No “One-Size Fits All”

Towards a principled approach for incentives in mobile crowdsourcing

John P. Rula* Vishnu Navda† Fabián E. Bustamante*
Ranjita Bhagwan† Saikat Guha†
*Northwestern University †Microsoft Research India

ABSTRACT

We are becoming increasingly aware that the effectiveness of mobile crowdsourcing systems critically depends on the whims of their human participants, impacting everything from participant engagement to their compliance with the crowdsourced tasks.

In response, a number of such systems have started to incorporate different incentive features aimed at a wide range of goals that span from improving participation levels, to extending the systems' coverage, and enhancing the quality of the collected data. Despite the many related efforts, the inclusion of incentives in crowdsourced systems has so far been mostly ad-hoc, treating incentives as a wild-card response fitted for any occasion and goal.

Using data from a large, 2-day experiment with 96 participants at a corporate conference, we present an analysis of the impact of two incentive structures on the recruitment, compliance and user effort of a basic mobile crowdsourced service. We build on these preliminary results to argue for a principled approach for selecting incentive and incentive structures to match the variety of requirements of mobile crowdsourcing applications and discuss key issues in working toward that goal.

1. INTRODUCTION

Mobile crowdsourcing leverage the pervasiveness of smartphones and other resource-rich devices to create a wide range of services from community sensing [7, 17, 20], to wireless network characterization [9, 10] and micro-task markets [5, 24].

Despite the rich potential of mobile crowdsourcing services, we are increasingly aware that their effectiveness critically depends on the whims of their participants, impacting everything from user engagement to their compliance with the crowdsourced tasks.

In response, a number of such systems have started to incorporate different incentive features including various forms of rewards, such as monetary, social and gaming-related, and incentive structures (e.g., lottery, per-time or per-task payments [15, 25]). Not surprisingly, the motivation behind these attempts varies as much as their services' goals, from improving participation levels, to extending the systems' coverage, and enhancing the quality of the collected data. While a campaign characterizing noise pollution levels [17], for instance, would be interested in large recruitments with little individual contribution, an application for tracking litter [3] could achieve its goals with only a few motivated users.

Our work is motivated by the observation that the inclusion of incentives in mobile crowdsourced systems has been, so far, mostly ad-hoc. In each new project, researchers explicitly or implicitly recognize the potential impact of participant engagement and embed some form of incentive, sometimes modeled on past efforts, generally treating incentives as wild-card solution appropriate whatever the occasion or goal. We believe that there is not a magic bullet when it comes to incentives.

In this paper we present results from a large scale experiment, run over 2 days with 96 participants, to compare the relative effects of two different incentive structures – *micro-payments* and *weighted lottery* (§ 4). We find that even with the same incentive type (money), the different structures induce significantly different behavior from participants. With respect to user recruitment, for instance, weighted lottery performed much better than micro-payments – a 46% improvement. Looking at task compliance on a per-user basis, however, the median number of tasks completed under micro-payment was twice that completed under weighted lottery. Building on these preliminary results, we argue the need for a principled approach for applying incentives in mobile crowdsourcing to best match incentive features to applications' requirements (§ 7).

The remainder of this paper is organized as follows. In the following section, we briefly discuss some relevant background on crowdsourced measurements and incentives. Section 3 describes our approach to experimentally evaluate the impact of incentive structures on mobile crowdsourcing. We describe our experiments and evaluation methodology in Section 4 and discuss our results in Section 5. In Section 6, we discuss some of our findings and early ideas toward a principled approach to incentive selection. We conclude in Section 8.

2. BACKGROUND

The effect of incentives has been the focus of research efforts in several domains, from economics and business management to public health, psychology and behavioral studies. Within computer science, the topic has been studied in the contexts of networks, time sharing, peer-to-peer systems and participatory sensing, among others [6, 11, 15, 26].

Incentives can be described along three different axes: type, magnitude and structure. *Incentive types* refer to the class of reward offered to the user and can be monetary, altruistic, social or game-related [13]. Incentive types can be further subdivided into *intrinsic* motivations, those which satisfy the individual such as enjoyment,

ACM HotMobile'14, February 26–27, 2014, Santa Barbara, CA, USA.

social interactions or helping others, and *extrinsic* motivations, those which originate externally including monetary payments (immediate or delayed) or perceived social pressures [19]. In a previous effort looking at the effects of incentive types, Shaw et al. compared different incentives and combination of incentive in the context of micro-task markets and find that some task framings may elicit higher quality performance than others [23].

