# Citizen Engineering: Methods for "Crowdsourcing" Highly Trustworthy Results *

Zhi Zhai, David Hachen, Tracy Kijewski-Correa, Feng Shen, Greg Madey
University of Notre Dame
Notre Dame, IN 46556, USA
zzhai, dhachen, tkijewsk, fshen1, gmadey@nd.edu

## Abstract

*Citizen Engineering seeks to leverage a large number of ordinary citizens to solve real-world problems. Emerging information technologies provide us with opportunities to answer a long-standing challenge in citizen engineering – can we effectively extract reliable results from a myriad of crowd inputs of varying quality? To investigate efficient approaches to achieving this "wisdom of crowds", we established a prototype site, where 242 students, acting as surrogate citizen engineers, signed up, logged in, and performed engineering tasks – tagging photographs of earthquake damage. Based on the analysis of user online behaviors, we developed an operable data mining algorithm to retrieve highly trustworthy results from thousands of limited size submissions collected from a cohort of contributors. By converging weight assignments and crowd consensus step-by-step, this extraction algorithm improves the quality of the results over time.*

## 1. Introduction

Emerging information technologies provide us with unprecedented opportunities to build transformative cyber-infrastructures. Characterized by broadband networks, remote and shared computational facilities, and large storage capacities, these new technologies can deepen and broaden participation by users scattered across different locations. As such, citizen engineering, where educationally diverse and physically dispersed users perform tasks, can exploit these new advances of information technology.

Meanwhile, in the domain of civil engineering, researchers and engineers are still restricted by yesterday's compartmentalized resources and solitary strategies, while facing today's multi-dimensional challenges – the information and expertise for complex system design like buildings,

bridges and other civil infrastructure are usually trapped inside proprietary systems, and thus these projects rarely benefit from the full capabilities already available in the larger engineering community [11]. At the same time, the aging civil infrastructure of the developed counties, and the underdeveloped infrastructures of the developing countries requires more engineering efforts, such as regular inspections.

We are thus motivated to investigate citizen engineering systems that address these challenges. We designed a prototype web-based system that channels individual efforts to crowdsource a broad range of tasks to *Citizen Engineers* – web-connected professionals, researchers, students, and even the public at large. However, given diverse education backgrounds and expertise of citizen contributors, we need to provide methods and algorithms to handle collective inputs that may have variable quality.

In this paper, we present a prototype citizen engineering photo-tagging site built to investigate the "wisdom of crowds" [22], with the vision of providing guidelines for successful future citizen engineering and aggregation systems.

After the 2011 Haiti Earthquake, to help local residents rebuild their homeland, civil engineers visited the country and took many photos of damage to various buildings to inform redesign and rebuilding efforts [18]. However, the number of these photos exceeded their capacity to classify the damage displayed in each scene. This motivated us to design a web platform that is able to leverage an online crowd to fulfill this photo classification task, which was previously only done by experienced professionals.

Based on the analysis of the data collected from this citizen engineering site, a data mining algorithm is developed that can provide researchers in related areas with a method for aggregating a myriad of limited size submissions and extracting highly trustworthy results from those inputs. Applying the algorithm to the data set derived from the crowds' work, it shows that the algorithm yields substantial improvements in photo classification accuracy.

Given the fact that our prototype citizen engineering site aims to solve a photo classification problem in the area of

civil engineering, it should be acknowledged that there are special requirements on highly reliable results in civil engineering projects.

## 2. Related Work

In developed economies, many people enjoy more spare time than ever before. However, much of that spare time is often spent on unproductive activities. This *Cognitive Surplus* [20] makes the idea of *Citizen Engineering* feasible, where well-designed mechanisms can engage and channel distributed human brainpower to solve time consuming problems that computers cannot yet handle well.

Open source software projects (e.g., Linux and Apache) serve as successful examples of citizen engineering, demonstrating that voluntary, collective, human effort of loosely organized individuals can generate useful intellectual products [17]. In such projects we see the power of collective intelligence [14], where crowds gradually enhance product quality over time. Such a harnessing of collective intelligence to achieve a common research or education purpose has many examples [23]: eBird [21], Galaxy Zoo [19], Stardust@home [4], and Foldit [9].

Public engagement in citizen-based projects has a long history, but new advances in information technology enable novel approaches and applications, such as urban planning [5], astronomical data analyzing [4][19], civil infrastructure flaw detecting [1][7], environment protecting [13], and socio-political movements [8][6].

On the other hand, a challenge in citizen engineering is the vastly diverse backgrounds of users. Some probably have many years of professional training, others may be amateurs or hobbyists. There are possibly even a few with malicious intent. This raises challenges concerning quality control, motivation and result aggregation. To motivate a large number of citizen engineers to perform meaningful tasks, it is essential to develop a practicable workflow to secure product quality and achieve highly trustworthy results, but not be so restrictive as to quench citizen engineers' enthusiasm to participate.

