Evaluating Participation and Performance in Participatory Sensing

Sasank Reddy, Katie Shilton, Jeff Burke, Deborah Estrin, Mark Hansen, Mani Srivastava
Center for Embedded Networked Sensing, University of California, Los Angeles (UCLA)
Email: {sasank, kshilton, jburke, mbs}@ucla.edu, destrin@cs.ucla.edu, cocteau@stat.ucla.edu

Abstract

Because participatory sensing – targeted campaigns where people harness mobile phones as tools for data collection – involves large and distributed groups of people, participatory sensing systems benefit from tools to measure and evaluate the contributions of individual participants. This paper develops a set of metrics to help participatory sensing organizers determine individual participants' fit with any given sensing project, and describes experiments evaluating the resulting reputation system.

I. Introduction

The rapid adoption of mobile phones over the last decade and an increasing ability to capture, classify, and transmit a wide variety of data (image, audio, and location) have enabled a new sensing paradigm – participatory urban sensing – where humans carrying mobile phones act as, and contribute to, sensing systems [1], [2], [3]. In this paper, we discuss an important factor in participatory sensing systems: measurement and evaluation of participation and performance during sensing projects.

In participatory sensing, mobile phone-based data gathering is coordinated across a potentially large number of participants over wide spans of space and time. We draw from three pilot projects to illustrate participatory sensing and describe the unique challenges to measurement and evaluation provided by "campaigns": distributed and targeted efforts to collect data. Project Budburst [4], Personal Environmental Impact Report (PEIR) [5], and Walkability all situate "humans in the loop", but have critical differences in their goals and challenges (Table I).

Campaign	Goal	Data Collection	Evaluation Challenges
Project Budburst	Gather data about the flowering of native plants to study climate change	Mobile phones upload time-stamped, geo-tagged plant photos	Ensuring data quality to meet scientific standards, including photo resolution and reliable capture of events of interest
PEIR	Enable individuals to collect data about environmental impact and exposure	GPS traces create estimates of carbon emissions/pollution exposure	Continual and long-term participation yields most meaningful results
Walkability	Gather data about neighborhood sidewalk hazards	Mobile phones upload geo-tagged photos of cracks, gaps, and impediments	Systematic coverage and verification to make a strong case to city councils

TABLE I
Goal, data collection, and evaluation challenges associated with campaigns.

These pilots have raised complex problems of reputation and reliability. "Human in the loop" sensing relies upon community expertise, drawing upon mass contributions in much the same way as web-based systems such as Wikipedia and Slashdot [6]. But these projects also demand data quality that meets criteria set by scientists or legislators. Participatory sensing reputation metrics must therefore incorporate expertise, data quality, credibility, and certainty, while encouraging participation and development of expertise among amateur volunteers.

To address this challenge, we have developed two indexes for measuring participation and expertise among campaign participants: cross-campaign and campaign-specific measurements. Cross-campaign metrics are résumé-like measures of previous experience and commitment. These include number of previous campaigns undertaken and the success of a participant in previous campaigns. Experience metrics can play a key role alongside other factors such as sensing modality, coverage, and cost in enabling campaign monitoring services to select participants who can achieve the highest effectiveness for a particular data collection initiative. Tracking participation in this way is not new; it has been employed widely by Internet businesses and services [7]. Systems that provide a marketplace for commissioned work, such as Amazon Mechanical Turk and GURU.com, keep detailed statistics tracking the performance of requesters and workers [8], [9]. Systems for question answering, such as Amazon Askville and Yahoo Answers, use credits to track a participant's performance [10], [11], and auctioning systems such as E-Bay have transaction ratings to help evaluate whether a particular participant is trustworthy [7].

Campaign-specific metrics provide project organizers with something existing systems do not offer: evaluation of participation during the deployment of a campaign over weeks or months. Instead of the résumé-like measures used by existing reputation systems, we compare campaign-specific metrics to a project review. As a campaign progresses, a "watchdog" module can observe quality and utility of a participant's contribution relative to campaign needs. The module can then send alerts and recommendations to participants, dynamically change incentives for existing participants, and recruit new participants to the campaign.

