Did they login? Patterns of Anonymous Contributions in Online Communities

COREY BRIAN JACKSON, Syracuse University, USA
KEVIN CROWSTON, Syracuse University, USA
CARSTEN ØSTERLUND, Syracuse University, USA

Researchers studying user behaviors in online communities often conduct analyses of user interaction data recorded in system logs e.g., an edit in Wikipedia. Such analysis relies on collating interactions by a unique identifier such as a user ID. However, if users can contribute without being logged-in (i.e., anonymously) analysis of interaction data omit part of a user's experience. Problematically, anonymous traces are unlikely to be randomly distributed, so their omission can change statistical conclusions, with implications for both research and practice. To understand the impacts on conclusions of leaving out anonymous traces, we conducted an analysis of system logs from two online citizen science projects. Attributing anonymous traces with user IDs, we found that (1) many users contribute anonymously, though with varied patterns; and (2) attributing anonymous traces diminishes empirical evidence used to support theory and change the results of system algorithms. These results suggest anonymous traces have implications for research on user behaviors and the practices associated with using such data to tailor user experiences in online communities.

CCS Concepts: • **Human-centered computing → User studies**; *Empirical studies in collaborative and social computing*;

Additional Key Words and Phrases: citizen science, user behavior, online communities, anonymity, Zooniverse

ACM Reference Format:
Corey Brian Jackson, Kevin Crowston, and Carsten Østerlund. 2018. Did they login? Patterns of Anonymous Contributions in Online Communities. *Proc. ACM Hum.-Comput. Interact.* 1, 1, Article 77 (January 2018), 16 pages. https://doi.org/10.1145/3274346

1 INTRODUCTION

An attractive approach for studies of large-scale distributed systems is to analyze the user interaction data recorded by the system in logs, known as trace data. For example, a system that supports commenting will record the text of the comment, the user ID of the individual who posted the comment, and a time stamp indicating the precise date and time the comment was posted. In some systems, viewing webpages or clicking links are also recorded, creating an extensive interaction history for each user. In aggregate, these traces allow researchers to construct and analyze a detailed record of users' interactions with the system and each other.

The potential of trace data [10] has not gone untapped as researchers rely on traces as empirical evidence in support of theories and construct behavioral models. For instance, Burke and Kraut [5, 6] used trace data (e.g., article edits and reverts) to predict users' likelihood to be promoted to

Authors' addresses: Corey Brian Jackson, Syracuse University, School of Information Studies, Syracuse, NY, 13244, USA, cjacks04@syr.edu; Kevin Crowston, Syracuse University, School of Information Studies, Syracuse, NY, 13244, USA, crowston@syr.edu; Carsten Østerlund, Syracuse University, School of Information Studies, Syracuse, NY, 13244, USA, costerlu@syr.edu.

2573-0142/2018/1-ART77
https://doi.org/10.1145/3274346

administrators in Wikipedia. Other studies employ trace data to understand socialization [6–8], motivation [17], learning [9], and role ascension [2, 5–7].

However, using trace data to understand user behavior can be problematic since some systems allow users to contribute anonymously. For example, in the English Wikipedia, visitors can edit most articles without having a registered account; in some Stack Overflow projects, visitors can post comments without having an account. As a result, user IDs are missing from some records posted to the database. Missing user IDs becomes problematic when users decide to registers for an account as the interactions recorded while they were anonymous are omitted from their history of interaction on the system. Results of analyses for research purposes or to tailor user experience on the system could be affected when anonymous traces are missing from the data. Anonymous traces seem unlikely to leave data missing at random, which is a requirement for the missing data to not bias an analysis. For example, some studies suggest that anonymous periods are more likely to include a user's initial interactions with a system [19–21, 25] as visitors lurk (reading materials or observing) before deciding if they want to create an account [9].

While challenging, it is sometimes possible to attribute anonymous traces to users. Some systems record users' IP addresses: those addresses may be unique to a user and so can be used to link anonymous traces with registered users. For instance, Panciera et al. [24] were able to link 20% percent of anonymous edits to registered user accounts in Cyclopath by using IP addresses. To learn more about anonymous traces and their impact on analyses of user behaviors, we studied the contribution of volunteers on Zooniverse, a website that hosts citizen science projects. Contributors to Zooniverse projects classify data on the site that are recorded in a database. However, volunteers are not required to log in to the system before contributing, meaning that some classification database records are anonymous.

Many studies of Zooniverse volunteers rely on trace data. However, to our knowledge, none mention how anonymous traces are handled. Additionally, the Zooniverse platform uses a volunteer's classifications history in algorithms that govern user experience. For instance, a volunteer should not be assigned the same object to classify more than once, which requires knowing which objects a volunteer has seen. The performance of such algorithms is also affected by anonymous contributions. The existence of anonymous traces and IPs make Zooniverse a suitable setting to examine anonymity and its impacts on research and practices. We address the following research question: *How might conclusions about user behaviors based on trace data be impacted by the omission of anonymous user interactions?*

2 ONLINE CITIZEN SCIENCE PROJECTS

Citizen science projects ask members of the public to participate in scientific research by collecting or analyzing data [4]. Citizen science tasks range from classification or transcription of pre–existing data to data gathering and idea generation. For instance, Galaxy Zoo, volunteers are asked to review images of galaxies and answer questions about its morphological characteristics i.e., shape, roundness. In eBird, volunteers are asked to report bird sightings by providing information about the quantity, species, and location of birds [30]. Ornithologists use the data to build better understand the migratory patterns of bird species.

Volunteers are crucial to citizen science projects. To better understand volunteers, studies have addressed topics including motivation [26, 27], socialization [18], learning [16], and engagement [12]. For instance, Mugar et al. [18] found that citizen scientists learn by observing the practices of more experienced members (legitimate peripheral participation). Jackson et al. [12] found that volunteer engagement in Planet Hunters (posting comments, asking questions) is dynamic from session to session. In another study, Reeves et al. [26] showed that motivational reinforcement using elements of gamification increases contributions. Luczak-Rösch et al. [16] found that volunteers' adoption of scientific terminology is dependent on both the project goals and supporting infrastructure.