## 3. The Experimental Investigation

This research is part of a study named Open Sourcing the Design of Civil Infrastructure (OSD-CI) [12]. In one OSD-CI citizen engineering investigation, we designed a web platform to attract citizen engineers and facilitate their contributions classifying earthquake damage photos. College students were recruited using announcements on mail lists and school-wide posters, resulting in 242 students participating in the experiment as surrogates for citizen engineers. Their work and online activities were recorded, including photo tagging classifications, the time spent tagging

each photo, and login/logout timestamps.[1]

During a 17 day period, Nov. 21, 2010 - Dec. 7, 2010, we received 9318 photo classifications of 400 photos (over 23 classifications per photo). Variable quality of these classifications was observed as the students displayed varying levels of seriousness on the tasks, and came from a broad range of backgrounds – some of the contributors were civil engineering majors, while others had little knowledge of engineering. This heterogeneity in expertise mimics what is commonly observed in crowdsourcing projects – highly diverse education levels of users and variable quality of work. In this study, we designed a data mining workflow, aiming to detect inputs from careless users, prune noising inputs, integrate valid inputs, and achieve highly trustworthy results from crowd classifications.

Using the typology of Malone *et al.* [16], our prototype system utilizes a cohort of online users, incentivized by financial rewards or moral motivations, to collaboratively tag damage building photos, where results are aggregated and retrieved automatically by computers.

This study also investigated the efficacy of different incentives on the quality and quantity of the work performed by citizen engineers. However, in this paper we ignore those experimental conditions and consider all the inputs from our subjects irrespective of their experimental treatment; those other results are reported elsewhere [10].

## 4. Experiment Procedures

### 4.1. Brief Description

Upon agreeing to a consent form, subjects were directed to a sign-up page, and asked to create their login credentials, where their campus login was verified to confirm that they are students. If the personal information entered was valid, a new account was created and a confirmation email was sent to his/her email address.[2] After viewing the introduction page, subjects were directed to a tutorial. They then proceeded to the task of classifying photographs. They could tag as many of the photographs as they wanted to within a seven-day period. The photographs depicted damage to buildings as a result of the Haitian earthquake.

### 4.2. Detailed Procedures

The web site consists of several components, including registration, an entry survey, tutorial, photograph classification, and an exit survey. The design goal was to let subjects

---

[1]The procedure for photo classification was developed by researchers from the Departments of Sociology, and Civil Engineering and Geological Sciences, University of Notre Dame.

[2]All subjects were recruited from the student population of the University of Notre Dame.

participate in the experiment from any place and at any time they chose.

1. **Registration** After subjects logged into the website, they saw a consent form with a brief description of the experiment: The task was to classify the type of earthquake damage depicted in 400 photos. Participants had the right to withdraw from the experiment at any time.

2. **Entry Survey** The purpose of this questionnaire was to collect demographic and attitudinal data from the subjects. Survey questions include: *Gender, Ethnicity, Moral Views, Voluntarism* (Attitude, Practice, etc.), *Education* (GPA, Major, etc.), *Workload* (Free time, Web-surfing time, etc.), *Religious background* (Affiliation, Religiosity, etc.).

3. **Tutorial** The goal of the online tutorial, which also embeds several self-quizzes, was to ensure that subjects had full information about the assigned tasks. The tutorial provides detailed information about how to successfully classify the damage depicted in a photo, and by using hyper-links, subjects could return to this tutorial to deepen their understanding as many times as they wish during the tagging process.

4. **Damage Classification** Subjects received a single, randomly chosen photo at a time, until they completed all the 400 photos in the database or the allocated time period expired. After submitting a classification of a photo they were not able to modify it. However, they could save their progress and return to the experiment at a later time until their time ran out (7 days after finishing the tutorial) or until they decided to opt out.

5. **Exit Survey** At the end of the seven-day period subjects were asked to complete a brief exit survey. We asked questions like why subjects decided to allocate time to classification work (motivation), the difficulty in classifying photos, the degree to which they found this to be an interesting task, and if they discussed the experiment with others.

### 4.3. Tagging Questions

As shown in Fig. 1, to classify a photo, subjects followed a five-step damage assessment process. These steps are:

1. **Image Content** Determine if an entire structure or only a part of the structure is visible in the image.

2. **Element Visibility** Identify which elements (*beams, columns, slabs, walls*) of the building are visible and can be assessed.

3. **Damage Existence** For each of these visible elements, determine if any of those elements are damaged.

4. **Damage Pattern** For each of the elements identified as damaged, identify the damage pattern.

5. **Damage Severity** For each of the elements identified as damaged, assess the severity of the damage (*Yellow* or *Red*).

Since we ask at most 25 classification questions for each photo, a user can get 25 points maximally from one photo. In particular, for each question, if this user's answer is same as the crowd consensus (defined using the algorithm discussed in Section 6), s/he receives one point. Otherwise, this user does not earn a point on the question. If the crowd consensus is that there is no damage on a certain element of the building, we do not further consider the user's inputs about the damage pattern and severity of that building element. As such, no matter what answers the user provides to the questions asking about the pattern and severity in that element, no points are assigned. In this regard, the maximal score a user can get from a photo is usually less than 25.

