasks to workers, by embedding public displays into a physical space and leveraging workers' serendipitous availability.

Public displays are becoming increasingly affordable, and researchers have systematically attempted to identify novel applications for this technology. Research on interactive displays in public spaces has often noted that users typically demonstrate playful and exploratory behaviour when using this technology. As a result, the collection of feedback on such displays is difficult, and results tend to be highly noisy [4, 13]. A reflection on the effective facilitation mechanism for public displays to motivate users to deliver reliable and meaningful feedback is lacking but also imperative. Most prior research has reported the use of public displays for hedonic services (*i.e.*, games, opinion disclosure) or information-based services (*i.e.*, information boards) that offer instant benefits to users [4, 7, 20]. There is a lack of deliberation on the possibility of using public displays in an altruistic manner, such as for non-paid crowdsourcing. Pragmatically, a successful demonstration of the potential of public displays for altruistic services implies a possible future direction for public displays research and practice.

In this paper we demonstrate that crowdsourcing performance on public displays is comparable to online crowdsourcing markets. More specifically, by replicating an experimental task in both we show that performance on public displays is similar to Amazon's Mechanical Turk and in fact task completion rate is an order of magnitude faster. Furthermore, we show how the use of motivation and fact-checking mechanisms can significantly improve task performance on public displays. Our results

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-1770-2/13/09…$15.00.
http://dx.doi.org/10.1145/2493432.2493481

demonstrate that public displays are a reliable channel for conducting crowdsourcing studies, and our extensive *in-situ* observations reveal important insights into how workers behave when they are completing tasks on public displays.

We must emphasize that our approach to crowdsourcing does not replace existing ones; it complements them. The *in-situ* nature of public displays offers a number of new opportunities for crowdsourcing studies, which we discuss further in this paper.

RELATED WORK

Studies on crowdsourcing have shown that workers often require some kind of motivation to participate and effectively donate their time to a cause. In other words, crowdsourcing performance and uptake is not just a matter of channel or medium, but also a matter of motivation [17] and instrumentation [18]. Kaufmann et al. [17] identify two major types of motivation for crowdsourcing: intrinsic and extrinsic. They note that *intrinsic motivation* can be *enjoyment-based* (related to the fun and enjoyment that the contributor experiences through their participation) and *community-based* (related to community participation, and include community identification and social contact). *Extrinsic motivation* can relate to having immediate or delayed payoffs (including material benefits) as well as social motivation (such as values and beliefs).

In terms of *instrumentation*, crowdsourcing performance can be substantially improved by taking a number of steps when designing the task itself. For example, previous work has suggested the inclusion of *explicitly verifiable questions* [18] when developing the experiment. The inclusion of these “fact-checking” questions has been shown to improve the quality of completed tasks – as workers become aware of prompt response verification. Furthermore, the actual *difficulty* of the task has an effect on task performance [27]. Specifically, for tasks with higher difficulty levels, workers may simply give up or provide an approximate answer for the task. Therefore, less difficult tasks are more likely to be completed [12].

Crowdsourcing beyond the desktop

Crowdsourcing with ubiquitous technologies is increasingly gaining researchers’ attention [31], especially on mobile phones. Targeting low-end mobile phones, txtEagle [8] is a platform for crowdsourcing tasks specific to habitants of developing countries. Similar platforms are MobileWorks [24] and mClerk [11] that specifically focus on asking users to convert handwritten words to typed text from a variety of vestigial dialects. Targeting smartphones, Alt *et al.* [1] explore location-based crowdsourcing for distributing tasks to workers. They focus on how workers may actively perform real-world tasks for others, such as giving a real-time recommendation for a restaurant, or providing an instant weather report wherever they are. Similarly, Vaataja et al. [29] report a location-aware crowdsourcing platform for authoring news articles by requesting photographs or videos of certain events from its workers. Mashhadi & Capra [22] suggest using contextual information, such as mobility, as a mechanism to ensure the quality of crowdsourced work.

Finally, a very active community has developed around the topic of crowdsourcing measurements and sensing. This participatory sensing movement is also referred to as “Citizen Science” [25] and relies on mobilizing large parts of the population to contribute to scientific challenges via crowdsourcing. Often this involves the use of mobile phones for collecting data [6, 9] or even donating computational resources while the phone is idle [3].

Despite the appeal of mobile phones, using them for crowdsourcing requires workers’ *implicit* deployment, configuration and use of the device. For example, in SMS-based crowdsourcing, participants need to explicitly sign up for the service, at the cost of a text message exchange. This challenges recruitment of workers, as a number of steps need to be performed before a worker can actually start contributing using their device. Contrary to mobile crowdsourcing, public displays crowdsourcing does not require any deployment effort from the worker to contribute.

Data collection with public displays

Not surprisingly, a number of previous studies have investigated the use of public interactive displays for the purpose of collecting data, most often collecting explicit human input [2, 4, 13]. *Ubinion* [13] was a service that employed large public interactive displays to capture youngsters’ personalized feedback on municipal issues to the local youth workers. While successful in engaging the young, the study reported a high level of *appropriation* of the public displays. The intended purpose of the system – to give feedback on youth related matters – took a backseat to user self-expression and playful interactions.

Brignull & Rogers’s [4] *Opinionizer* is a system designed and placed in two authentic social gatherings (parties) to encourage socialization and interaction. Participants could add comments to a publicly visible and shared display. During the study the authors found that a major deterrent preventing people from participating is *social embarrassment*, and suggest making the public interaction purposeful. The environment, both on and around the display, also affect the use and data collected, as the environment produces strong physical and social affordances which people can easily and unambiguously pick up on. Hence they argue for facilitating the public in its needs to rapidly develop their conceptions of the purpose of the social activity, and to be able to move seamlessly and comfortably between being an onlooker and a participant.

A further study that considered public displays as data collection mechanisms was *TextTales* [2]. Here the authors attempted to explore the connection between story authorship and civic discourse by installing a large, city-scale, interactive public installation that displays a 3-by-3 grid of image-text combinations. A discussion on a certain photograph would start with SMSs sent by users, displayed in a comments stream. The comments of TexTales users deviated significantly from the “intended” topic of discourse, *i.e.*, the theme set by the photographs. More importantly, this study highlights the challenges in harnessing the general public in natural usage settings for a tightly knit purpose.