The studies mentioned above relied, in part, or wholly on trace data extracted from system logs. Both Mugar et al. [18] and Jackson et al. [12] analyzed classification data from a volunteers' first registered interaction and Luczak-Rösch et al. [16] collected comment logs. The dataset in Jackson et al. [12] contained page views and classifications for each volunteer during their tenure in Planet Hunters. The most granular data contains the user ID of the individual whose interaction is represented and the time the event occurred. Collating the records in system logs allowed the researchers to build and analyze a detailed history of each volunteer in Planet Hunters. Other studies have performed similar aggregations of data.

To our knowledge, anonymous traces have not been studied in citizen science. Kawrykow et al. [15] acknowledges work from some sizable fraction of non-registered citizen scientists and Jay et al. [13] showed more users contribute when they can do so anonymously, however, no attempt was made at systematically analyzing anonymous traces in these studies.

3 ZOONIVERSE: HIGGS HUNTERS AND GRAVITY SPY

Our study is set in the context of Zooniverse [28]; an online citizen science platform that currently hosts more than eighty projects and has more than 1.5 million registered volunteers. Professional scientists use Zooniverse to crowdsource the analysis of data by asking volunteers to classify or transcribe pre-existing data, a popular mode of citizen science work [31]. To minimize the barrier to entry, volunteers can classify without registering for or logging into a Zooniverse account. If volunteers are not logged-in, after a few classifications the system will prompt them to login or register for an account. Nevertheless, registration is always optional.

We researched anonymous traces in two citizen science projects: Higgs Hunters and Gravity Spy. Higgs Hunters is a particle physics citizen science project launched in 2014 that helps physicists searching for exotic particles in data from the Large Hadron Collider. Volunteers are shown an image of a collision in which charged particles are represented as lines and asked to mark off-center vertices, which indicate the creation of new particles. A screen shot of the classification interface is shown in Figure 1 (top).

In Gravity Spy, volunteers classify data from the Laser Interferometer Gravitational-Wave Observatory (LIGO) project. Scientists in LIGO are searching for evidence of gravitational waves using interferometry. The instruments are sensitive to the slightest perturbations around the instrument (e.g., small earthquakes or nearby vehicles) as well as noise created by interactions within the detectors [32]. These glitches in the data can mask detections and mapping them can help LIGO scientists search for and sometimes eliminate those glitches. The classification interface is shown in Figure 1 (bottom).

The Gravity Spy project has several advanced functionalities not found in Higgs Hunters. Gamification elements are present, since participation is scaffolded, with volunteers able to advance to more challenging levels with more glitch classes (currently there are 5 levels). Access to higher levels is granted by a promotion algorithm that assesses volunteer accuracy on gold standard data (i.e., data for which the system knows the correct response). The system, however, does not require that volunteers be logged-in to see gold data. When volunteers classify gold standard data anonymously, those answers are not used as input for the promotion algorithm, which can delay promotion for some volunteers.

4 METHODS

Data originated from classifications submitted to Higgs Hunters and Gravity Spy. The data in Higgs Hunters was from November 18, 2014 to June 20, 2015 (214 days) and Gravity Spy from October

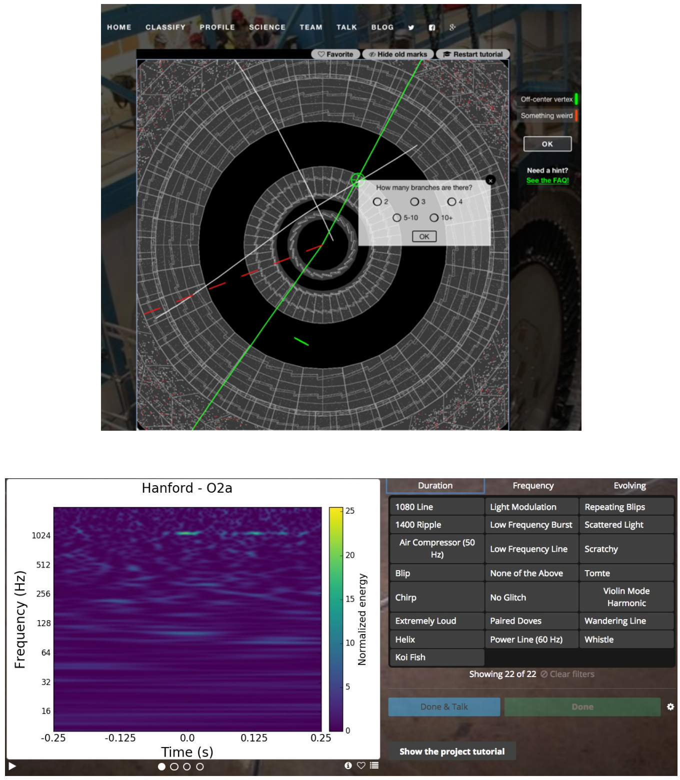

Fig. 1. The Higgs Hunters (top) and Gravity Spy (bottom) classification interfaces. In Higgs Hunters, volunteers are asked to search the images for decay anomalies or appearances of off-centre vertex lines, which are indications of new particles created from the decay of unseen ones. In Gravity Spy volunteers are asked to mark the similarities between a set of glitch options and the spectrogram in on the left.

12, 2016 to April 27, 2017 (197 days). As volunteers classify data, the system records the user ID of the classifier, IP address, response, and time. For records submitted anonymously the user ID field is blank. The Higgs Hunters dataset contained 793,188 classifications and Gravity Spy dataset contained 1,663,093. Gravity Spy underwent a beta testing phase where the project was open only to a selected pool of experienced volunteers from other projects on the Zooniverse platform, we removed records of volunteers who classified during the beta testing period. We argue that volunteers in the beta phase experience the project differently than volunteers who started after the launch so their behaviors could be different. In total, 318,079 classifications (86,988 occurred prior to launch and 231,091 after the launch date) were removed from the dataset.