Compared to the similar image classification work conducted in [2], this paper presents a more sophisticated photo tagging schema with great potential to generate new knowledge because of its detail.

## 5. Data Collection and Cleaning

### 5.1. Data Collection

Fig. 2 shows the daily number of new registered users and the cumulative number of users. Also, the daily new classifications and cumulative number of classifications are shown in Fig. 3. The peaks correspond to Thanksgiving break in 2010, when students had more time to participate.

### 5.2. Data Cleaning

The first challenge we encountered was that there were some freeloaders, who just clicked through all the photos without seriously thinking about their answers. To identify these users, we evaluated several noise-pruning methods, finally using the average time spent on each photo for the following reasons.

It takes roughly 40 seconds to classify one photo. For an unambiguous photo taken from a close-up position, 20 seconds is enough to obtain an accurate answer, while more ambiguous ones may take as long as 3 minutes. In the experiment, users saw these photos in a random order, and as such the average tagging time fell in the range of 20-60 seconds. Given these averages, if a user's average tagging time was less than 15 seconds, we were highly doubtful that s/he was a serious photo tagger.

Another complication is that some photo classifications had abnormally long tagging times, such as the ones in the
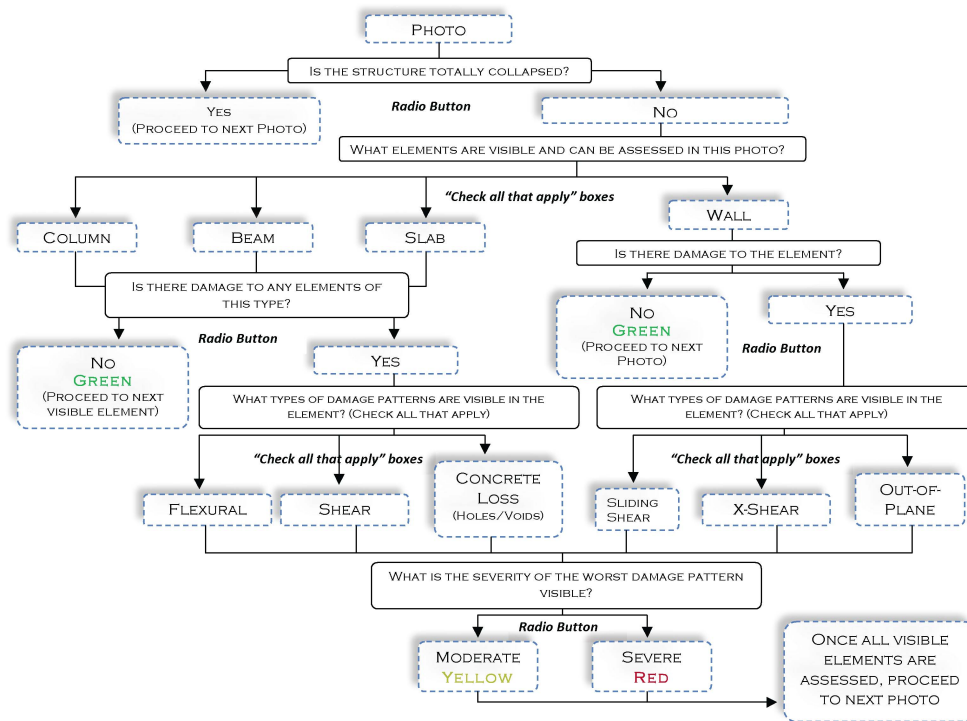
**Figure 1. Classification schema. As online users went deeper along the tree, their answers diversified.**

rightmost bin shown in Fig. 4. Causes of these outliers may be that users may also conduct other activities while classifying photos. For example, replying to emails and talking on the phone may be going on, resulting in photo tagging time being artificially prolonged. For the analysis reported in the paper, we consider tagging periods longer than 300 seconds as outliers, and do not take them into account when calculating the user average tagging time.

After pruning these outliers, the average tagging time is depicted in Fig. 5, from which we can observe that 8 subjects fall into the first bin, which means they spent less than 10 seconds classifying a photo on average. These 8 subjects and 2 additional from the second bin, which have much lower tagging time than others, are identified as mischievous clickers, and their inputs are removed from the aggregation table in the database. After data cleaning, we obtained 6186 valid photo classifications from 194 users.

## 6. Result Extraction Algorithm

### 6.1. Algorithm Principles

Our ultimate goal is to obtain trustworthy results from crowdsourced efforts. In this study, the workflow we de-

signed for determining the crowd consensus is shown in Fig. 6, and, similar to Galaxy Zoo, our strategy is to pay more attention to inputs from users who tend to agree with the crowd consensus [15]. Specifically, the algorithm progressively increases skillful users' weights over low-performers by assigning them into different rating groups. Within each group all users have the same weight, and different groups have unequal weights. In this manner, by taking into consideration not only the number of users, but also their weights, we make crowd consensus tilt towards to the opinion of the more reliable users, who are from heavier-weighted groups.