Literature suggests that people are interested to use public display deployments [2, 4, 13], but with personal motives in mind resulting in strong appropriation of the technology. In our study we aim to leverage people’s willingness to engage with this technology, but motivate the appropriate use of the system by through the use of *motivational* and *instrumentation* mechanisms.

STUDY

The aim of the study is to establish whether crowdsourcing can successfully be used to harness the willingness of users to spend time on public displays to produce a valuable input. In addition, it aims to investigate the impact of intrinsic *motivational* techniques and *fact-checking* mechanisms on the performance of crowdsourcing workers.

Apparatus

Four 46” full-HD LCD displays with a touchscreen overlay (Figure 1) were placed throughout our university campus. The four chosen locations had a steady flow of people passing by and were effectively busy walkways (*i.e.*, main corridors). The campus has about 18000 registered students and staff, but we expect that a subset of these visit the university on a daily basis.

Experimental Task

The task used in our experiment is the counting task proposed by Rogstadius *et al*. [27] in which workers are asked to count malaria-infected blood cells on images of a petri dish generated algorithmically. This task was chosen because it has a set of desirable characteristics:

- The complexity of the task is *varied systematically*;
- The *correct answer* for each task can be objectively determined; and
- The task is *realistic* enough to convince users that it offers *value*.

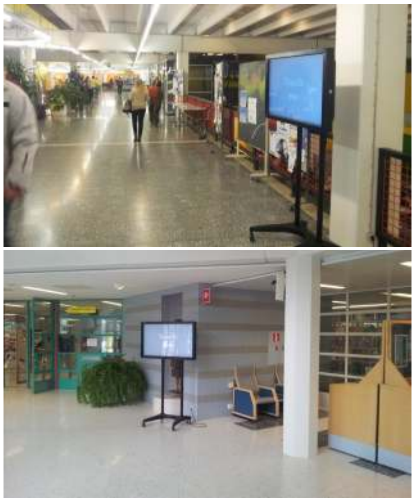


Figure 1. In-situ photographs of two displays used in our experiment.

	Fact-check	
	present	**absent**
No motivation	Condition 1	Condition 2
Enjoyment	Condition 3	Condition 4
Community	Condition 5	Condition 6
Enjoyment & Community	Condition 7	Condition 8

Table 1. The 8 conditions derived from the 4x2

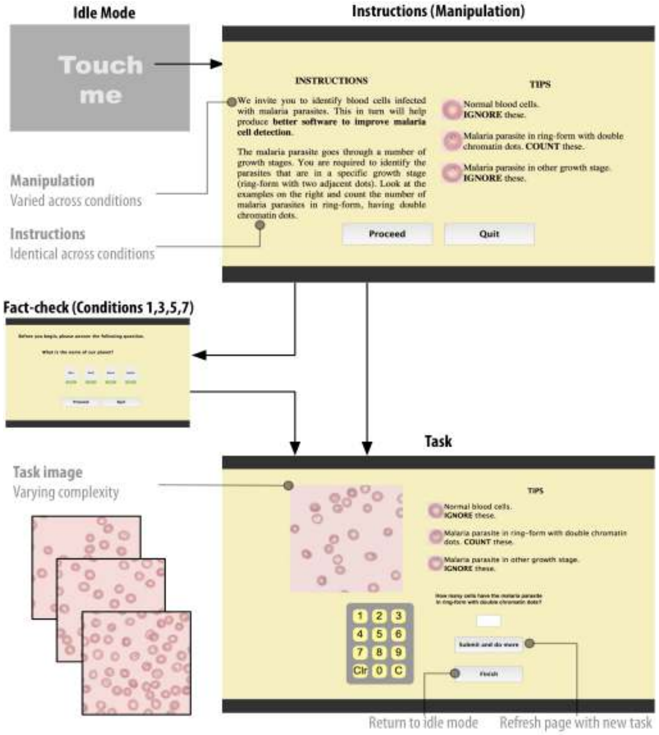
Furthermore, the adoption of this task allows us to compare our results with equivalent results obtained online from Mechanical Turk. We decided to reuse the set of images that Rogstadius *et al*. [27] made available, along with the obtained results from Mechanical Turk. These images present an increasingly complex counting task, as shown in Figure 2.

During a pilot study we observed that the images of high difficulty were relatively time-consuming and challenging on a public display. For this reason, we chose a subset of the original image set, *i.e.*, the 30 images with lowest complexity.

The displays showed initially a set of instructions and subsequently a legend (Figure 2) with each task for easy reference. Both the instructions and the legend were identical across all conditions in the experiment to remove any potential bias:

The malaria parasite goes through a number of growth stages. You are required to identify the parasites that are in a specific growth stage (ring-form with two adjacent dots). Look at the examples on the right and count the number of malaria parasites in ring-form, having double chromatin dots.

To complete the task, participants had to use an onscreen dialpad to type their estimate of the number of cells with the “*malaria parasite in ring-form with double chromatin dots*”. When submitting their answer they could decide to do more tasks, or to terminate their session (Figure 2).



We invite you to identify blood cells infected with malaria parasites. This in turn will help produce **better software to improve malaria cell detection**.

The malaria parasite goes through a number of growth stages. You are required to identify the parasites that are in a specific growth stage (ring-form with two adjacent dots). Look at the examples on the right and count the number of malaria parasites in ring-form, having double chromatin dots.

Normal blood cells. **IGNORE** these.

Malaria parasite in ring-form with double chromatin dots. **COUNT** these.

Malaria parasite in other growth stage. **IGNORE** these.

Before you begin, please answer the following question.

What is the name of our planet?

How many cells have the malaria parasite in ring-form with double chromatin dots?

Figure 2. Screenshot flowchart of the experimental task. Only half the conditions had the fact-check page, the other half skipped it. A worker was allocated to a condition when clicking the Idle Mode page.