Our study was approved by our University's Institutional Review Board (IRB). We note that identifying the source of online contributions (i.e., "doxing") is generally quite controversial. However, in this study, we only connect contributions to a user ID, not to a known individual and no data about individual volunteers are revealed in this paper or other publications. Furthermore, the content of the contributions (a classification of a data object) is unlikely to be controversial and the data are pre-existing and non-identified (i.e., with user IDs rather than real names). Finally, while the identification we describe could be problematic in another setting, it is also easily circumvented, e.g., by using a VPN or a service such as Tor. In summary, while we do not minimize the risk of studying anonymous work in general, there is minimal risk to volunteers from this study.

4.1 Data Pre-processing

To link anonymous classifications to a registered account, we adopted the approach outlined in [22]. We labeled classifications as *Logged-In* if the classification was submitted while the volunteer was logged-in, *Identified* if the IP was associated with one and only one user ID, *Ambiguous* if the classification was submitted from an IP address associated with more than one user ID, and *Anonymous* if the IP address was not associated with any user ID. To attribute anonymous classifications (those *Identified*) we determined a possible connection between a user ID and an IP address from the Logged-In classification records. If an IP address was associated with only one user account we simply appended the user ID to the classification record when the user ID field was blank.

Once classifications were sorted, we grouped the classification records of each volunteers into sessions, the set of classifications that seemed to be performed by a volunteer in a single sitting. Session boundaries were determined by looking for larger gaps between sequential classification records. The intuition is that volunteers typically come to the system, perform some number of activities separated by a short gap over a some period of time (a work session) and then quit until later, e.g., until the same time the next day, leaving a larger gap between the activity at the end of one session and the start of the next. Sessions are interesting because they indicate an ongoing level of contribution to a project. Following Mao et al. [17], we define a session as the sequence of classification records separated by less than 30 minutes. If the gap between two classifications is greater than 30 minutes, we mark the beginning of a new session. Next, we defined a time spent classifying variable by computing the time between the current and next classification. In aggregate, this variable represents an approximate amount of time volunteers spent reviewing the data.

4.2 Data Analysis

First, we describe the datasets and note the patterns of contribution found by the group of volunteers who contribute anonymously. We compared the contribution statistics using two datasets, the first with only Logged-in classifications (henceforth, referred to as D1) and the second adding Identified anonymous classifications (henceforth, referred to as D2). We compared the distribution of the number of classifications that were contributed per volunteer, time spent classifying and the number of sessions in the two datasets. The non-parametric Wilcoxon signed-rank test was used for testing the significance of differences in the datasets. The Wilcoxon signed-rank test is an alternative to t-test for matched pairs and assumes the data are not normally distributed (a feature of our data).

Second, we address our main research question by presenting two scenarios in which we use system log data to draw conclusions about volunteers. These cases are used a evidence that excluding anonymous events can lead researchers to contrasting conclusions and alternate algorithmic outcomes. For the first comparison, we address a reoccurring question in the CSCW community concerning whether first session activities can be used to predict future participation. We apply the approach outlined in Panciera et al. [22] to determine whether different conclusions are reached. In the second comparison, we examine the impact of anonymous traces on the promotion algorithm in Gravity Spy. All data were processed and analyzed using R and Python.

5 RESULTS

The contribution statistics for volunteers in both projects are shown in Table 1. Both projects have approximately the same number of users accounts (6,354 in Higgs Hunters and 6,761 in Gravity

Spy) and projects were active for a similar number of days (214 vs. 197). In contrast, the data show differences in the number of classification types with respect to logged-in and non-logged in classifications as Gravity Spy has almost twice as many classifications (1,345,014) as Higgs Hunters (793,188).

	Higgs Hunters	Gravity Spy
Volunteers		
IPs	22,507	15,188
User IDs	6,354	6,761
Classifications	**793,188**	**1,345,014**
Logged-in	684,087 (86.3%)	1,246,351 (92.6%)
Identified	27,075 (3.4%)	11,697 (0.9%)
Anonymous	80,357 (10.1%)	69,851 (5.2%)
Ambiguous	1,668 (0.2%)	17,115 (1.3%)

Table 1. The break down of volunteers and classification in the two datasets we analyzed.

5.1 Patterns of Anonymous Contributions

In this section, we report on several patterns that emerged from our analysis. First, we find differences in the proportion of anonymous classifications. There were 109,101 (13.7%) anonymous contributions in Higgs Hunters and 98,664 (7.4%) in Gravity Spy. We were able to identify 27,075 anonymous classifications in Higgs Hunters and 11,697 in Gravity Spy. In terms of raw numbers and as a percentage of the total number of classifications, Gravity Spy had fewer anonymous classifications. These differences are likely the result of the scaffolded participation described in section 3. Anonymous users are only shown level 1 of the system with two glitch classes; in order to access the more interesting higher levels, volunteers must be logged in. These results suggests the design of user experience has an impact on the quantity of anonymous contributions in a system.

Second, when we assigned the anonymous classifications, we found 3,112 (49%) volunteers who had contributed anonymously in Higgs Hunters and 1,212 (22%) in Gravity Spy. In Higgs Hunters, adding identified anonymous classifications increased the average number of classifications per volunteer by 8.7 (σ = 21.1); the median increase was 5. The average increase in Gravity Spy was 9.65 (σ = 18.53), and the median increase was also 5. We note that these are conservative estimates of the impact of including anonymous work, as we were able to identify only 25% (Higgs Hunters) and 12% (Gravity Spy) of anonymous contributions.

Third, as noted in the review of literature, exploring a community is an important activity in newcomers' early interactions with a system. We find that 79% of anonymous classifications occur during a volunteer's first session. Nevertheless, anonymous classifications can be found throughout volunteers' tenure in a project. In Higgs Hunters, 547 (17.6%) volunteers contributed anonymously after the first session, totaling 5,682 classifications; in Gravity Spy, 210 (17.3%) volunteers contributed anonymously, totaling 2,386 classifications. Finally, anonymity appears to be a regular mode of participation for some newcomers, as 370 volunteers in Higgs Hunters didn't register until their second session, 58 in their third, and 16 in their fourth session or later. In Gravity Spy, 119 volunteers had their first logged-in session only during their second session and 6 during their third session or later. These results show that anonymous classifications were missing from the contribution history of a sizable number of volunteers.