Incentive magnitude refers to the amount or degree of each reward offered. A common example of incentive magnitudes is the different wages, or monetary compensation for labor. Numerous works in economics and psychology have explored the response to different levels of payment for work [2,12,14,22]. Within the realm of crowdsourcing, several efforts have looked at effectively pricing simple micro-tasks [8,18] and the effect of differential prices in this environment [16]. In mobile crowdsourcing environments, Reddy et al. [21] compares different denominations of micro-payments within participatory sensing campaigns.

Last, *incentive structure* represents the set of rules which govern how different incentives are distributed to users, including any performance metrics used in compensation, and the risk associated with compensation [2]. In incentive structures known as tournaments, uncertainty is introduced into the compensation structure. Other more performance driven structures have a guaranteed payment related to the participant's performance. The latter type is common for crowdsourcing environments such as Amazon Mechanical Turk [1] and GigWalk [5].

3. INCENTIVES & CROWDSOURCING

In trying to understand the relative effect of different incentive mechanisms in mobile crowdsource services, we run a 2-day experiment at a corporate conference using a collection of instrumented mobile devices.

We identify three key performance metrics of interest to mobile crowdsourcing: recruitment, compliance and user-effort. *Recruitment* tries to captures the ability to attract participants. While most crowdsourcing applications could benefit from greater recruitment levels, particular tasks which rely on passive or background measurements such as opportunistic urban sensing critically depend on the number of active participants.

Compliance describes the extent to which a participant carries out the required task. In open task markets typical in crowdsourcing, users can undertake a task they never complete or complete "badly" (e.g. a lower quality noise sample or a sample at the wrong location).

Our third metric, *user effort*, tries to capture the amount of energy invested by each participant during the crowdsourcing activity. We would like to differentiate, for instance, applications that require minimal or no participant effort from those requiring a number of consecutive measurements or micro-tasks that involve non-trivial travel distances.

4. EXPERIMENTAL EVALUATION

Focusing on these metrics, we conducted preliminary experiments to evaluate the relative impact of two different incentive structures: micro-payments and weighted lottery.

Our experiments were run over two days in March, 2013 at the Microsoft Research Technical Conference in the Redmond (USA) campus. We use a collection of 50 Windows Mobile devices loaded with a virtual scavenger hunt game and we distributed them among the conference attendees. The hall where the conference was held included 151 individual demo booths of different research projects from the lab. The hall was in operation from 12pm to 5pm, each day, for a total of 10 hours. As mentioned, our experiment participants consisted of conference attendees who volunteered to check out a mobile device in exchange for the incentive structure offered that day.

In our virtual scavenger hunt game, *Tech Hunt*, users are given tasks which contained clues corresponding to one of the demo booths at the conference. When a user feels they have correctly decoded the clue, they proceed to the associated demo booth and scan the 2-D bar code through our application. If successful, the user is awarded the allotted points.

From the set of 151 demo booths at the conference, we randomly chose a subset of 51 to use as part of our experiments. For each of the booths, we generated an easily identifiable "clue" referring to it. An example of these would be "Are your fists clenched? Kinect knows.", which referred to the demo booth about hand gestures with Microsoft Kinect. These were designed to be easily linked to their respective booth based on the title of the demo booth in the conference program.

Each user running our application was given a random subset of 10 unique tasks selected and displayed within the application. The tasks were presented to the user as a list of clues. A user was only given a total of 10 tasks, and, once completed these tasks were not replenished. In addition to displaying the requested tasks, the application supported the QR code scan and presented a leaderboard showing scores of other participants playing that day.

Participants were presented with a different incentive structure each day: micro-payments and weighted lottery. With *Micro-Payments* (MP), participants were compensated for each set of 5 tasks they completed. For every set completed by a user, a $5.00 gift card to a national coffee chain was given. In addition, players were given a $5.00 gift card for checking a device out for more than an hour, for a maximum compensation of $15.00. With *Weighted Lottery* (WL) participants received a raffle ticket for every 5 tasks they completed, as well as a ticket for checking out a device for more than an hour. At the end of the day, tickets were drawn with winners receiving a $50.00 gift card to same national coffee chain. Twenty winners were chosen from the set of participants, with each participant only able to win once.