Finally, we decided not to actively promote our study in any way to avoid response bias, as we chose to solely rely on the serendipitous nature and attractiveness of public displays [7, 20]. For this reason we simply added a generic enticement to interact with the display as proposed in [21]. When our displays were idle, *i.e.*, not in use, they showed a screen with the words "Touch Me" as seen in Figure 2.

Design

The study was a 4x2 between-subjects experimental design with two variables: **motivation** and **fact-check**. The experimental conditions are shown in Table 1. Each condition contained the same set of images (n=30) for the counting task.

The first independent variable was the *motivation* given to the participants, by manipulating the text on the instruction page that appeared when users first touched the display. We decided to *avoid extrinsic motivation* (such as a monetary reward) and explicitly focus on the altruistic use of the displays. From literature, we identified two types of *intrinsic motivators*: *enjoyment*-based and *community*-based [17]. We used one construct per motivator, to enable reliable testing on a public display.

- **Task identity**: A worker performs a task because he knows that his work will be used (*e.g.,* writing a product description for a website) [17].
- **Community identity**: A worker who only accepts tasks from requesters with a good reputation because they are known as a valuable supporter of the community [17].

Using these two constructs we derived 4 levels of motivation for our experiment. The introductory text on the instruction page (Figure 2) was manipulated based on the condition:

- **Control (no motivation)**: "We invite you to identify blood cells infected with malaria parasites."
- **Enjoyment-based (task identity)**: "We invite you to identify blood cells infected with malaria parasites. This in turn will help produce better software to improve malaria cell detection."
- **Community-based (community identity)**: "We invite you to identify blood cells infected with malaria parasites. This in turn will help Oulu medical scientists on their research."
- **Both enjoyment & community-based (task & community identity)**: "We invite you to identify blood cells infected with malaria parasites. This in turn will help produce better software to improve malaria cell detection and help Oulu medical scientists on their research."

The second independent variable was **fact-check**, with two levels: *present* and *absent*. This can be an effective way to filter out non-serious answers and improve the overall quality of the answers given [18]. An important characteristic of this step is that the question must be easy to answer, and it must be clear to the respondent that the experimenters also know the answer. In those conditions where the fact-check was present, we showed users a question just before starting their first task. The question was: "*What is the name of our planet?*" Users would then have to select their answer out 4 possibilities (Earth, Saturn, Mars, Jupiter) ordered randomly to avoid bias.

Finally, we decided to balance the attributed condition and images shown. This was done to guarantee an even spread amongst conditions and number of times each picture was answered for each condition in order to provide a fair comparison between our manipulations. Our goal was for each picture (n=30) to be answered 5 times on each of the 8 conditions, which meant a total of 1200 answers. To achieve this, the least answered tasks were assigned when a user touched the screen.

Data Collection

All interactions between users and the display were logged. This includes: a unique session ID, condition assigned, answer to fact-check, image ID, answer(s) to the task(s), timestamp for each touch and time spent on each page. We added an inactivity timeout (60s) on every screen so that if a user did not complete the whole process the display would revert to idle mode. Two metrics we calculated were *accuracy* of each response and *complexity* of each image. We calculate the accuracy of a response as a number between 0 and 1 using the formula adapted from [27]:

$$accuracy = 1 - \left(\frac{|p_{est} - p_{real}|}{p_{real}}\right),$$

where p_{est} is the number parasites reported by the user and p_{real} is the actual number of parasites on the particular image.

Furthermore, the complexity of each image was calculated using the formula presented in [27]:

$$complexity = c_{real} + 3p_{real},$$

where c_{real} is the total number of cells in an image and p_{real} is the actual number of parasites in the image. The formula gives greater weight to parasites than to cells because users had to consider the growth stage of the parasite when counting them.

Finally, we obtained the data from [27] for the 30 images we used in our own experiment. This data was originally collected on Amazon's Mechanical Turk following a typical web-based crowdsourcing process. We obtained data on the time workers spent on each task, the rate at which tasks were completed during that study, and using the above formulas we calculated the *accuracy* of each response. The *complexity* of each image was identical across the studies.

Observations and Interviews

We conducted **unobtrusive observations** totalling 24 hours (6 hours per display). During these observations we noted all interactions with and around the screen, including whether tasks were completed, if the user was alone or in a group, and the social dynamics of the interaction.

Additionally, we conducted several semi-structured **interviews** (n=24) during busy hours. These were conducted with people that had interacted with the display regardless of having completed tasks or not. To reduce the likelihood of any bias as suggested by [21], we approached the interviewees after they had walked away from the display in order not to obstruct the display and ensure that passers-by could not infer that people using the display were being interviewed. We discussed with interviewees their experience and motivation for using the displays and their perceptions regarding the tasks.

Finally, we installed a **camera** to constantly record interactions with one of the displays. All displays were in public walkways on our campus, and we were granted ethical and legal permission to video-record without audio. Cameras are also operating in the same area by the security team as well as computer vision researchers. The resolution of the recorded video was downsized to reduce the captured details. At the end of the study we conducted a workshop in which 5 researchers analysed and categorized the video snippets (n=123) of users interacting with the display and cross-referencing this with the logged data from the display. Through our combination of human observations and video recording we aimed to gather insights on users' behaviours when completing our crowdsourcing tasks.

RESULTS

We ended our data collection after 25 days when our goal of 1200 answers spread equally among conditions and pictures was reached. In total our application was launched 1790 times and 482 (27%) of those resulted in at least one answer being given. The overall average accuracy was .74 (SD=.34) and the average time spent completing tasks (excluding the instructions page and fact-check) was 18.69s (SD=16.41, min=1.97, max=166.86). The fact-check question was answered 493 times (82%) correctly and 107 times (18%) incorrectly. Finally, workers completed 2.49 tasks on average (SD=2.35 min=1, max=17) before leaving the display (i.e. per session). Note that we discard sessions where no tasks were completed, because we cannot reliably determine whether workers actually attempted to read the instructions or the display simply timed out.