5.1.1 Shifting Data Distributions. Next, we examine how attributing anonymous classifications to a volunteer's history changes statistical conclusions about the community's production. We compared the distributions of classifications, time spent classifying and number of sessions per volunteer. The analysis was conducted using the Wilcoxon signed-rank test to statistically compare two data distribution (shown in Figure 2). While the counts of classification and time spent in the second dataset are always equal or larger for any volunteer (since the quantity of classifications and time spent classifying can only increase), variation exists among volunteers, so it could be that the changes due to adding identified classifications are insignificant compared to the natural variation in the data.

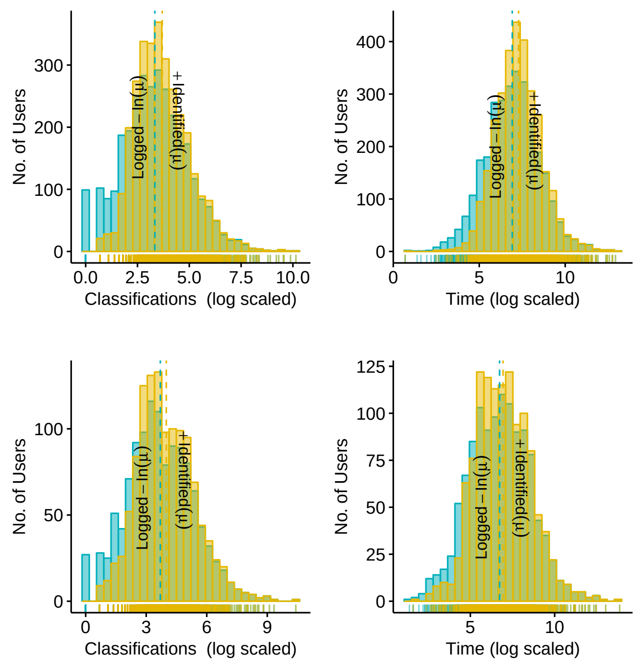

Fig. 2. The distributions of number of classifications and classifying time for Higgs Hunters (top) and Gravity Spy (bottom). The distribution in the foreground (yellow) represents data from D2 (logged-in and identified data), while the D1 (only logged-in data) is represented in the background in blue. Both variables are log-transformed to correct for skew.

Classifications. The first comparison uses the total population of volunteers (both those who did and did not contribute anonymously). When comparing the distributions of classifications the Wilcoxon signed-rank test indicated classifications were significantly higher in D2 than in D1 in Higgs Hunters (D1 μ= 108, $\sigma = 701$, D2 μ= 112, $\sigma = 699$ at Z = 21479000, $p < 0.001$). In Gravity Spy, classifications were higher in D2 (μ= 186, $\sigma = 727$) than in D1 (μ= 184, $\sigma = 727$), however, the difference was non-significant (Z = 22442000, $p < 0.06$). Next, we compared only the population of volunteers who contributed anonymously. In the Higgs Hunters D1 dataset, the average volunteer made 136 ($\sigma = 797$) classifications compared to 145 ($\sigma = 799$) in D2, an average increase of 8.7 classifications. The comparison of mean ranks between the two samples shows significant differences at Z = 4215800, $p < 0.001$. In Gravity Spy, the mean rank of classifications between D1 (208, $\sigma = 1148$) and 218 ($\sigma = 1158$) in D2 was also significant (Z = 657440, $p < 0.001$).

Classifying Time. Next, we wanted to see how much additional time volunteers were spending classifying when anonymous classifications are included. Again, using the total population we found a statistically significant difference for Higgs Hunters, but not for Gravity Spy. The Higgs Hunters D1 and D2 datasets had means of 58 mins. (σ = 6 hrs.) vs. 61 mins. (σ = 6 hrs.) and the mean for D2 was significantly higher (Z = 21695000, $p < 0.001$). The mean increase in time for Gravity Spy was 1 minute, 74 mins. (σ = 6 hrs.) vs. 75 mins. (σ = 6 hrs.), but Z = 22552000, $p = 0.18$ non-significant. Using the subset of volunteers who have contributed anonymously points to significantly higher means in both Higgs Hunters (Z = 4002700, $p < 0.001$) and Gravity Spy (Z = 666000, $p = 0.0101$). The difference in time was +5 minutes and +1 minute in Higgs Hunters and Gravity Spy, respectively.

We also compared the number of sessions in each dataset, since new sessions could emerge as a result of entirely anonymous sessions. For instance, if a volunteer classified anonymously in a single session and returned the next day to create an account (using the same IP address), grouping by user ID would omit the anonymous sessions from their contribution history. However, grouping classification records by IP address means both sessions would be attributed to the volunteer. By using IP addresses we found 485 new sessions in Higgs Hunters and 112 sessions in Gravity Spy. In summary, omitting anonymous traces leads to a statistically significantly different datasets for volunteers who contribute anonymously. And even though only a fraction of volunteers contribute anonymously, when comparing the distribution for all volunteers, differences are still statistically significant in the project in which we were able to identify a sufficient fraction of the anonymous work.

5.2 Case studies on the impacts of omitting anonymous work

The previous analysis shows that in many cases including anonymous classifications significantly alters descriptive statistics. In this section, we present case studies of analyses conducted using user data and how evidence could be altered when anonymous traces are considered. First, we consider an often discussed topic in on-line communities, that is, the extent to which a user's first session activities can predict future activity. Second, as a practical example of how user experience could be altered, we evaluate the results of an algorithm used to promote volunteers to levels in Gravity Spy to determine whether volunteers experience delays in promotion when their anonymous classifications are not attributed to their user ID.