### 6.2. Algorithm Implementations

#### 6.2.1 Crowd Consensus Calculation

At the very beginning, we assign equal weights, say 1, to all users. In the first iteration, since all 194 users have the same weight, the value of the weight actually does not have an effect on the crowd consensus calculation. To determine the crowd consensus, the basic guideline is that the higher the weight a user has, the larger proportion his or her opinion will take in the calculation. Every photo has up to 25
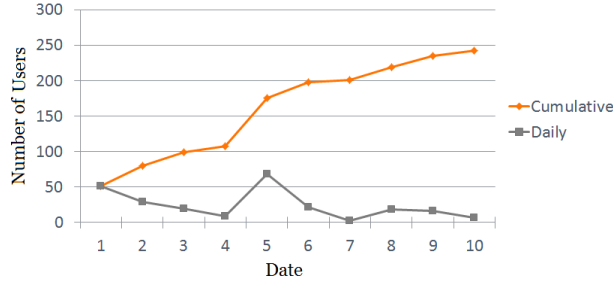
Figure 2. Daily and cumulative number of users. The peak corresponds to Thanksgiving vacation, 2010.
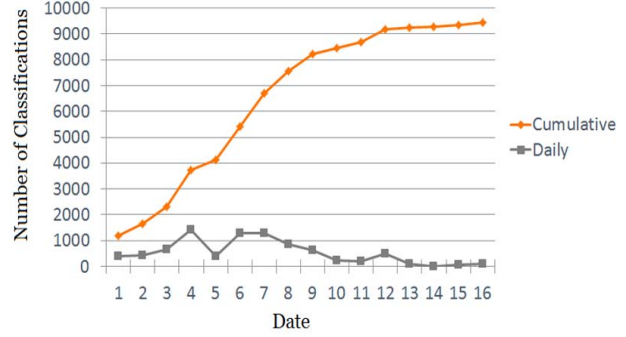


Figure 3. Daily and cumulative number of classifications. The peak corresponds to Thanksgiving vacation, 2010.

Table 1. Group assignment

| Group Index | Number of Users | Weight |
|---|---|---|
| 1 | 40 | 5 |
| 2 | 40 | 4 |
| 3 | 40 | 3 |
| 4 | 40 | 2 |
| 5 | 34 | 1 |



Figure 4. Photo tagging time (equal-width discretization).

questions, with each of them having 2 or 3 candidate options. To determine which option crowds converged on, we need to calculate how many votes each option obtained by using Equ. (1), where Option $i$ from Question $j$ on Photo $k$ receives $V_{ij}^k$ votes in total.

$$V_{ij}^k = \sum_{m=1}^{n} (N_m * W_m) \qquad (1)$$

$N_m$ represents the number of users in Group $m$ that vote for Option $i$ from Question $j$ on Photo $k$, and $W_m$ is the weight of Group $m$. All users within the same group have an equal weight.

Here is another example illustrating the consensus calculation: for instance, 4 users classified Photo $k$, where user $A$ has weight 2, user $B$ has 4, user $C$ has 6, and user $D$ has 8. When answering Question $i$, user $A$ and $B$ selected Option $x$ as their answers, and user $C$ and $D$ selected Option $y$. In this manner, Option $x$ will get (1*2+1*4) = 6 votes, and Option $y$ gets (1*6+1*8) = 14 votes. Therefore, the crowd consensus on Question $i$ of Photo $k$ is Option $y$, since Option $y$ obtained more votes than Option $x$.

If the numbers of votes that Options $x$ and $y$ obtained are equal, there would be a two-way tie. In this case, both Option $x$ and $y$ are considered to be the crowd consensus.
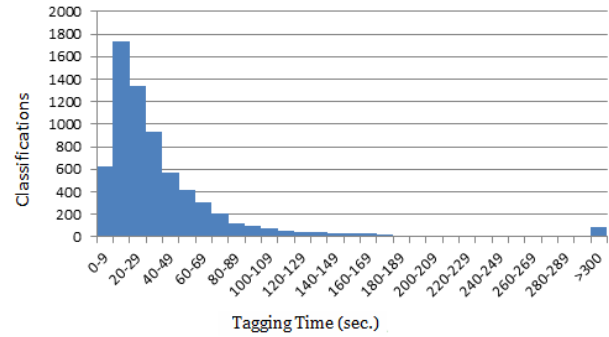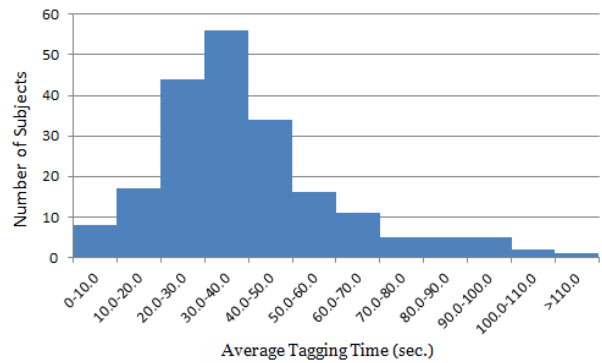