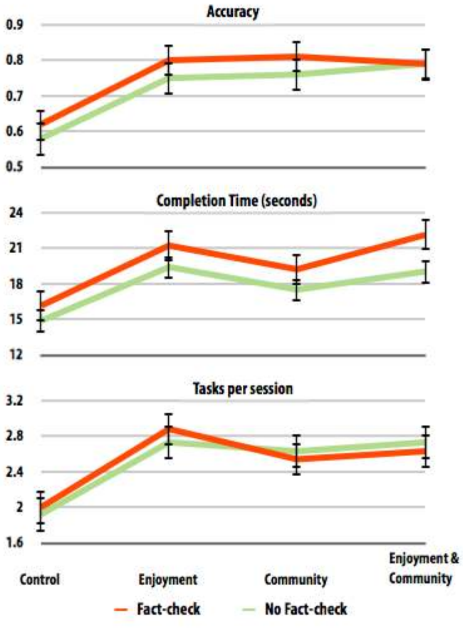


Figure 3. Performance metrics for each condition (averages). Accuracy and time spent refers to individual task; tasks per session refers to the average number of tasks completed by each worker in each visit to the display.

Effects of the Manipulations

Next, we report the effects of our manipulations through the use of between-subjects ANOVAs.

Accuracy

The main effect of fact-check on accuracy was not significant ($F(1,1192) = 0.16$, $p = .69$) but the main effect of motivation on accuracy was significant ($F(3,1992) = 13.93$, $p < .01$). *Post-hoc* comparisons using the Tukey HSD test indicated that workers whose instructions had no motivational approach scored significantly lower accuracy than workers who had a motivational approach present ($p < .05$). These differences can be verified in Figure 4. There was no significant interaction between motivation and fact-check regarding accuracy ($F(3,1192) = 1.592$, $p = .19$).

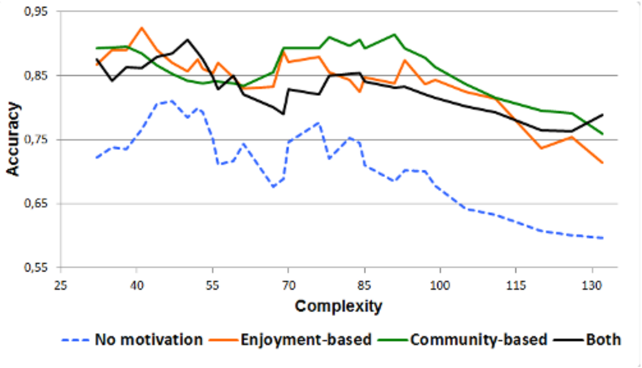


Figure 4. Differences in accuracy across complexity for each level of motivation used.

Completion Time

In terms of task completion time, there was a significant main effect of motivation ($F(3,1192) = 6.07$, $p < .01$) but not for fact-check ($F(1,1192) = 3.47$, $p = .06$). *Post-hoc* comparisons using Tukey HSD test indicated that workers whose instructions had no motivational approach spent significantly less time on each task than workers who had a motivational approach present ($p < .05$). These differences can be verified in Figure 5. There was no significant interaction between motivation and fact-check regarding completion time ($F(3,1192) = .19$, $p = .90$).

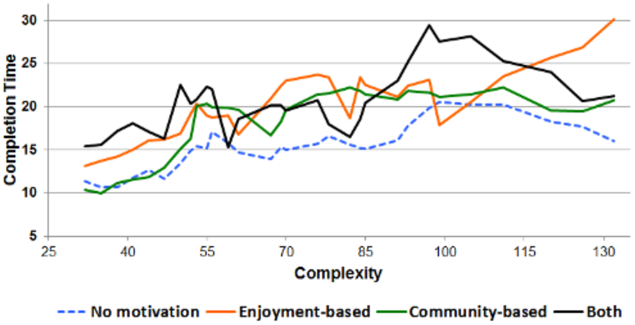


Figure 5. Differences in completion time across complexity for each level of motivation used.

Tasks Completed

We analysed the number of tasks completed per session by workers. We found a significant main effect of motivation ($F(3,474) = 3.45$, $p = .02$) but not for fact-check ($F(1,474) = 0.03$, $p = .87$). *Post-hoc* comparisons using the Tukey HSD test showed that workers whose instructions had no motivational approach completed significantly less tasks than workers who had a motivational approach present ($p < .05$).

Responses to the Fact-check

Considering just the conditions that had the fact-check, we analysed workers' performance based on whether their response to the fact-check was correct or wrong, shown in Figure 6. We found a significant effect of fact-check response on accuracy ($F(1{,}598) = 344.83$, $p < .01$), number of tasks done per session ($F(1{,}241) = 22.07$, $p < .01$) but not on completion time ($F(1{,}598) = .37$, $p = .51$).

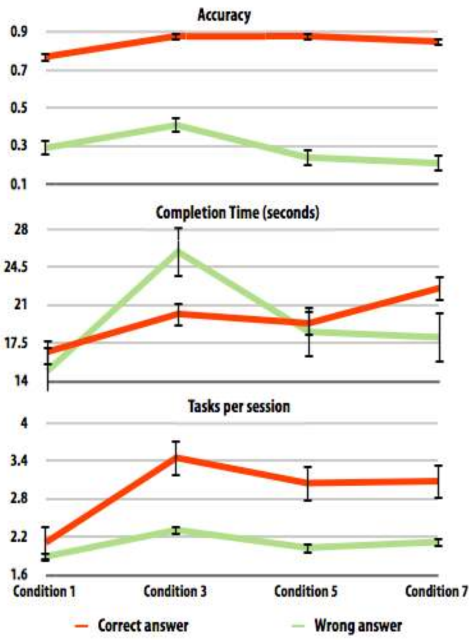


Figure 6. Breakdown of workers' performance based on the correctness of their answers to fact-check.

Comparison to Amazon's Mechanical Turk

Since we adopted a standardized task for our experiment, we are able to compare the performance of our participants in this study to the performance of workers on Amazon's Mechanical Turk reported in [27]. Because some participants on Mechanical Turk were rewarded, we report these results separately: participants who were not paid (0 cents per task), those who received 3 cents per task, and those who received 10 cents per task. Also, here we average all results from our own study into a single category to make visual comparison easier.