The data for our analysis comes from the mobile devices themselves, as well as an anonymized listing of badge swipes from the conference. The mobile devices were loaded with our virtual scavenger hunt as well as a WiFi location daemon. The badge swipes came from RFID readers at each booth which recorded employee id and booth number and were used to measured booth popularity and conference dynamics. From the booth swipes, we observed 1596 unique users on the first day of the conference where micro-payments were tested and 2011 unique users on the second day of the conference where weighted lottery was tested.

5. RESULTS

In this section, we analyze our experiment results for the two incentive structures in terms of the introduced metrics: *recruitment*, *compliance* and *user effort*.

5.1 Recruitment

We define recruitment as the number of users who choose to participate in the mobile crowdsourcing application. High levels of recruitment are critical to the success of most mobile crowdsourcing applications, particularly those which do not involve explicit user interaction such as opportunistic mobile sensing applications. In our experiments, we differentiate between the number of participants who checked out phones from our booth, *total recruitment*,

	Micro-payments (MP)	**Weighted Lottery** (WL)
Participants	39	57
Active	23	39
Completed	99	120
User Mean	4.3	3.24
User Median	4	2
User Variance	6.65	6.61

Table 1: Summary of user performance between incentive structures. The weighted lottery was able to recruit more users, but with a lower average output than micro-payments.

and those who actively contributed by completing at least a single task, *active recruitment*.

We observed weighted lottery achieving greater total recruitment, as well as larger active recruitment than micro-payments a 46% increase and 19% increase respectively. Over the course of our two day experiments, we had 96 users who checked out phones from us: 39 with MP and 57 with WL. Of those 96 participants, 62 completed at least a single task within our game: 23 with MP and 39 with WL, showing a participation rate of 57% for the MP to 68% for WL. In order to compensate for the greater numbers of visitors seen on the second day of the conference, we normalize these recruitment levels across the number of unique users seen each day to generate a recruitment rate as the number of recruited participants over the total unique visitors seen each day. When we compare this recruitment rate, 0.024 for MP and 0.028 for WL, we still see a 16% increase in total recruitment with WL over MP.

Recruitment levels, however, do not seem to explain the whole story of incentive selection. While we saw more combined recruitment from WL, we found that recruitment did not necessarily coincide with higher compliance rates of participants.

5.2 Compliance

We measure compliance by the total number of tasks completed by a participant. We saw differences in the output of users between the two incentive structures. Table 1 presents the summary of these results. Compliance rates were higher on average from MP compared to WL, 0.43 to 0.32. In addition, the median number of tasks completed per user for WL was half that of the MP.

The compliance of tasks within mobile crowdsourced services goes beyond the total compliance rate, and also depends on the value of the tasks completed. In mobile crowdsourcing, a major issue is obtaining compliance in areas with typically low user coverage. Within our experiments, we used the popularity of conference booths as a substitute for spatial coverage density in mobile crowdsourcing environments.

Within the convention center, there were booths and areas that received more traffic than others due to the presence of other popular booths or high trafficked corridors. Figure 1 shows the distribution of badge swipes at each booth combined across both days, normalized from the maximum number of visits at any booth. There was a clear distribution between popular booths, those with high traffic, and unpopular booths with low traffic. The figure shows a noticeable distinction between the set of unpopular booths as those which have a normalized visit count of 0.2 or less.

To analyze the ability of each incentive to drive users to unpopular booths, we calculated the correlation of tasks compliance with regards to each booth's popularity. We find no correlation between booth popularity and task compliance, with MP and WL having Pearson correlation coefficients of 0.048 and 0.002 respectively, indicating no correlation between these sets [4].