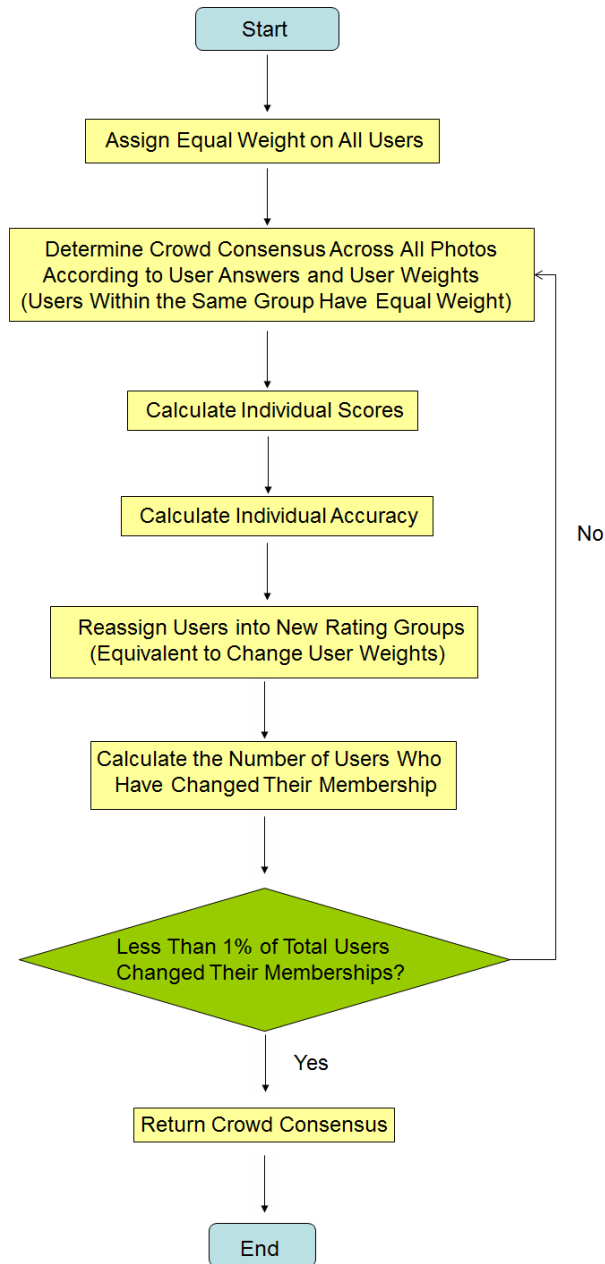Figure 5. Subject distribution on average tagging time.
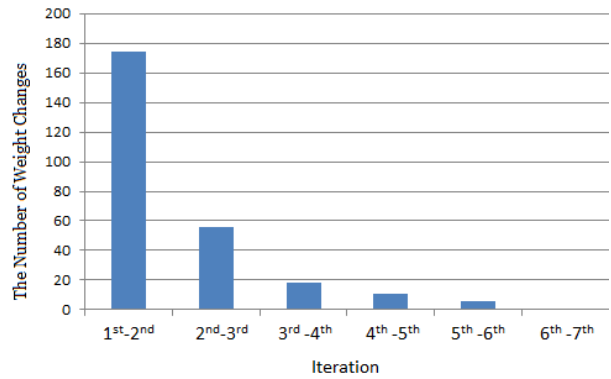
Figure 6. The result extraction algorithm.



**Figure 7. The number of weight changes between iterations.**

### 6.2.2 Individual Score

After determining the crowd consensus for each item in the photo, we calculate the number of points a user obtained from each photo classified. For a question, if the user answer agrees with the crowd consensus, s/he is given one point. Otherwise, this user does not receive any points from this question.

The overall score for an individual is calculated by Equ. (2), where $S_u$ is the overall score of User $u$, $s_{ku}$ is the number of points User $u$ obtained from a single Photo $k$, and $p$ is the total number of photos classified by User $u$.

$$S_u = \sum_{k=1}^{p} s_k \qquad (2)$$

### 6.2.3 Individual Accuracy

Since we knew which user classified which photo, namely user-photo pair in the data collection step, we are able to calculate a user's accuracy by dividing the total points s/he accumulated over the maximal points this user could have possibly obtained across all photos that s/he classified if all her/his answers were the same as the crowd's consensus.

For example, if user Alice tagged 2 photos, and got 13 points out of 20 in photo A, and 2 points out of 10 in photo B, her accuracy will be 50% , since (13+2)/(20+10)=50%.

We believe this is an effective way to do the accuracy calculation, since different photos have variable difficulties. In a pancake collapse [3], which is believed to be the simplest case, one just needs to answer the first question: *Can-*

---

[3] The entire building was badly damaged, and no single building element can be accurately assessed.

*not Determine*. On the other hand, for some complicated scenarios, users had to identify damage patterns associated with each building part. Therefore, photos with different levels of difficulty should have different weights in the final accuracy calculation. So, in the above example, it is not proper for us do the calculation this way: ((13/20)+(2/10))/2 = 35%, where all photos have equal weights in the determination the overall accuracy.