We compare the two datasets in terms of: accuracy (Figure 7) and rate of uptake of tasks (Figure 8). The first focuses on individuals' performance when completing tasks, while the rate of uptake is indicative of how quickly the tasks were completed, and how much time it takes to have a large number of tasks completed. We did not compare individual task completion times because workers on Amazon's Mechanical Turk also had to estimate the total number of cells in the images.

Figure 7 shows that accuracy results obtained on Mechanical Turk were higher. However, we note that this is for all public display conditions averaged, and certain public display conditions perform higher. In addition, the figure shows that workers on the public display are more likely to "give up" after a certain point of complexity.

Finally, Figure 8 shows that rate of uptake of tasks on the public display was much higher than on Mechanical Turk reaching 1200 tasks completed in 25 days compared to the non-paid version on Mechanical Turk that only reached 100 tasks completed in over 45 days.

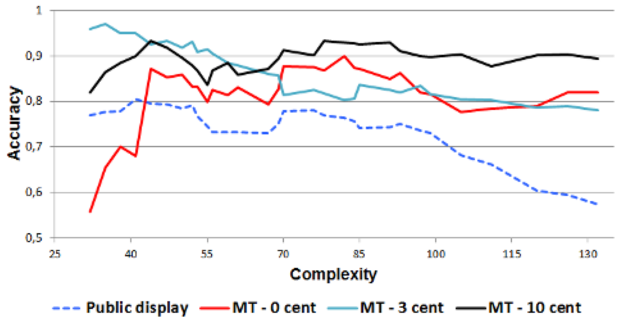


Figure 7. Comparison of public display and Amazon's Mechanical Turk accuracy across complexity.

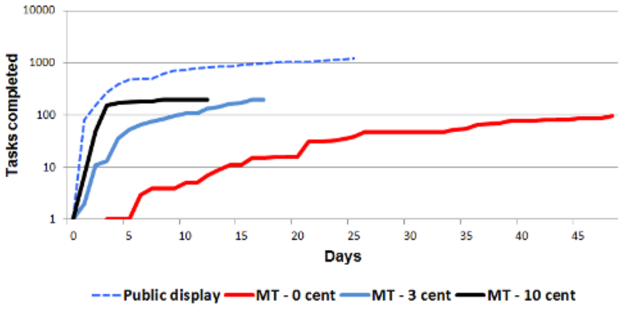


Figure 8. Comparison of public display and Amazon's Mechanical Turk rate of uptake of tasks.

In-situ Observations, Video and Interviews

We conducted 24 hours of direct observations during this study (6 hours per display). During this time we observed 149 physical interactions with the display in addition to other behaviours happening around the screen. Out of the 149 interactions, 52 (35%) involved people who approached the screen alone, while 97 (65%) involved groups of least 2 people. In total 77 (52%) interactions led to at least one task being completed. Interestingly, users that were alone had relatively more interactions that led to at least one task being completed (81% vs 36%) and also performing more than 1 task (21% vs 6%) therefore spending more time on the display. In some cases where groups approach the screen, each member would actively give opinions on what they believed was the correct answer resulting in mild arguments on what to input to the system.

However, in other instances of group usage they would input obviously wrong answers (e.g. "666", "999") in a joking manner. In addition, we observed passersby just touching screen and keep moving without stopping, while other simply overlooking the displays [23]. Our video observations included 123 instances of interaction. Our analysis confirmed instances of the behaviours that we initially noted in our in-situ observations, but also revealed several new behaviours that people exhibited when using the display (Figure 9).

In most cases we observed **ignorers:** passers-by walking by the display and completely ignoring it. This behaviour has often been referred to as display blindness, and has been reported in literature [23]. Often, however, we observed passers-by looking at the display but deciding not to interact with it. In certain cases, we observed **unlockers**: passersby walking past the display and in the process just clicking it once to see what happens. Our observations of these people's body language suggests that these people did not have an intention of actually using the display, because they do not slow down as they approach the display. Rather, they maintain a constant pace, and use the opportunity to touch the display once to "unlock" the screen.

Focusing on instances where crowdsourcing tasks were completed, we found that by far the most "active" workers were **loners**: individuals who use the display on their own, typically for longer periods of time, and who are not distracted by passersby. Often we observed that loners would terminate their activities when a person they were waiting for would come to find them, and they would walk away together.

In certain cases, loners became **attractors.** We observed individuals using the display, and then attracting others to join them on the display, also known as the 'honeypot' effect [4], and complete tasks jointly. Often we found that people who joined attractors left the display before attractors decided to leave. Our analysis suggests that this often caused a disturbance and delay in the completion of tasks, although in some of cases the attracted individuals did appear to be actively contributing to the tasks.

We also observed **repellers**: individuals who are applying social pressure on their peers to leave the display. For instance, we often observed a pair of people walking past the display, and one of them becoming interested in the display and completing a small number of tasks. In the meanwhile, the repeller would adopt a stance and body language that implied they were impatiently waiting for their friend to finish, often by standing in their field of view but in a position where they could not use the display. Interestingly, we also observed instances of **herders**: individuals who appear to lead a small group of friends. In these cases, we observed herders walking up to the display and using in a way suggesting they were demonstrating or explaining to their friends the interaction. In the meanwhile, the friends would adopt a passive position behind the herder, in a way that suggested they were not applying social pressure but rather observing. Ultimately, the herder would complete a small number of tasks and walk away with their friends following. Using the categories above, we labeled all interactions that we recorded on video. The relative frequency of the behaviour patterns that actively interacted with the display was: Loner 19%, Attractor 11%, Herder 6%, Repeller 14%, Unlocker 44%. The remaining 6% of interactions did not fit the description of the aforementioned behaviour patterns.