5.2.1 Predicting future participation from early activity. An ongoing area of conversation in research on user behavior in online communities is whether an individual's future contribution can be predicted based on early activities. Some studies suggest that users are consistent in the types of behaviors and level of engagement exhibited [14, 22]. For instance, Panciera et al. [22] found that Wikipedians who contribute many edits in their first and second sessions have much higher likelihood of becoming a power editor (defined as making more than 250 edits). Understandably given the challenges associated with such studies, these analyses did not accounted for anonymous work. Thus, it is possible that users are engaged with the system before they become fully visible in the data and would look different if anonymous traces could be examined.

To examine this possibility, we conducted the same analysis as Panciera et al. [22], dividing citizen science volunteers into strata based on the quantity of contributions (in our case classifications) during their first session and then computed the proportion that those volunteers would become "power classifiers" (i.e., > 250 classifications, as Panceria et al. used > 250 edits).

The results of our analysis are presented in Table 2. As in Panciera et al. [22], both Gravity Spy and Higgs Hunters show small and somewhat gradual increases in the likelihood that a volunteer will become a power classifier as the number of classifications contributed in the first session

Classifications in First Session	No. of Volunteers	Proportion of Volunteers	No. Becoming Power Classifiers	Likelihood of Power Classifiers
First Session Classifying in Higgs Hunters				
Logged-In Classifications				
1	127	0.04	2	0.02
2-3	229	0.08	4	0.02
4-5	223	0.07	3	0.01
6-10	434	0.15	5	0.01
11-20	578	0.19	17	0.03
21-40	570	0.19	20	0.04
Over 40	832	0.28	177	0.21
Logged-In & Identified Classifications				
1	108	0.04	13	0.12
2-3	132	0.04	5	0.04
4-5	127	0.04	9	0.07
6-10	399	0.13	11	0.03
11-20	659	0.22	22	0.03
21-40	662	0.22	15	0.02
Over 40	907	0.30	172	0.19
First Session Classifying in Gravity Spy				
Logged-In Classifications				
1	40	0.04	1	0.02
2-3	69	0.06	1	0.01
4-5	64	0.06	1	0.02
6-10	133	0.12	6	0.05
11-20	215	0.19	10	0.05
21-40	216	0.19	12	0.06
Over 40	400	0.35	108	0.27
Logged-In & Identified Classifications				
1	26	0.02	7	0.27
2-3	56	0.05	5	0.09
4-5	43	0.04	2	0.05
6-10	101	0.09	3	0.03
11-20	227	0.20	12	0.05
21-40	265	0.23	15	0.06
Over 40	419	0.37	99	0.24

Table 2. Likelihood of power user status. We adopted the analysis approach in Panciera et al. [24] where evidence for users being born versus made was evidence of proportions of users in different edit strata becoming power users (more than 250 edits). In our dataset, excluding anonymous classifications leads to over-estimates of power classifiers status based on first session activity.

increases. In Higgs Hunters, for example, volunteers making more than 40 classifications during their first session have a .21 likelihood of becoming a power classifier, while those making only one classification have a less than .02 chance. Similar statistics are found for Gravity Spy.

Including anonymous classifications (i.e., D2) yields different results. The number of volunteers making more than 40 classifications in their first session in Gravity Spy decreases by nine (i.e., 9 of these volunteers had earlier sessions that were completely anonymous) and the likelihood of being a power classifier decrease by .03. The most striking difference is in the large increase in the likelihood of becoming a power contributor for volunteers contributing only one classification in their first session. A similar change was observed in Higgs Hunters. In other words, while it is still the case that those contributing a lot in their first session are likely to be major contributors, so are a sizable proportion of those contributing only one classification in their first (possibly anonymous) session.

5.2.2 Delays in Promotion. As a second case study, we consider the impact of including anonymous work on decisions made by the promotion algorithm in Gravity Spy (see description in Section 3). In short, if classifications used to assess volunteer performance in an algorithm are omitted, a volunteer's promotion could be delayed. Since volunteers can only contribute anonymously in Level 1 we examine promotion from Level 1 to Level 2. There were 7,136 volunteers who contributed at Level 1, of whom 3,336 (47%) were promoted to Level 2, and 388 (10.1%) contributed classifications while not logged-in that could be linked to their user account using IP address matching. The remainder of the analysis focuses on this subset of volunteers.

Gravity Spy Session Contribution Statistics	
Logged-In	
Classifications	40.94 (σ = 101.56)
Gold Classifications	16.35 (σ = 40)
Anonymous	
Classifications	12.52 (σ = 15.9)
Gold Classifications	5.37 (σ = 7.44)
Time to Promotion (hours)	
Logged-In	156 (σ = 586.423)
Logged-In & Anonymous	12 (σ = 48.56)

Table 3. Average Level 1 contribution statistics for the volunteers in Gravity Spy who were promoted to Level 2 *and* had identified anonymous classifications.

Current Promotion. On average, the 388 volunteers made 40.94 (σ = 101.56) classifications and saw 16.35 (σ = 40) gold images (Figure 3) in Level 1. The median time to promotion considering only logged-in classifications was 370 minutes (or 6.1 hours). We report median as well as mean values because the data are quite skewed, and the mean is impacted by volunteers who contribute sporadically. For example, one volunteer contributed four classifications (not enough to be evaluated and promoted to Level 2), but didn't have their second session until five months later, at which time they were promoted.