II. Cross-Campaign Participation: The résumé

Our work builds on ideas of monitoring participant behavior and provides metrics to evaluate participation and performance in sensing campaigns. To suggest metrics useful to filter or rank potential participants based on past performance, we consider the requirements volunteers must meet before joining a campaign and how the metrics that make up these requirements are delineated. Cross-campaign metrics can relate to either campaign participation or campaign performance.

A. Campaign Participation

Participation requirements allow a campaign organizer to recruit participants that have a certain level of experience or have been active recently. Participation metrics include: a) number of campaigns volunteered for, b) number of campaigns accepted for, c) number of campaigns participated in, and d) number of campaigns abandoned. Individual metrics can be associated with other information about a campaign, such as size, lifetime, and type of sensing required. Examples of requirements could include: participants who have been accepted for image-sensing campaigns in the last 6 months, have participated in a certain number of location-sensing campaigns, or have less than 10% abandonment rate.

B. Campaign Performance

To create metrics that represent a participant's performance, campaign organizers need a language to express the campaign contributions they require. A successful contribution can be defined according to a number of qualities, including: what sensor type should be used and what modalities employed; the spatial or temporal context in which the sample is taken; and timeliness, relevance and quality of the sampled data. Timeliness represents the latency between when a phenomenon is sampled (or occurs) and when it is available to a data processing module. Relevancy indicates how well the sample describes the phenomenon of interest ("did the participant photograph a flower?") Quality describes the ability of the system to determine a particular feature in a sample ("can the system detect a sidewalk hazard in this photograph?") Quality includes the probability of detection, probability of a false positive, or probability of a false negative. Campaign performance may also include metrics that describe the responsiveness of a participant, or similarly, the amount of responsibility taken by a participant. Campaign organizers might consider how frequently a participant checks in with the system, whether a participant uploads regularly, or whether the participant takes privacy precautions with their data such as blurring third party images in photographs. Because the meaning and importance of each of these variables can vary based upon the needs of the campaign, it will be important for campaign organizers to have control over setting definitions and levels of cross-campaign metrics.

Using the above metrics, campaign organizers can define a useful campaign entry and set benchmarks that indicate participant success. Performance benchmarks can be absolute, set limits for performance categories, or relative, based on performance of other members. Also, the scales can be translated into user-friendly forms for querying such as the 5-star system popular on many Internet platforms [12]. As future work, we will consider enabling participants to set their own benchmarks for success.

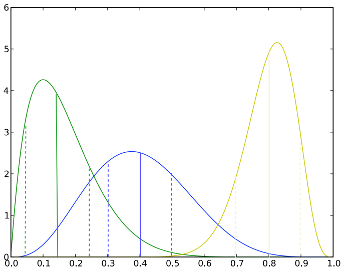
III. Campaign-Specific Participation: The Project Review

Analysis of pilot campaign participation has shown that it is important to monitor participant contributions while a campaign is running, as well. Participants may realize during the course of a campaign that data collection does not fit their schedule or interests, or organizers may discover that participants are not keeping up with their data collection duties. Campaign-specific monitoring enables a monitoring service to adapt the participant list, coordinate reliable contributors to collect or verify information, or adapt feedback and incentive mechanisms.

To generate campaign-specific participation and performance measures, campaign organizers could choose several mechanisms. Campaigns could incorporate a "calibration" phase paired with reoccurring "check-ups" where experts or campaign organizers obtain ground truth information for a particular area of interest. Participants would then be coordinated to monitor the same area. The observations made by participants could then be compared to the ground truth to obtain a reliability measure. For campaigns where obtaining ground truth information is not practical or possible, indirect "collective" observations made by participants can generate a reliability score. In this case, geographic coverage overlaps between participants would be found, and reliability scores can be calculated by measuring the similarity of observations in these areas. Evaluating participant reliability by comparing overlapping results is similar to the social science practice of calculating interrater reliability: consistency among responses when assigning values to subjective data [13].