Finally, we conducted in total 24 semi-structured interviews (17 male, 7 female) across the 4 locations of the study. All participants claimed that they had interacted with public displays before. When asked why they decided to touch the screen their answers were mixed. Nine participants stated that curiosity was the main factor for them to touch the screen while another 9 said they were waiting for someone else or a class to start so they decided to kill some time. A further 6 participants said that the reason they approached the display was because friends had told them about it and they wanted to check it out.

Next, we enquired about their perceptions regarding the tasks shown. The majority (n=16) thought the idea was interesting and valuable completing at least one task. Some participants even directly referred to the motivational approach employed (i.e. building better software or helping medical research) showing that instructions influenced their behaviour in some way. Four participants noted that the reason they did not complete any tasks was the "fear" of answering incorrectly on such a sensitive task even though they knew it was for a good cause – something that would also hold in the case of web-based crowdsourcing. The remaining 4 participants either thought the tasks were too complex or that they did not feel like performing tasks. Interestingly, 3 of them were from either condition 1 or 2, i.e. with no motivation (Table 1).

DISCUSSION

Here we discuss i) the benefits and drawbacks of conducting crowdsourcing tasks on public displays, ii) how we can improve serious data collection on public displays, iii) users' behavioural patterns when crowdsourcing on public displays in natural usage settings, and iv) the potential for altruistic behaviour on public displays.

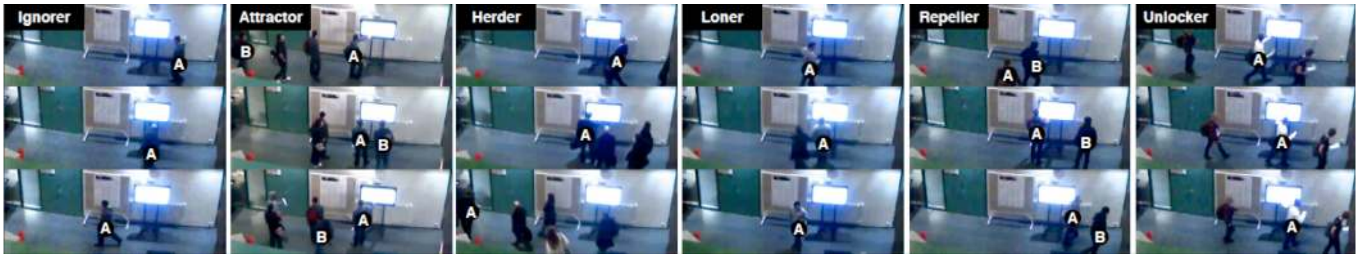


Figure 9. The different types of behaviour frequently observed around the display. Ignorer: This is the most typical scenario where a passerby completely ignores the display. Attractor: Person "A" starts using the screen, person "B" becomes attracted and approaches, and eventually "B" leaves while "A" remains on the display. Herder: "A" approaches and uses the display while a group observes him. Loner: "A" approaches and uses the screen for a relatively long period of time, while passersby ignore him. Repeller: "A" starts using the screen while "B" uses body language to apply social pressure to "A" to leave. Unlocker: "A" briefly interacts with the display without stopping his walk.

Benefits and Drawbacks

Our study demonstrates that crowdsourcing on public displays can produce comparable performance even to paid studies on Amazon's Mechanical Turk. But beyond task completion and accuracy rates, what other benefits or drawbacks does this medium bring to crowdsourcing studies?

One benefit that was made clear in our experiment was the much higher rate of task uptake on the public display (Figure 8). In the study by Rogstadius et al. [27] non-paid workers took over 45 days to complete 100 tasks while those were paid (3/10 cents) took over 10 and 15 days to complete 200 tasks respectively. On the public display it took us 25 days to complete 1200 tasks – without any monetary compensation given. This sharp difference can be mostly explained by the affordances of public displays. Two such affordances are the serendipitous and self-advertising nature of public displays [20]. After being exposed to a public display, passersby get attracted by public displays to complete tasks. This process itself helps raise the awareness of crowdsourcing among the local community acting as a self-renewable "workforce" by constantly attracting new workers. In a study by Gupta et al. [11] it became evident how other forms of crowdsourcing, in this case on mobiles, have difficulties in sustaining participation. While their participation started strong, when they reduced the payments the number of active users dropped 53% within a day. Most users who left reported that they could not invest more time and work for lower compensation, even though they were still getting paid.

Another key difference between online crowdsourcing markets and our deployment is the need to login. The login mechanism on Amazon's Mechanical Turk is a form of quality control that denies access to tasks for workers who perform poorly or attempt to cheat [26]. We believe that this additional barrier is not necessary on a public display as "bad" workers have no monetary incentive to lose time trying to cheat the system. In this case, potential workers could just approach the public display and start performing tasks right away, instead of going through an authentication mechanism that would most likely greatly diminish the amount of answers gathered.

Finally, Amazon's Mechanical Turk finds it challenging to recruit workers that speak a particular language or live in a particular city [15]. The strategic placement of public displays could help mitigate this issue by, for example, going directly to people that speak a specific language. Another example in which public displays could be used to improve crowdsourcing capabilities would be to target a specific audience with specialized skills that might be difficult to reach otherwise. For example by placing a medical crowdsourcing task (such as the one presented in this paper) on public displays located on a medical school campus it would be possible to reach users at the exact moment when they have free time to do the tasks.

Despite the various benefits of public displays for crowdsourcing, there are some serious drawbacks. For instance, the walk-up-and-use nature of public displays can result in limited usability and accessibility of tasks, with less rich interface controls than a standard desktop environment. This means that not all types of tasks can be crowdsourced on a public display. Another drawback is that the maintenance of public displays is more difficult than maintaining an online server and can incur higher initial costs.

Our results also suggest that beyond a certain level of task complexity workers on a public display will "give up", leading to a decline in performance (Figure 7). This threshold is lower on public displays than on Mechanical Turk, suggesting that very hard tasks probably cannot be reliably crowdsourced on public displays. Furthermore, it is not clear whether such crowdsourcing tasks could coexist with other content as previous work has shown that application discoverability plays a key role on multipurpose public displays [14, 19]. Finally, we note the limited multitasking capability on public displays in that it is not convenient to open another browser window to look for information.