Hypothetical Promotion. When analyzing dataset D2, we found volunteers who had identifiable anonymous contributions made on average 12.52 anonymous classifications and 5.37 (σ = 7.44) anonymous gold classifications. We computed when a volunteer would have been promoted had their anonymous work been taken into account by assuming that volunteers would have been promoted after the same number of gold classifications. Work done after that point is therefore above what should have been necessary to be promoted. Figure 3 shows the number of classifications a volunteer contributed in Level 1 after the hypothetical promotion when taking anonymous classifications into account. On average, volunteers contributed 41.1 (σ = 70.12) additional classifications (including gold and non-gold subjects) after the hypothetical promotion point. Figure 3 also shows the additional time volunteers spent classifying after their hypothetical promotion. On average, volunteers spent an additional 143.66 hours. Again, as means are sensitive to outliers, the median is

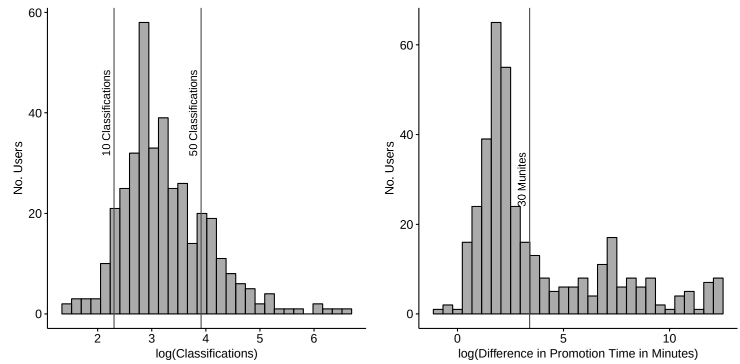

Fig. 3. A histogram showing the number of additional classifications volunteers contributed after their hypothetical Level 2 promotion date. The hypothetical promotion date reflects when a volunteer would have been promoted had anonymous gold classifications been linked to their account. The number of the x-axis is displayed on a log scale. The additional time volunteers spent classifying in Level 1 after their hypothetical promotion date. The number of the x-axis is displayed on a log scale.

more representative. When anonymous classifications were added to the dataset the median time to promotion decreases by approximately one hour (to 311.75 minutes = 5.2 hours).

5.2.3 Classification Accuracy. In assessing the point of hypothetical promotion, we assumed that volunteers would classify gold standard data correctly. We examined the classifications to validate this assumption. Level 1 is intended to introduce newcomers to the system and so is not a challenging level. Accuracy on logged-in Level 1 gold standard data is high (98%) and it is equally high for the anonymous classifications we were able to link to volunteers (98.3%). The results of the paired t-test revealed non-significant differences at t(382) = 0.564, $p = 0.573$ indicating that anonymous classifications are not of a worse quality than logged-in classifications.

6 DISCUSSION

The results presented above emphasize the importance of anonymous traces for understanding user behavior. This research makes several important findings about anonymous classifications in citizen science. First, anonymous classifications comprise a small, but non-trivial volume of work in citizen science projects as volunteers' contributions increased by approximately 10 classifications. Anonymous classifications comprised 13% and 7.5% of all classifications in Higgs Hunters and Gravity Spy respectively. Approximately one-half of volunteers in Higgs Hunters and one-quarter in Gravity Spy contribute anonymous classifications. This might be explained by the varied amount of publicity given to projects. Higgs Hunters is a project frequently mentioned in Zooniverse advertising campaigns, while Gravity Spy doesn't receive similar attention. As a result, there are many more "dabblers" [27] in Higgs Hunters (7,319 more unique IPs than Gravity Spy); volunteers who check out the project but don't continue. We also found anonymous activity is not limited to initial interactions with the system as 17% of volunteers contributed anonymously after their first session.

Second, our findings provide evidence that aggregating data using only logged-in traces results in underestimations of the volume of classifications and time spent analyzing data. For instance, when analyzing the population of anonymous contributors the mean ranks of classifications and time spent classifying was significantly higher in D2. However, when comparing datasets using the total population of volunteers, the mean rank difference was only significant in Higgs Hunters, we suspect these results are influenced by variances in system design. In Gravity Spy, participation is scaffolded where volunteers need to log-in to participate in Level 2 and above. As a result, a smaller percentage of classifications in Gravity Spy (7.4%) were anonymous when compared to Higgs Hunters (13.7%). Thus, the issue of anonymous traces might be mitigated by scaffolding participation or including elements of gamification to encourage volunteers to login or create an account.

The case studies provide the most substantial evidence for why anonymous traces are important for research and practice. First, we showed that omitting anonymous classifications could cause evidence to become incompatible with the assumptions of a theory. When applying Panciera's [22] analytic approach to predicting the likelihood of become a "power editor" several interesting changes emerged when comparing D1 to D2. Most obvious little support emerge for early activity predicting future participation. For instance, our analysis showed users who had only one classification with D1 (only logged-in classifications) was underreported since they were contributing without registering for an account. Once we linked anonymous classifications this pushed many volunteers into higher classification strata. Additionally, some sessions were included in D2 that were absent in D1. This is because some volunteers contributed anonymously during their first interaction and did not register until their second visit. Thus, aggregating by user ID excludes the initial session. As a result, we see more volunteers who contributed one classification in their first session becoming power classifiers.

Second, we showed that algorithms designed to govern user experience could cause delays in granting access to system functionality. The results showed that in Gravity Spy, a small number of Level 2 volunteers contributed anonymous classifications in Level 1. Not linking these classifications to their user IDs delayed their promotion by a median of one hour and as a result caused them to contribute on average 41 additional classifications before being promoted. For volunteers, promotion can impact motivation resulting in an increased sense of achievement. Additionally, promotion brings about work that is more challenging and complex, a known motivator in work design. If promotion and thus access to more rewarding activities is delayed, volunteers might find the task overly monotonous, become disengaged and leave the community. For the community, specifically the scientists who rely on data classified by volunteers, the delay in promotion reduces the number of volunteers who can classify images in Level 2, where more challenging work exists. The delayed volunteers contributed 15,453 classifications in Level 1 that could have been contributed in Level 2. Level 2 introduces additional classes of glitch, so the delay also means delays to processing those additional classes of glitch.

6.1 Anonymous Contributions in Other Production Communities

Our study has been set in the context of online citizen science, nevertheless, we expect that anonymous data not being missing at random may generalize to many online communities.