We propose a mathematical model to represent campaign-specific performance, update it based on measures of participant reliability, and translate it into a participant trust metric. Existing reputation systems used in applications include: cumulative, where a user's reputation ratings are summed; average, where the reputation score is computed by averaging all scores; blurred, where a weighted sum is used to down weight old ratings; and adaptive, where the current reputation score affects to what degree new observations make a difference [14].

The above mentioned reputation systems only capture stochastic uncertainty (due to randomness of the system). We instead want a reputation system that captures both stochastic and epistemic uncertainty (due to lack of knowledge about the randomness of the system) so we adopt the Beta distribution to model the reputation of a participant. The distribution is indexed by two parameters, alpha and beta, which define the number of successful and unsuccessful transactions that have occurred in the past. A participant's reputation can be found by calculating the expectation of the Beta distribution, E(alpha,beta) which is the stochastic uncertainty. The confidence factor is the posterior probability given the actual expectation value lies within an acceptable level of error [15]. Prior to any interaction, alpha and beta are set to 1, which results in a uniform distribution where all values are considered equally likely. We refer the readers to [16] for a full description of the Beta distribution.



$$f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$E(\alpha,\beta) = \frac{\alpha}{\alpha+\beta} \qquad \begin{array}{l} \alpha = \text{successful entries} + 1 \\ \beta = \text{unsuccessful entries} + 1 \end{array}$$

Beta plot generated for three participants (alpha, beta): (2,10), (4,6), (20,5) with mean of (0.16), (0.40), (0.80) and confidence value of (0.67), (0.48), and (0.80) when 0.10 is used for the acceptable level of error.

Fig. 1. Density of Beta functions for various types of participants.

Figure 1 shows the density of Beta function for different participants. The acceptable error level was defined as 0.1. By having more evidence for a certain hypothesis, as is the case with the participant with alpha of 20 and beta of 5, the level of confidence is large, 0.80. A participant with only a few entries, alpha of 4 and beta of 6, the confidence factor is significantly lower, 0.48. The Beta formulation for reputation affords us other features that are useful in monitoring participants as well. For instance, an aging factor can be introduced to account for quality variations over time by discounting past contributions as a campaign executes, and higher weights can be introduced for contributions that are made in high priority contexts (discussed in Section IV) [12]. Also, contributions do not have to be binary: ratings between 0 and 1 can be made by appealing to the Dirichlet process [17].

IV. Preliminary Evaluation

Since we are in the preliminary stages of several campaigns, we focus on evaluating campaign-specific participation metrics using information gathered from pilot studies of Walkability, PEIR, and Project Budburst campaigns.

A. Walkability and Image Quality

To pilot the Walkability campaign, we asked 6 participants to walk the Westwood neighborhood of Los Angeles and take geo-tagged images of sidewalk segments with visible damage such as cracks. The campaign ran for 2 weeks. Figure 2 shows an analysis of whether participants' contributions were adequate to assess the state of the sidewalk. Images deemed too blurry or dark by a human were considered inadequate. A sense of the epistemic uncertainty helps analyze whether a given participant would be useful as a campaign continues. For example, participant #1 has a high mean (likely to contribute adequate information) but our confidence in his ability is low since the number of contributions he made is small. The confidence factor, however, for the other participants is high because have contributed much more data. We can consequently be much more certain about the abilities of the other participants. Note that we used an error level of .1 when calculating the epistemic uncertainty.