Improving Data Gathering on Public Displays

Although technology deployments in authentic environments are always complex and practically impossible to replicate [5], our experiences highlight two issues to consider when collecting data with public interactive displays. Typically, censoring and moderation are laborious and costly challenges to tackle in such deployments [2]. Our study suggests that by carefully motivating the users, it is possible to improve the input of public display users. Workers who were motivated by *task* and *community based* identity [17] within the application description scored significantly higher accuracy in completing a task than workers with no motivational text present (Figure 4). Also, they spent more time on each task, suggesting that they were also more serious in their input (Figure 5).

Secondly, much of the related work tends to suggest that supporting group use on public displays is always beneficial, because it lowers the barrier to interaction [13]. In groups people feel less socially awkward; a common problem when using public displays [4]. However, our results suggest exactly the contrary as the right thing to encourage when collecting data from users: solo users, or *loners*, were clearly willing to spend more time with the system. We agree that in groups it is socially easier to approach and use a display, but we did observe more serious use of the system in the case of loners. This difference is partially explained by the social need of building "comradeship" or mutual trust by breaking rules or doing something that is against the accepted norms together [28]. In the case of playful applications or games, this does not matter and can even act as a catalyst to use, but when collecting meaningful data from the public it may be beneficial to attract more loners than groups.

Engaging Public Displays

In online and mobile crowdsourcing studies workers are treated as a black box in that we rarely get an opportunity to directly observe people completing crowdsourcing tasks. Here, we had the unique opportunity to observe participants' attitudes and social context when completing tasks. Understanding the reasons behind workers' altruistic behaviour and observing social dynamics around the display is crucial for the improvement of crowdsourcing in this medium.

In our study, we identified six categories of users. During our observations we observed that each category contributed differently to the amount of crowdsourcing tasks. Ignorers and unlockers did not complete any tasks, repellers prevented the completion of tasks, and herders demonstrated to others how to complete tasks. Loners completed the most tasks, with attractors coming second, despite sharing crowdsourcing tasks with the attracted friends.

Besides the need for breaking norms when in groups, we argue that the performance of loners can be explained by the absence of *peer pressure* [16]. Those that performed tasks alone did so on their own terms without being pressured by their friends to perform in a certain way or getting distracted. Another potential explanation is that these workers were not influenced by repellers, in others words, there was nobody else with them to actively try to get them away from the display. On the other hand, those that approached the screen in groups would sometimes, as seen in our observations, engage in *explicit subversive performances* [28] in which the primary user deliberately performed badly in an attempt to be humorous towards their friends.

Altruistic Behaviour on Public Displays

Whilst prior works investigate the hedonic use of public displays, this study demonstrates both the flexibility and the potential of public displays for altruistic services, such as crowdsourcing. While altruism should be enough of a motivator since it appeals to people's desire to help [26], we found that appropriate motivational and fact-checking mechanisms in the design are an important prerequisite for collecting accurate responses from users. Specifically, for people giving a correct response to the fact-check question after being exposed to motivational information, the system achieved an average accuracy of 88%, as shown in the Figure 3. This accuracy is comparable and even higher than that of workers on Amazon's Mechanical Turk – even those motivated by money (Figure 7).

Without a motivational approach and fact-check, altruistic crowdsourcing can only work if participants actually think the problem being solved is interesting and important [26] which is in most cases hard to achieve. For instance, when computer scientist Jim Gray went missing during a sailing trip in early 2007, thousands of online volunteers combed through over 560,000 satellite images (www.helpfindjim.com) covering nearly 3,500 square miles of ocean hoping to determine Gray's location. Sadly the effort was not successful, but the altruistic efforts of these volunteers nevertheless demonstrated that people will expend significant time and effort for the right cause [26].

On the other hand, "implicit" crowdsourcing work like reCAPTCHA [30] can be just as effective as altruistic crowdsourcing. reCAPTCHA is used online to verify that the user is human and at the same time provide benefit to a greater cause by helping digitise books, albeit without the user being aware of this. One other example of implicit crowdsourcing is Google Maps traffic information, provided by the millions of devices running Google Maps. As the devices move around a city, speed and location are crowdsourced to indicate road congestion [10].

We argue that public displays present themselves as ideal vehicles for both *altruistic* crowdsourcing and *implicit* crowdsourcing during their everyday use or with crowdsourcing tasks. One way to achieve the former would be through developing fact-checking questions for unlocking the public displays (just like smartphone lock screens), which would offer the dual benefit of demonstrating seriousness but also providing a crowdsourcing mechanism.

Limitations

We compared the performance of our participants in this study to the performance of workers on Amazon's Mechanical Turk reported in [27]. While the photographs used between the two studies were identical, there are a number of differences between the studies that we must highlight. First, the deployment settings are different. People visit crowdsourcing markets explicitly to complete tasks and to make money. In our case, individuals approached the displays in an impromptu fashion. Furthermore, in crowdsourcing markets tasks compete for the attention of workers, and therefore reward has an important effect on performance [27]. In our case, our display did not compete with other displays in terms of crowdsourcing, but it did compete with all other stimuli in the environment in terms of attention, including the fact that workers were often distracted by other people in the environment. Therefore we acknowledge the fact that our comparison of the two sets of results is not ideal, but nevertheless it is helpful in establishing the overall levels of performance we obtained. Finally, we note that that

CONCLUSION

This is the first crowdsourcing study on interactive public displays, where users could serendipitously perform tasks. During this study we demonstrate that it is indeed feasible to leverage this medium for such purposes, complementing rather than replacing existing crowdsourcing approaches. We argue that this is a clear step forward from the current crowdsourcing state-of-art.

In addition, we also demonstrate that it is possible to leverage public displays for serious data collecting, something that several researchers have struggled with. Further recommendations to these serious data collection mechanisms are given thanks to user behavioural patterns identified throughout our study.

Finally, we demonstrate that it is possible to promote altruistic behaviour on these displays and for crowdsourcing purposes. However this requires careful design with appropriate motivational approaches and worker quality signalling through verifiable questions.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland, TEKES, the Nokia Foundation, the Walter Ählstrom Foundation, the Tauno Tönning Foundation and Infotech Oulu.