First, anonymous work is not a rare phenomenon. While anonymous contributions have been discussed in only a handful of studies, a significant portion of the content on sites allowing anonymous contributions may be generated by users that either do not have an account or are not logged into their account when contributing. For example, in the English version of Wikipedia, approximately 100,000 anonymous editors make at least one edit a month, and currently account for about 13% of persisting words contributed[1]. The 2011 Wikipedia Survey found that 59% (N=6,657) of users in Wikipedia made anonymous edits and 20% contributed between 11 and 50 edits anonymously.

Second, the prior literature about newcomer behavior in online communities suggests that anonymous work is not taking place at random but rather prevalent at specific points in a user's

[1]https://meta.wikimedia.org/wiki/Research:Measuring_edit_productivity

engagement with a community. Anonymous work may also be more likely for certain classes of users. For example, Anthony's [1] research suggests that expert participants care less about reputation and thus might be more likely to work anonymously.

Finally, many studies of communities like Wikipedia or free/libre open source software developers rely on quantitative analysis of trace data to understand user behavior, e.g., [5, 6, 22, 23]. For instance, Burke et al. [5] collected the edit histories of users on English Wikipedia who submitted request for administrator status (RfA) applications to predict the success of the request. While it may be possible to make accurate predictions based only on visible data, interpretations of these models about user behavior may be misleading if the data used to build them do not include anonymous work.

While our findings stress the importance of taking anonymous contributions into account in user studies, we do not argue that the results of our case study will translate directly to the findings of studies such as Panciera et al. [22]. For example, if expert users are more likely to be anonymous, including anonymous work might actually increase the number of top initial contributors. More studies are needed in other online communities detailing the impacts of anonymous traces on research and practice.

6.2 Limitations

We acknowledge that the strategy for attributing anonymous work to users has limitations. First, the data are based on user IDs and IP addresses. Regarding user IDs, we presume that most volunteers use a single ID, but cannot rule out users having multiple IDs or multiple users using a single ID, though we do not believe that either situation is common. Regarding IP addresses, our strategy for assigning anonymous work assumes that volunteers contribute regularly from a computer with a single IP. But a single volunteer may have multiple IP addresses, e.g., someone contributing from multiple locations, from a mobile device or through a system such as Tor. In the Gravity Spy, 789 volunteers had more than one IP associated with their account. If such a user contributes from yet another unique IP and does not logged in, there is no way to attribute those classifications to their account, leading to an over-estimate of the number of anonymous volunteers in the project. Conversely, in the data we have 74 IP address that were used by multiple volunteers, making it impossible to attribute anonymous work from those IP addresses.

The worst case for our analysis would be multiple users contributing from a single IP address but where only one user logs in, leading to an over-estimate of that user's anonymous work. Such a situation is imaginable: e.g., a classroom behind a NAT with a single IP address where the teacher has an account and the students contribute anonymously. However, we do not believe such situations are common. Further, we use robust statistical analysis techniques that are not sensitive to outliers, so even if there were a few such cases, they should not alter our conclusions.

Second, we were able to attribute only a fraction of the anonymous classifications to a user, raising the concern that the classifications we did attribute might not be representative of anonymous traces in general. For instance, if a user only contributes anonymously via their mobile device and never logs in from that device our sample might exclude the unique ways in which user behaviors manifest on mobile devices. However, we cannot test this possibility directly.

7 CONCLUSIONS

We conclude with some implications for research and for practice. Given the results described above, researchers should state explicitly the ability of users to contribute anonymously and (1) include anonymous activities if possible (perhaps in collaboration with the system operators) or (2) indicate how their research might be impacted by omitting anonymous contributions. If anonymous traces are accessed, researchers should consider the ethical implications of including them in a study.

While no foreseen harm is expected to come from coupling anonymous traces in exploring citizen science classification data, there are settings where users intentionally remain anonymous to avoid disclosing sensitive data [3, 11].

For practice, system operators may consider how users "claim" prior contributions made anonymously. The Zooniverse system prompts users to log-in after a few anonymous classifications, but does not retrospectively attribute those classifications. Previous versions of Wikipedia allowed users to re-attribute edits to a single user account, however this feature ceased to exist in 2005. Several proposals have been suggested to handle anonymous edits; one workaround is for users who contribute from a static IP to create a new editor page (or subpage) and manually copy their list of edits. However, if a user contributes from many IPs, it may take time to assemble the list[2]. Other systems could consider allowing similar functionality for users so their work can be attributed to their user accounts.

Finally, our work suggests several avenues for future research. Research might explore alternative approaches to assign anonymous work to a user. For example, [29] found browser characteristics (e.g., browser version, plug-ins, screen size, etc.) of individuals were consistent across sessions. These browser characteristics could be used to assign anonymous activities in other online settings, though requiring researchers to consider new user privacy and research ethics concerns. Our findings also point to the need for additional research examining the role and characteristics of anonymous events across a range of online communities. Additional knowledge is needed to understand what factors influence users to contribute anonymously. In summary, this research shows the importance of more precise accounting of anonymous events in online communities.

ACKNOWLEDGMENTS

We thank the citizen science volunteers and Zooniverse team for access to data. The work is supported by the U.S. National Science Foundation INSPIRE under Grant No.: 15-47880.

REFERENCES

[1] Denise Anthony, Sean W Smith, and Tim Williamson. 2007. The quality of open source production: Zealots and good Samaritans in the case of Wikipedia. *Rationality and Society* (2007).

[2] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. 2015. Functional Roles and Career Paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1092–1105. https://doi.org/10.1145/2675133.2675257

[3] Michael S Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. 1–8.

[4] Rick Bonney, Caren B Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V Rosenberg, and Jennifer Shirk. 2009. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience* 59, 11 (2009), 977–984.

[5] Moira Burke and Robert Kraut. 2008. Mopping Up: Modeling Wikipedia Promotion Decisions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. 27–36.

[6] Moira Burke and Robert Kraut. 2008. Taking Up the Mop: Identifying Future Wikipedia Administrators. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems (CHI EA '08)*. 3441–3446.