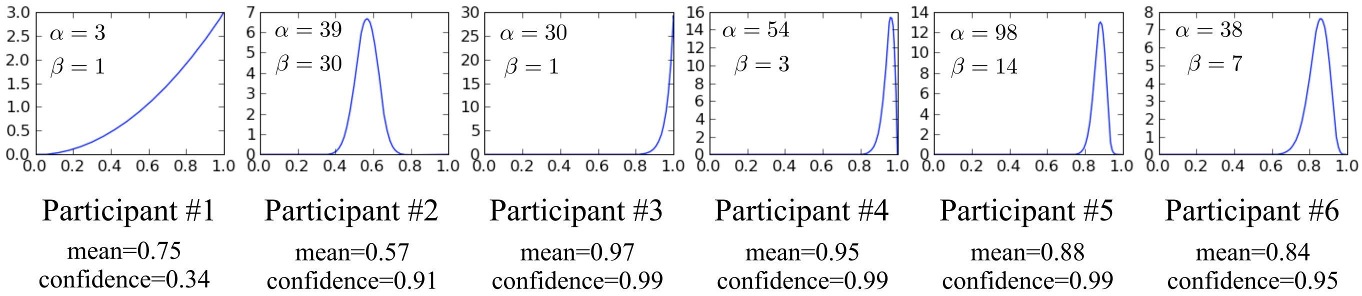


Fig. 2. Walkability pilot analysis showing the likelihood of contributing adequate contributions.

B. PEIR and Long-Term Participation

During the PEIR technical pilot campaign, participants were encouraged to contribute location traces as frequently as possible to test the performance and accuracy of the system. Figure 3 shows the analysis of participation over 61 days for the 26 participants in the pilot. PEIR participants can be clustered into three types of users: consistent, bursty, and sporadic. Consistent contributors contributed data in a dedicated manner throughout the campaign. Bursty participants showed concentrated bursts of contribution, perhaps due to reminders sent to solicit participation. Sporadic participants were very inconsistent or even one-time contributors. By breaking up the campaign into intervals and evaluating participants using the Beta distribution, we could cluster participants according to participation and send feedback based on this information.

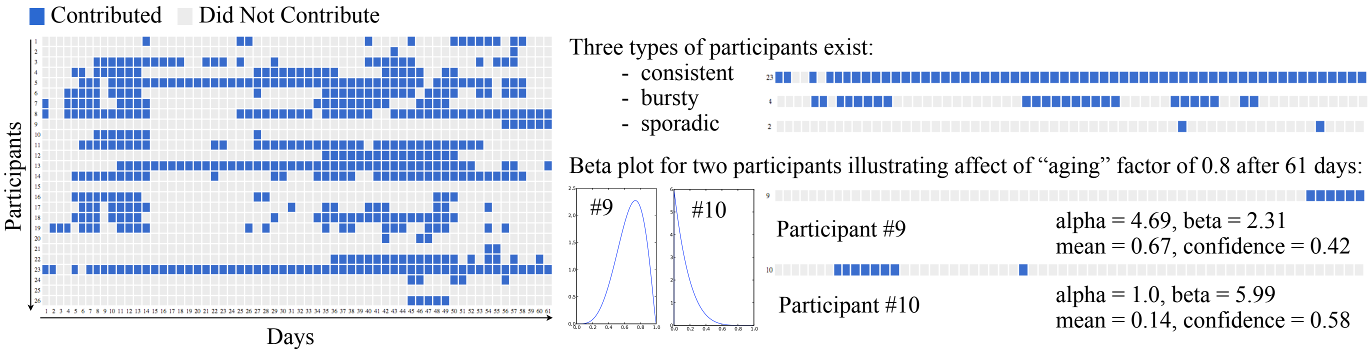


Fig. 3. PEIR analysis showing long-term participant contributions and aging factor.

A feature of the Beta distribution beneficial for long-running campaigns is the aging factor, which we demonstrate by analyzing PEIR participants #9 and #10. Participants #9 and #10 contributed roughly similar amounts of data, but did so during different periods of the campaign. Participant #9 contributed during the tail end of the campaign while participant #10 was heavily involved during the beginning. By having a weight of .8 for the aging factor, we see that participant #9 would be more likely than #10 to contribute data (thus giving #9 a higher mean) in the immediate future. Incorporating an aging factor into the reputation mechanism indicates reliability over time, important since campaigns with a long temporal duration may experience bursty participation.