### 6.2.4 Group Assignment

Based on the user accuracy, the 194 valid users can be divided into 5 groups, with around 40 users in each group, except for the Group V, which has 34 group members. The top-performing group, Group I, has the highest weight, 5, and the low-end group, Group V, has the lowest weight, 1. The 3 groups in between have weights 2-4 in the next iteration, as shown in Table 1. In this example, users are presumably divided into 5 groups, but later we also investigated different group sizes to achieve better results (Fig. 8).

### 6.2.5 Next Iteration

After reassigning weights, we go to the second iteration. Based on the new weight assignment, a new crowd consensus, new individual scores and accuracies, and new group assignments are calculated. As shown in Fig. 6, this loop continues until the stop criterion is met.

### 6.2.6 Stop Criterion

Between two consequent iterations, if there are few changes on users weights, that means crowds consensus and user scores are stable. In practice, we set the stop criterion with a *1% User Rule*, which means if there are less than 1% of users who have to change their weight assignments between two consequent iterations, then the loop terminates, and eventually the algorithm outputs the crowd consensus as the final result.

Implemented in Matlab, when we ran the program, this procedure terminated at the end of the sixth iteration where there were 5 groups. The number of users who changed weights between iterations is shown in Fig. 7.

## 7. Experimental Results

## 7.1. Ground Truth

Intuitively, the more votes one option gets, the more likely it is correct. We expect a high quality in the crowd consensus generated from the result retrieval algorithm. To evaluate the crowd consensus, we employed 3 graduate students from the Department of Civil Engineering to provide professional classifications of all 400 photos. Those 3 experts were asked to classify those 400 photos following the exact same procedure as used in the experiment. We collected this authoritative data with 3 foci:

- *What is the correct answer for each question on each photo?*

- *What is the number of maximal points crowds can obtain on each photo?*

- *What is the normal average classification duration across all photos?*

The classifications of the professional taggers is considered as the ground truth when they agree. We disregarded the ambiguous questions on which 3 experts did not agree with one another. Based on this ground truth, the number of maximal points across all 400 photos is 4905. At this point, to compute crowd scores, we can compare the answers from the crowd and the assessed ground truth after each iteration. It is similar to the way we calculated individual scores before: for each question, if the crowd consensus and the professional taggers' answer are the same, the crowd receives one point. Otherwise, they do not receive a point on this question. If there is a 2-way tie in the crowd consensus, crowds will receive a half point, and if there is a 3-way tie, the crowds will receive one third point.

According to these evaluating metrics, the ideal scenario of crowd performance will be like this: crowd scores progressively improve, as the crowd consensus gradually gets closer to the ground truth by focusing on the inputs from more skillful photo taggers, who achieved high scores in the previous iteration.

## 7.2. Crowd Performance

The actual crowd scores are shown in Fig. 8, where we also varied the group size to investigate the best group denomination. In 4 separate runs, 194 users are divided into 5, 10, 20, and 40 groups respectively.

- *5-Group*: 194 users are divided into 5 groups, and approximately 40 users in each group; the highest group weight is 5, and the lowest group weight is 1.

- *10-Group*: 194 users are divided into 10 groups, and approximately 20 users in each group; the highest group weight is 10, and the lowest group weight is 1.

- *20-Group*: 194 users are divided into 20 groups, and approximately 10 users in each group; the highest group weight is 20, and the lowest group weight is 1.

- *40-Group*: 194 users are divided into 40 groups, and approximately 5 users in each group; the highest group weight is 40, and the lowest group weight is 1.

**Table 2. User statistics according to classification accuracy - *Top* group.**

| Rank | Num. of Photos | Avg. Time (Sec.) | Accuracy (%) |
|------|----------------|------------------|--------------|
| 1 | 6 | 51.5 | 95.4 |
| 2 | 4 | 45.0 | 94.3 |
| 3 | 4 | 37.0 | 89.1 |
| 4 | 13 | 68.4 | 89.1 |
| 5 | 10 | 55.9 | 89.0 |
| 6 | 62 | 85.8 | 88.6 |
| 7 | 23 | 64.2 | 88.2 |
| 8 | 25 | 35.0 | 87.8 |
| 9 | 9 | 79.3 | 87.5 |
| 10 | 12 | 81.1 | 86.2 |

**Table 3. User statistics according to classification accuracy - *Bottom* group.**

| Rank | Num. of Photos | Avg. Time (Sec.) | Accuracy (%) |
|------|----------------|------------------|--------------|
| 1 | 27 | 22.1 | 46.5 |
| 2 | 10 | 26.7 | 45.0 |
| 3 | 37 | 12.8 | 44.3 |
| 4 | 14 | 13.7 | 43.6 |
| 5 | 9 | 18.3 | 41.9 |
| 6 | 39 | 13.4 | 40.5 |
| 7 | 15 | 27.0 | 39.3 |
| 8 | 15 | 26.9 | 37.3 |
| 9 | 41 | 28.3 | 31.0 |
| 10 | 4 | 19.5 | 21.9 |

From Fig. 8, we observe that the highest crowd accuracy is achieved when 194 users are divided in 5 groups after the second iteration. So, hereafter the discussion will be based on the 5-group division. As Fig. 8 shows, in the beginning, the crowd score does increase as we predicted. However, starting from the third iteration, the score flattens out and slightly declines in later iterations. As we interpret it, after the second iteration, the opinions of top performers are over-represented by being assigned exceedingly heavy weights. Since we overlooked other users' inputs, the "wisdom of crowds" is insufficiently harnessed.