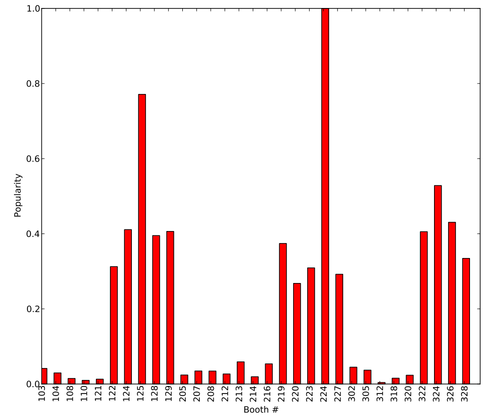


Figure 1: Booth visits from all conference goers normalized against the maximum booth visits. A clear distinction exists popular and unpopular booths at 0.2

5.3 User Effort

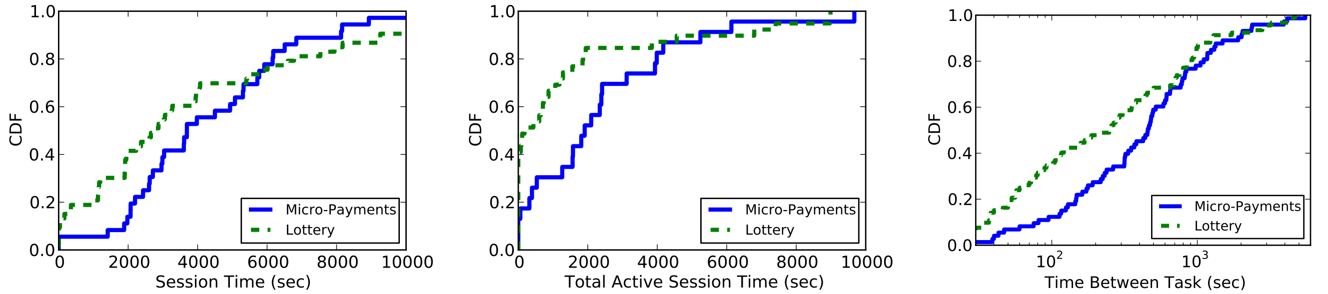
User effort refers to the amount of interaction spent by each user within the mobile crowdsourcing system, and is useful for applications which require user interaction, or applications which need multiple measurements from the same user. We calculate user effort from our participants as the total amount of time spent with the checked-out device, the amount of time spent actively using the device, and the area covered by each user.

Total check-out time is measured as the time from the user checking out a device from our booth to the user returning the device to us, measuring physical possession of the device. This metric can be used to determine a user's passive availability, and would be applicable to many background and opportunistic sensing applications. Figure 2a shows a cumulative distribution of check-out times for both micro-payments and lottery. The figure shows that micro-payments had a longer check-out time up until the 80th percentile, where lottery participants had significantly longer checkout times.

The *active* time spent within the application is measured as the time between the first and last completed task for a given user. This metric is important because it gives the time window where a user is engaged and producing within the crowdsourcing application, and can additionally gauge a user's temporal coverage window. Figure 2b shows a cumulative distribution of active user sessions for both incentive structures. Micro-payments significantly out perform lottery for almost all participating users with regards to active session times.

Similarly, we measured the time between consecutive tasks to look at the evenness of temporal coverage of participants. Figure 2c shows the time between completed tasks for users given micro-payments and lottery incentive mechanisms. The figure shows that nearly 40% of users within the lottery mechanism completed tasks within one hundred seconds of each other, confirming the results from the previous two session metrics.

We can also view user effort by looking at the aggregated mobility traces of users across the two days. Figure 3 shows a heatmap of mobility for each incentive mechanism collected from our WiFi localization service running on each device. From the figure we can observe the larger area covered and with greater density by participants given MP than WL. These participants spent



(a) Device checkout times.

(b) Active user sessions.

(c) Time between completed tasks for users.

Figure 2: User effort statistics for participants shown for both incentive structures. In all metrics of user effort, micro-payments out perform weighted lottery for total session time, active session time, and inter-task time.

more time traversing the conference area than those of the lottery, and were able to accomplish this with 46% fewer users than the weighted lottery.

6. DISCUSSION

From our experiments, we observed larger overall recruitment along with lower average output of users with weighted lottery compared to micro-payments. Users appeared to be motivated by the higher potential payouts to join the lottery pool, however, completing more than 2 missions (the median) did not seem worth the effort based on user's perception of probability of success. An analysis of the expected payouts for WL show that checking out a phone gave a user a probability of winning of 0.28, thus an expected payout of $14.08, and a user completing 5 tasks would double their expected payout to $28.16. Given that only 12 participants completed at least 5 tasks, it seems the initial expected payment from merely checking out a device was sufficient given the minimal effort required for it.