C. Project Budburst and Reliability

For the Project Budburst pilot we recruited 11 participants. Because we did not have any prior measure of their reputation as data collectors, we initiated a short calibration exercise. The campaign organizer documented flowering plants using geo-tagged images along three specific routes. Participants then traversed the same routes looking for flowering plants. Routes #1 and #2 were short and consisted of 15 and 7 flowering instances, respectively. Route #3 was longer and had 23 instances of flowering. By comparing participant contributions to the calibration phase, we can quickly identify highly reliable participants as well as ones who may required feedback to adjust their data collection practices. For instance, by looking at the results from routes #1 and #2, we could predict which users would be most effective in route #3. The participants who achieved mean of .88 and mean of .92 (with confidence level greater than .90 when the acceptable error is .1) using the Beta distribution for the first two routes, contributed samples that most closely matched the ground truth for route #3. On the other hand, the participants who performed poorly, mean of .58 and .71, on the first two routes followed up with samples that were least consistent with the ground truth. This illustrates that using the Beta distribution with a calibration phase could effectively provide an initial measure of a participant reliability useful for adaptive recruitment or feedback.

V. Discussion and Future Work

Both the challenge and the promise of participatory sensing emerge from involving people in the sensing process. The ultimate "smart sensors," people can make decisions that increase data accuracy, but they also vary according to participation and performance. This paper presents a set of participation and reputation metrics along with a model to help organizers of sensing campaigns determine the reputation and fit of potential participants and whether adjustments are needed during campaign execution. This work is just a first exploration; further study will take place as we incorporate an adaptive participant recruitment system. We will explore how best to communicate reputation ratings, provide actionable feedback to improve a participant's reputation, and determine whether the metrics can become incentives. Another area for future work is exploration of how attributes used in recruitment, such as social network membership or external credentials, may affect reputation measures.

References

[1] J. Burke, et. al., "Participatory Sensing," *ACM Sensys World-Sensor-Web*, 2006.

[2] S. Eisenman, et. al., "MetroSense Project: People-Centric Sensing at Scale," *ACM Sensys World Sensor Web Workshop*, 2006.

[3] E. Paulos, R. Honicky, and E. Goodman, "Sensing Atmosphere," *Workshop on Sensing on Everyday Mobile Phones in Support of Participatory Research*, 2007.

[4] National Phenology Network, "Project budburst," 2008. [Online]. Available: http://www.windows.ucar.edu/

[5] E. Agapie, et. al., "Seeing Our Signals: Combining location traces and web-based models for personal discovery," *HotMobile*, 2008.

[6] S. David, "Toward Participatory Expertise," *Structures of Participation in Digital Culture*, 2007.

[7] P. Resnick, et. al., "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

[8] Amazon, "Amazon mechanical turk," 2008. [Online]. Available: http://www.mturk.com

[9] GURU.com, "Guru.com - freelancers at online service marketplace." 2008. [Online]. Available: http://www.guru.com

[10] Amazon, "Amazon askville," 2008. [Online]. Available: http://www.askville.com

[11] Yahoo, "Yahoo answers," 2008. [Online]. Available: http://answers.yahoo.com

[12] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Systems*, vol. 43, no. 2, pp. 618–644, 2007.

[13] E. Babbie, "The practice of social research," 2007.

[14] A. Schlosser, M. Voss, and L. Bruckner, "Comparing and Evaluating Metrics for Reputation Systems by Simulation," *Paolucci*, 2004.

[15] W. Teacy, J. Patel, N. Jennings, and M. Luck, "Coping with inaccurate reputation sources: experimental analysis of a probabilistic trust model," *Conference on Autonomous Agents and Multiagent Systems*, pp. 997–1004, 2005.

[16] A. Jøsang and R. Ismail, "The beta reputation system," *15th Bled Electronic Commerce Conference*, pp. 324–337, 2002.

[17] S. Ganeriwal, L. Balzano, and M. Srivastava, "Reputation-based framework for high integrity sensor networks," *IEEE TOSN*, 2008.