## 8 Discussion

Our citizen engineer surrogates are college students, who may be collection of citizens with above average aptitude,

so our experimental results should be further evaluated before being generalized to average online crowds. Also, we find other interesting research questions suggested by the data:

- *Is the average tagging time spent on photos significantly correlated with user performance (score)?*

Average tagging time may be a good indicator of classification quality; if users spend more time on each photo, they may be more careful and responsible about their inputs. However, there is a complication; spending more time may imply that those users are inefficient, and often have to consult tutorials or reference books to confirm their solutions.

- *Is the number of photos a user classifies significantly correlated with user performance (score)?*

Intuitively, we would think users who have classified more photos are more enthusiastic, and thus more serious about the project. However, another possibility is that users who classify large numbers of photos are not essentially interested in the project, and they are just curious about what those photos are, and therefore simply make some random selections to get through, which leads to an artificially high number of classifications.

To address these two questions, we conduct a *Student's t-test*. We sort users according to their classification accuracies. We disregard the users who classified less than 3 photos, since their accuracies and average tagging times may be deemed as unstable. After data pruning, all photo taggers are sorted by their individual accuracies, and then stratified into two layers – *High-accuracy* layer and *Low-accuracy* layer, with equal number in each layer. Then 10 taggers from the high-accuracy layer are selected into the *Top* group, and 10 taggers from the low-accuracy layer are sampled into the *Bottom* group. The user statistics of the Top and Bottom groups are shown in Table 2 and Table 3, and the following unpaired student's t-test is conducted based on this 20-user sample, from which we make inferences about the entire population.

### 8.1. Average Tagging Time vs. Accuracy

The student's t-test is a technique to test if there is a statistically significant difference between the means of two populations. With the users in the High-accuracy and Low-accuracy layers as the two targeted populations, we have two hypotheses:

- *Null Hypothesis* There is no statistically significant difference between the average photo tagging times associated with users from the High-accuracy layer and Low-accuracy layer.

- *Alternative Hypothesis* There is a statistically significant difference between the average photo tagging
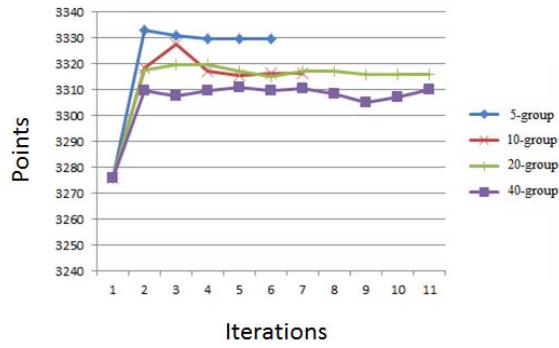
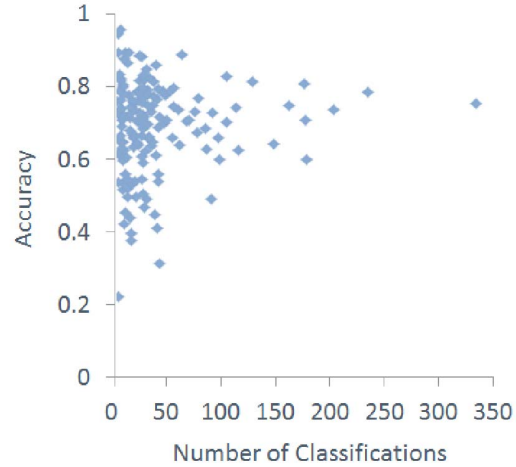**Figure 8. The crowd score after each iteration.**



**Figure 9. Average tagging time vs. classification accuracy. Each point represents one user (n=194).**



**Figure 10. Number of classifications vs. accuracy. Each point represents one user (n=194).**

## 8.2. Number of photos classifications vs. Accuracy

We conducted another t-test following similar procedures on the parameters of the number of photos vs. accuracies, where two hypotheses are established:

- *Null Hypothesis* There is no statistically significant difference between the average number of photo classifications associated with users from the two layers.

- *Alternative Hypothesis* There is a statistically significant difference between the average number of photo classifications associated with users from the two layers.

In this testing, at the 95% significance level we have no sufficient evidence to reject the null hypothesis that there is no significant difference between the average number of photo classifications from the users in the two different layers, which we may intuitively observe in Fig. 10.