The micro-payments, while attracting fewer users, had more productive participants. The payout from micro-payments required more time (more task completions), yet had a guaranteed rate of return. This is surprising considering the large disparity between expected payments between the two incentive structures. While completing all tasks in micro-payments would pay $15.00, completing the same amount under our weighted lottery would yield an expected payment of $42.24.

While we attempted to estimate equal expected payouts for both incentive structures, the our experiments clearly favored the weighted lottery with its much larger expected payments for the same effort. However, if we look at user performance as a return on investment, micro-payments greatly outperformed the weighted lottery. Even without considering the fewer overall visitors on the first day of the conference, micro-payments only suffered an 18% reduction in *total* production while only spending a quarter the amount of the lottery ($255 vs. $1000).

We observed that both incentive mechanisms provided sufficient motivation for participants to visit booths with low popularity. While indeed the energy needed to visit an unpopular booth was low in the experiment context, the incentive mechanisms still caused users to visit booths in unpopular areas.

Our participants were not randomly selected, instead all of our participants volunteered to check-out a device, thus accepting the incentive structure offered. While it is possible that this lead to biases in our participant base, we believe that our experiment setup most closely matches the conditions in which mobile crowdsourcing applications exist. Most crowdsourcing participants are themselves volunteers, therefore, our experiments are comparing those same users who would be involved in crowdsourcing elsewhere.

7. TOWARD A PRINCIPLED APPROACH

We build on our experimental results to make the case for a principled approach for applying incentives in mobile crowdsourcing. We believe the choice of incentive type and structure should depend on the requirements of the application (e.g., coverage, required effort) as well as constraints on the publisher themselves (e.g., budget, technical capability).

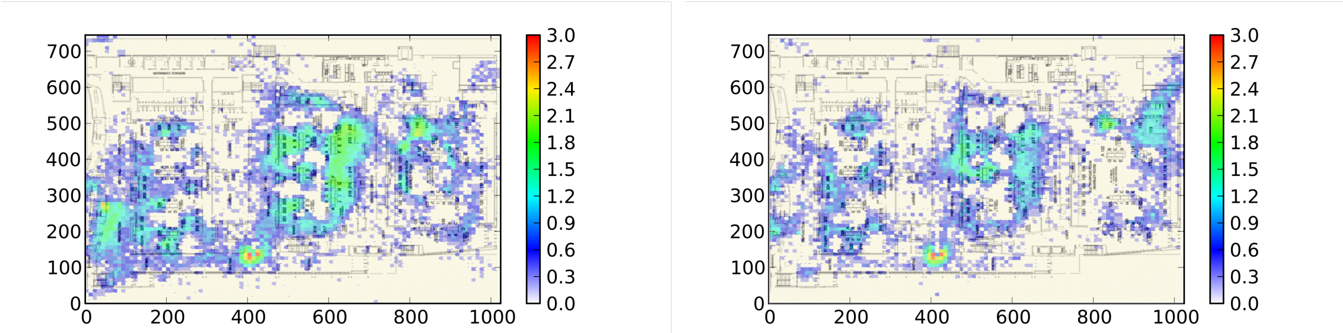
Our preliminary results showed different strengths and weaknesses for each of the evaluated incentive structures, some of which would be more suitable for different crowdsourcing applications. A principled approach would incorporate both the incentive type and magnitude along with the structures themselves, and allow application designers to match the appropriate incentive to their particular application.

Clearly the "right" incentive would also depend on the capabilities and constraints of the mobile crowdsourcing publisher. For instance, a publisher which is financially constrained would not be able to offer substantial financial incentives, and instead may opt for entertainment based incentives through gamification of their crowdsourcing campaign. Within each gamification element, the structures described before have the potential to offer the same benefits as they would for monetary incentives.

8. CONCLUSION

The effectiveness of mobile crowdsourcing systems critically depends on the whims of their human participants, impacting everything from user engagement to their compliance with the crowdsourced tasks. In response, a number of mobile crowdsourcing systems have started to incorporate different incentive features aimed at a wide range of goals depending on the particular application or service. In all past efforts, however the inclusion of incentives has been mostly ad-hoc, treating incentives as a wildcard response fitted for any occasion and goal.