REFERENCES

1. Alt, F., Shirazi, A., Schmidt, A., Kramer, U., Nawaz, Z. Location-based crowdsourcing: extending crowdsourcing to the real world. In *Proc.* NordiCHI '10, ACM (2010), 13-22.
2. Ananny, M., Strohecker, C. TexTales: Creating interactive forums with urban publics. In: Foth, M. (Ed.) *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*, IGI Global (2009), 68-86.

3. Arslan, M., Singh, I., Singh, S., Madhyastha, H., Sundaresan, K., Krishnamurthy, S. Computing while charging: building a distributed computing infrastructure using smartphones. In *Proc*. CoNEXT '12, ACM (2012), 193-204.
4. Brignull, H., Rogers, Y. Enticing People to Interact with Large Public Displays in Public Spaces. In *Proc*. INTERACT '03 (2003), 17-24.
5. Brown, B., Reeves, S., Sherwood, S. Into the wild: challenges and opportunities for field trial methods. In *Proc*. CHI'11, ACM (2011), 1657-1666.
6. Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivanstava, N. Participatory sensing. In *Proc*. WSW'06, ACM SENSYS (2006).
7. Churchill, E., Nelson, L., Denoue, L., Hefman, J., Murphy, P. Sharing Multimedia Content with Public Displays: A Case Study. In *Proc*. DIS'04, ACM (2004), 7-16.
8. Eagle, N. txteagle: Mobile Crowdsourcing. In *Proc*. IDGD '09, Springer-Verlag (2009), 447-456.
9. Goncalves, J., Kostakos, V., Karapanos, E., Barreto, M., Camacho, T., Tomasic, A., Zimmerman, J. Citizen Motivation on the Go: The Role of Psychological Empowerment. *Interacting with Computers*, online first (2013).
10. Google. The bright side of sitting in traffic: Crowdsourcing road congestion data. http://googleblog.blogspot.com/2009/08/bright-side-of-sitting-in-traffic.html (2009).
11. Gupta, A., Thies, W., Cutrell, E., Balakrishnan, R. mClerk: enabling mobile crowdsourcing in developing regions. In *Proc*. CHI '12, ACM (2012), 1843-1852.
12. Horton, J., Chilton, L. The labor economics of paid crowdsourcing. In *Proc*. EC '10, ACM (2010), 209-218.
13. Hosio, S., Kostakos, V., Kukka, H., Jurmu, M., Riekki, J., Ojala, T. From School Food to Skate Parks in a few Clicks: Using Public Displays to Bootstrap Civic Engagement of the Young. In *Proc*. Pervasive 2012, Springer (2012), 425-442 (2012).
14. Hosio, S., Goncalves, J., Kostakos, V. Application Discoverability on Multipurpose Public Displays: Popularity Comes at a Price. In *Proc*. PerDis '13, ACM (2013), 31-36.
15. Ipeirotis, N. Demographics of Mechanical Turk (2010), http://archive.nyu.edu/bitstream/2451/29585/2/CeDER-10-01.pdf
16. Kandel, E., Lazear, E. P. Peer pressure and partnerships. *Journal of Political Economy* (1992), 801-817.
17. Kaufmann, N., Schulze, T., Viet, D. More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. In *Proc*. AMCIS'11 (2011).
18. Kittur, A., Chi, E., Suh, B. Crowdsourcing user studies with Mechanical Turk. In *Proc*. CHI '08. ACM (2008), 453-456.
19. Kostakos, V., Kukka, H., Goncalves, J., Tselios, N., Ojala, T. Multipurpose public displays: How shortcut menus affect usage. IEEE Computer Graphics and Applications 33, 2 (2013), 50-57.
20. Kukka, H., Kostakos, V., Ojala, T., Ylipulli, J., Suopajarvi, T., Jurmu, M., Hosio, S. This Is Not Classified: Everyday Information Seeking and Encountering in Smart Urban Spaces. *Personal and Ubiquitous Computing* 17, 1 (2013), 15-27.
21. Kukka, H., Oja, H., Kostakos, V., Goncalves, J., Ojala, T. What Makes You Click: Exploring Visual Signals to Entice Interaction on Public Displays. In *Proc*. CHI'13, ACM (2013), 1699-1708.
22. Mashhadi, A., Capra, L. Quality control for real-time ubiquitous crowdsourcing. In *Proc*. UbiCrowd'11, ACM (2011), 5-8.
23. Müller, J., Wilmsmann, D., Exeler, J., Buzeck, M., Schmidt, A., Jay, T., Krüger, A. Display Blindness: The Effect of Expectations on Attention towards Digital Signage. In *Proc*. Pervasive 2009, Springer (2009), 1-8.
24. Narula, P., Gutheim, P., Rolnitzky, D., Kulkarni, A., and Hartmann, B. MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid. In *Proc*. HCOMP '11 (2011).
25. Paulos, E., Honicky, R. J., Hooker, B. Citizen Science: Enabling participatory urbanism. *Handbook of Research on Urban Informatics*, IGI Global (2009), 414-436.
26. Quinn, A., Bederson, B. Human computation: a survey and taxonomy of a growing field. In *Proc*. CHI '11, ACM (2011), 1403-1412.
27. Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. Vukovic, M. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *Proc*. ICWSM'11, AAAI (2011), 321-328.
28. Schwarz, O. Subjectual Visibility and the Negotiated Panopticon: on the Visibility-Economy of Online Digital Photography (2011).
29. Väätäjä, H., Vainio, T., Sirkkunen, E., Salo, K. Crowdsourced news reporting: supporting news content creation with mobile phones. In *Proc*. MobileHCI '11, ACM (2011), 435-444.
30. von Ahn, L., Maurer, B., McMillen, C. Abraham, D., Blum, M. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science 321, 5895 (2008), 1465-1468.
31. Vukovic, M., Kumara, S., Greenshpan, O. Ubiquitous crowdsourcing. In *Adj. Proc*. Ubicomp '10, ACM (2010), 523-526.