## 9   Conclusions and Future Work

To design and deploy more effective citizen engineering projects, we developed a robust and operable workflow to effectively aggregate users inputs and extract highly trustworthy results from the "wisdom of crowds." Based on these initial results, we find that some interesting research topics are worth further investigation. For example, what are the motivations behind users' volunteer work? What is the optimal number of users to work on the same photo to secure a quality result? How should we rate and group online users based on their performance? Answers to these

times associated with users from the High-accuracy layer and Low-accuracy layer.

In Fig. 9, the scatter plot shows the distribution of data points representing the average tagging time vs. classification accuracy. Testing with the statistical tool R [3], we conclude at the 95% significance level that we have enough evidence to reject the null hypothesis that there is no significant difference between the average photo tagging times in users from High-accuracy layer and Low-accuracy layer, and accept the alternative hypothesis, that there is significant difference between the two layers. In other words, average tagging time is likely a meaningful indicator of classification accuracy.

questions could help to guide future research and development on how best to extract the wisdom of the crowd from large numbers of inputs that vary in their quality.

# References

[1] Infrastructure Sensing. `https://firefly.cse.nd.edu/infra`, Retrived Jan. 2010.

[2] ImageCat. `http://www.imagecatinc.com/`, Retrived Aug. 2011.

[3] R Language and Environment. `http://www.r-project.org/`, Retrived Jul. 2011.

[4] D. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer. SETI@home: an experiment in public-resource computing. *Communication of the ACM*, 45:56–61, Nov. 2002.

[5] J. Burke, D. Estrin, M.Hansen, A. Praker, N. Ramanathan, S. Reddy, and M. Srivastava. Participatory sensing. In *ACM Sensys World Sensor Web Workshop*, Boulder, CO, USA, Oct. 2006.

[6] V. Coelho and B. Lieres. *Mobilizing for Democracy: Citizen Action and the Politics of Public Participation*. Claiming Citizenship: Rights, Participation and Accountability. Zed Books, 2010.

[7] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan. The pothole patrol: Using a mobile sensor network for road surface monitoring. In *the Sixth Annual International conference on Mobile Systems, Applications and Services (MobiSys 2008)*, Breckenridge, CO, USA, Jun. 2008.

[8] S. S. Fainstein. Spotlight: Urban social movements, citizen participation, and trust networks. In M. Hanagan and C. Tilly, editors, *Contention and Trust in Cities and States*, pages 175–178. Springer Netherlands, 2011.

[9] J. Jésior, A. Filhol, and D. Tranqui. *FOLDIT* (*LIGHT*) – an interactive program for Macintosh computers to analyze and display Protein Data Bank coordinate files. *Journal of Applied Crystallography*, 27(6):1075, Dec 1994.

[10] Z. Kertcher and D. Hachen. Online work motivation: An experiment of instrumental and moral incentives. In *the 23rd Annual Meeting of the Society for the Advancement of Socio-Economics*, Madrid, Spain, Jun. 2011.

[11] T. Kijewski-Correa et al. `http://www.ne.edu/˜opence/`, Retrived Jul. 2011.

[12] T. Kijewski-Correa et al. Open sourcing the design of civil infrastructure (OSD-CI): A paradigm shift. In *Proceedings of Structures Congress*, Las Vegas, NV, USA, Apr. 2011.

[13] T. S. Lena, V. Ochieng, M. Carter, J. Holgun-Veras, and P. L. Kinney. Elemental carbon and pm2.5 levels in an urban community heavily impacted by truck traffic. *Environmental Health Perspectives*, Oct. 2002.

[14] P. Levy. *Collective intelligence: Mankind's emerging world in cyberspace*. Perseus Books, 1999.

[15] C. J. Lintott1, K. Schawinski1, A. Slosar1, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389:1179–1189, 2008.

[16] T. Malone, R. Laubacher, and C. Dellarocas. The collective intelligence genome. *MIT Sloan Management Review*, 51(3), 2010.

[17] R. McMillan. The great dictator Linus Torvalds: The benevolent, brilliant keeper of the kernel. *FEATURES (Linux Magazine)*, Dec. 2002.

[18] D. Mix, T. Kijewski-Correa, and A. A. Taflanidis. Assessment of residential housing in leogane, haiti after the january 2010 earthquake and identification of needs for rebuilding. *Earthquake Spectra*, Oct. 2011.

[19] M. Raddick, G. Bracey, P. Gay, C. Lintott, P. Murray, K. Schawinski, A. Szalay, and J. Vandenberg. Galaxy Zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review*, 9(1), 2010.

[20] C. Shirky. *Cognitive Surplus: Creativity and Generosity in a Connected Age*. Allen Lane, 2010.

[21] B. L. Sullivan, C. L. Wood, M. J. Cliff, R. E. Bonney, D. Fink, and S. Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.

[22] J. Surowiecki. *The Wisdom of Crowds*. Doubleday, 2004.

[23] A. Wiggins and K. Crowston. From conservation to crowdsourcing: A typology of citizen science. In *Proceedings of the 44th Annual Hawaii International Conference on System Sciences*, Koloa, HI, USA, Jan. 4-7 2011.