We presented results from a large, 2-day experiment with 96 participants, comparing the effect of two difference incentive structures on recruitment, compliance and user effort on the assigned task. Our preliminary results show the different impact each incentive structure has along the three observed metrics of performance. Building on these observations, we argue for a principled approach for incentive selection in crowdsourcing environments.



(a) Micro-payments.

(b) Lottery.

Figure 3: Heatmap of mobility traces aggregated over all our participants over the two day period. The yellow and red colors indicate levels of high traffic centered around the experiment station (lower center), and several popular demo booths.

In the immediate future, we plan to expand the list of application requirements evaluated as well as the number of incentive mechanisms compared, as we work toward defining the basis for a principled approach to incentives in mobile crowdsourcing.

9. REFERENCES

[1] Amazon. Amazon mechanical turk. `https://www.mturk.com/mturk`.

[2] G. P. Baker, M. C. Jensen, and K. J. Murphy. Compensation and incentives: Practice vs. theory. *The Journal of Finance*, 43(3):pp. 593–616, 1988.

[3] U. center for Embedded Networked Systems. Participatory sensing / urban sensing projects. `http://urban.cens.ucla.edu/projects/garbagewatch/`, January 2011.

[4] J. Cohen. *Statistical Power Analysis for the Behavioral Sciencies*. Lawrence Erlbaum Associates, Incorporated, 1988.

[5] Gigwalk. Gigwalk: Hire your smartphone army. `http://www.gigwalk.com`.

[6] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang. Tube: time-dependent pricing for mobile data. In *Proc. ACM SIGCOMM*, 2012.

[7] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson. Virtual trip lines for distributed privacy-preserving traffic monitoring. In *Proc. of MobiSys*, 2008.

[8] J. J. Horton and L. B. Chilton. The labor economics of paid crowdsourcing. In *Proc. of EC*, 2010.

[9] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. A close examination of performance and power characteristics of 4g lte networks. In *Proc. of MobiSys*, 2012.

[10] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl. Anatomizing application performance differences on smartphones. In *Proc. of MobiSys*, 2010.

[11] S. Jun and M. Ahamad. Incentives in bittorrent induce free riding. In *Proc. of P2PECON*, 2005.

[12] E. Kamenica. Behavioral economics and psychology of incentives. *Annu. Rev. Econ.*, 4(1):427–452, 2012.

[13] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing: a study on mechanical turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems*, pages 1–11, 2011.

[14] E. P. Lazear. Performance pay and productivity. Technical report, National Bureau of Economic Research, 1996.

[15] J.-S. Lee and B. Hoh. Dynamic pricing incentive for participatory sensing. *Journal of Pervasive and Mobile Computing*, 6:Pages 693–708, 12/2010 2010.

[16] T. Y. Lee, C. Dugan, W. Geyer, T. Ratchford, J. Rasmussen, N. S. Shami, and S. Lupushor. Experiments on motivational feedback for crowdsourced workers. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[17] N. Maisonneuve, M. Stevens, M. Niessen, and L. Steels. Noisetube: Measuring and mapping noise pollution with mobile phones. *Information Technologies in Environmental Engineering*, pages 215–228, 2009.

[18] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proc. HCOMP*, 2009.

[19] E. Massung, D. Coyle, K. F. Cater, M. Jay, and C. Preist. Using crowdsourcing to support pro-environmental community activism. In *Proc. of CHI*, 2013.

[20] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proc. of MobiSys*, 2009.

[21] S. Reddy, D. Estrin, M. Hansen, and M. Srivastava. Examining micro-payments for participatory sensing data collections. 2010.

[22] S. A. Ross. Compensation, incentives, and the duality of risk aversion and riskiness. *The Journal of Finance*, 59(1):207–225, 2004.

[23] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *In Proc. CSCW*, 2011.

[24] TaskRabbit. Taskrabbit | your tasks, done. `http://www.taskrabbit.com`.

[25] T. Yan, V. Kumar, and D. Ganesan. Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In *Proc. of MobiSys*, 2010.

[26] D. Yang, G. Xue, X. Fang, and J. Tang. Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. In *Proc. of MobiCom*, 2012.