[7] Moira Burke, R Kraut, and E Joyce. 2010. Membership Claims and Requests: Conversation-Level Newcomer Socialization Strategies in Online Groups. *Small Group Research* 41, 1 (Jan. 2010), 4–40.

[8] Boreum Choi, Kira Alexander, Robert E. Kraut, and John M. Levine. 2010. Socialization Tactics in Wikipedia and Their Effects. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 107–116. https://doi.org/10.1145/1718918.1718940

[9] Alexandra Eveleigh, Charlene Jennett, Ann Blandford, Philip Brohan, and Anna L. Cox. 2014. Designing for Dabblers and Deterring Drop-outs in Citizen Science. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2985–2994. https://doi.org/10.1145/2556288.2557262

[2] https://en.wikipedia.org/wiki/Wikipedia:Changing_attribution_for_an_edit

[10] R Stuart Geiger and David Ribes. 2011. Trace ethnography: Following coordination through documentary practices. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. 1–10.

[11] Matthew Heston. 2016. (In)visible Cities: An Exploration of Social Identity, Anonymity and Location-Based Filtering on Yik Yak. In *iConference 2016*. iSchools, 1–10.

[12] Corey Jackson, Carsten Østerlund, Veronica Maidel, Kevin Crowston, and Gabriel Mugar. 2016. Which Way Did They Go?: Newcomer Movement Through the Zooniverse. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. 624–635.

[13] Caroline Jay, Robert Dunne, David Gelsthorpe, and Markel Vigo. 2016. To Sign Up, or Not to Sign Up?: Maximizing Citizen Science Contribution Rates Through Optional Registration. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1827–1832. https://doi.org/10.1145/2858036.2858319

[14] Nicolas Jullien. 2015. Big from the beginning: Assessing online contributors' behavior by their first contribution. *Research Policy* 44, 6 (July 2015), 1226–1239.

[15] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Phylo players, Luis Sarmenta, Mathieu Blanchette, and Jérôme Waldispühl. 2012. Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. *PLOS ONE* 7, 3 (03 2012), 1–9.

[16] Markus Luczak-Rösch, Ramine Tinati, Elena Paslaru Bontas Simperl, Max Van Kleek, Nigel Shadbolt, and Robert J. Simpson. 2014. Why Won't Aliens Talk to Us? Content and Community Dynamics in Online Citizen Science. *At Eighth International AAAI Conference on Weblogs and Social Media Eighth International AAAI Conference on Weblogs and Social Media*.

[17] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why Stop Now? Predicting Worker Engagement in Online Crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[18] Gabriel Mugar, Carsten Østerlund, Katie DeVries Hassman, Kevin Crowston, and Corey Brian Jackson. 2014. Planet Hunters and Seafloor Explorers: Legitimate Peripheral Participation Through Practice Proxies in Online Citizen Science. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 109–119. https://doi.org/10.1145/2531602.2531721

[19] Michael Muller. 2012. Lurking As Personal Trait or Situational Disposition: Lurking and Contributing in Enterprise Social Media. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 253–256. https://doi.org/10.1145/2145204.2145245

[20] Blair Nonnecke and Jennifer Preece. 2000. Lurker demographics: counting the silent. In *CHI '16: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 73–80.

[21] Blair Nonnecke and Jennifer Preece. 2004. Shedding Light on Lurkers in Online Communities. *Ethnographic Studies in Real and Virtual Environments Inhabited Information Spaces and Connected Communities*. (Feb. 2004), 1–7.

[22] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work (GROUP '09)*. ACM, New York, NY, USA, 51–60. https://doi.org/10.1145/1531674.1531682

[23] Katherine Panciera, Mikhil Masli, and Loren G Terveen. 2014. CrèMe De La CrèMe: Elite Contributors in an Online Community. In *Proceedings of The International Symposium on Open Collaboration (OpenSym '14)*. ACM, New York, NY, USA, Article 21, 10 pages. https://doi.org/10.1145/2641580.2641609

[24] Katherine Panciera, Reid Priedhorsky, Thomas Erickson, and Loren Terveen. 2010. Lurking? Cyclopaths?: A Quantitative Lifecycle Analysis of User Behavior in a Geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1917–1926. https://doi.org/10.1145/1753326.1753615

[25] Jennifer Preece, Blair Nonnecke, and Dorine Andrews. 2004. The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior* 20, 2 (March 2004), 201–223.

[26] Neal Reeves, Peter West, and Elena Simperl. 2018. " A game without competition is hardly a game": The impact of competitions on player activity in a human computation game. *In Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing* (2018).

[27] Dana Rotman, Jennifer Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. 2012. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 217–226.

[28] Robert Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: Observing the World's Largest Citizen Science platform. In *Proceedings of the 23rd conference on the World Wide Web*.

[29] Jessica Su, Ansh Shukla, Sharad Goel, and Arvind Narayanan. 2017. De-anonymizing Web Browsing Data with Social Networks. In *Proceedings of the 26th International Conference on World Wide Web*. 1261–1269.

[30] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (Oct. 2009), 2282–2292.

[31] Andrea Wiggins and Kevin Crowston. 2011. From Conservation to Crowdsourcing: A Typology of Citizen Science. In *Proceedings of the 2011 44th Hawaii International Conference on System Sciences (HICSS '11)*. IEEE Computer Society,

Washington, DC, USA, 1–10. https://doi.org/10.1109/HICSS.2011.207

[32] Michael Zevin, Scott Coughlin, Sara Bahaadini, Emre Besler, Neda Rohani, Sarah Allen, Miriam Cabero, Kevin Crowston, Aggelos Katsaggelos, Shane Larson, Tae Kyoung Lee, Chris Lintott, Tyson Littenberg, Andrew Lundgren, Carsten Oesterlund, Joshua Smith, Laura Trouille, and Vicky Kalogera. 2017. Gravity Spy: Integrating Advanced LIGO Detector Characterization, Machine Learning, and Citizen Science. *Classical and Quantum Gravity* 34 (2017), 064003.

Received April 2018; revised June 2018; accepted September 2018