Working in the Shadows: Anonymous Contributions by Users in Citizen Science

Corey Brian Jackson, Kevin Crowston, Carsten Østerlund
School of Information Studies, Syracuse University, Syracuse, NY, 13210, USA
{cjacks04, crowston, costerlu}@syr.edu

ABSTRACT
Researchers studying the behaviour of users of online community systems often base their analyses on the logs of activities captured by the system (e.g., the record of a comment post or of an edit). The history of users' interactions can reveal, for example, how users move from novices to experts and the steps they take as they learn to contribute to the community. However, some systems allow users to contribute without logging in or even having an account. Since these anonymous events are not associated with a particular user in the logs, they are generally not included in analyses of user behaviour. Omitting anonymous events may bias findings of studies of user behaviour related to trajectories of membership or learning. We investigated the characteristics of anonymous work in an online citizen science project. Our analysis suggests that at least 50 percent of users with accounts also contributed anonymously, for an average of 8.9 anonymous events.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
Online communities, anonymous work

INTRODUCTION
Understanding users' interaction with systems and how their behaviors change over time is an important area for computer human interaction (CHI) and computer-supported cooperative work (CSCW). There are many methodological options for studying a single user and system, such as lab experiments, observation or interviews of the user. However, many of these options are difficult or impossible to apply to online community systems with large number of users whose use is spread out geographically and temporally, and their application often leaves concerns about the representativeness of the sample of users included in the study.

An attractive alternative for studies of large-scale systems is to analyze the data collected by the system that record user interactions on the site. For example, a system that supports commenting will record the text of the comment, the user name of the individual who posted the comment, a time stamp indicating the precise date and time the comment was posted and perhaps other metadata. Other kinds of interactions are similarly recorded, creating an extensive interaction trail for each user. In aggregate, these data allow researchers to analyze a rich record of how users interact online.

The potential of such trace data [11] has not gone untapped. Models of users' interaction have relied heavily on the traces of activity captured in system logs. In Wikipedia for example, Burke et al. [6, 7] used historical information such as article edits and reverts to predict users' likelihood to be promoted to administrators. Other researchers have addressed questions such as socialization [5, 7, 8], engagement with site features [14, 15, 17, 22], or the impacts of contribution behaviors on the community [10, 12, 13].

However, systems often allow some amount of anonymous interaction. For example, in Wikipedia, visitors can read and search articles anonymously. Twitch, an online gaming platform, allows gamers to watch other gamers play games anonymously. Some systems even allow users to create content without registering for an account. In some Wikipedias, anonymous users can edit articles and some online discussion fora (e.g., 4chan and some Stack Overflow projects) allow users to post comments without having an account. As a result of these anonymous interactions, traces associated with a particular user ID may not capture the full interaction of that person with the system.

These omissions are problematic, as anonymous work does not happen randomly: rather studies suggest that it is more likely to include a user's initial interactions with a system, which prior studies suggest are particularly formative. First, part of learning about a system comes through anonymous observation of other's behaviours. Many authors [19–21, 24] point to lurking as a formative activity. Antin et al. [2] suggests reading is a gateway activity for newcomers to learn about the site. Preece and Shneiderman [25] also noted the importance of observing, describing the trajectory of reader of online content in a community to leader of the community. From interviews with experienced Wikipedia editors, Bryant [4] found that many newcomers begin contributing as passive consumers of content, reading articles and pages to understand community norms of participation and only later in their Wikipedia tenure do they start focusing on functional roles [3].

CSCW'18 TBD

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Initial contributions and the feedback received have also been shown to be valuable for understanding ascension to new roles in the community [3, 5–7], retention [18], and learning [9]. If an important period of a user's history is missing from the traces, then subsequent analyses of the data may lead to inaccurate and/or misleading conclusions about how users evolve.

While understanding what people do prior to creating an account is potentially important, studying the phenomena is challenging. To learn more about anonymous activities, we studied Higgs Hunters, an online citizen science project hosted on the Zooniverse citizen science platform [26]. The software developers who designed the Zooniverse platform sought to remove as many barriers to participation as possible; thus an account is not required to contribute. As a result, Zooniverse users contribute classifications (the primary work in the project) without an account or while not logged-in, i.e., anonymously.

The purpose of this research is to understand the characteristics of anonymous work and in particular, to assess the impact on findings of user studies of omitting anonymous work. Our first set of research questions are simply: a) how much work is done anonymously, b) in what pattern and c) does considering anonymous work lead to a significant difference in a users contribution statistics?

Our research makes two contributions:

- building on research Panciera et al. [23] who studied the anonymous work of Cyclopath volunteers, we assess the importance and volume of anonymous work in the Higgs Hunters project.

- we compare the differences in the the data distribution when anonymous activities are included in their profiles versus when anonymous work is excluded.

ANONYMITY ONLINE

Discussions of anonymous work is largely absent from the literature on user behavior in online communities, having been addressed in only a handful of studies. And yet, on sites that allow anonymous contributions, a significant portion of the content may be generated by users that either do not have an account or are not logged into their account when they contribute. 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 reported making anonymous edits and 20% contributed between 11 and 50 edits while anonymous.

A few studies have examined why online community members might prefer to remain anonymous [19]. Noting that registration was necessary to develop a reputation on a site, Anthony [1] argued that visitors who register for an account have different participation intentions than those who do not register. Specifically, they described intentionally anonymous contributors as good Samaritans and suggested they are in two groups: (1) experts who don't care about reputation and likely contribute only to articles in topic areas in which they have expertise, and (2) users who are simply correcting editing errors. Anthony et al. [1] also hypothesized that users in the first group would produce quality work and those in the second group were likely to have shorter and less substantive edit histories.

The benefits of anonymity for the community's production have also been noted. Jay et al. [16] found that removing the registration requirement in Manchester's Museum citizen science project increased the number of visitors who contribute by 62%. They therefore suggest that online communities should provide an option to register, but that it should not be a requirement, as registration creates a barrier to participation.

The above discussion has considered anonymity as deliberate and all or nothing, but in practice, anonymity may be only temporary. For example, a user may contribute anonymously for some time and then register for an account, or may have an account but contribute occasionally while not logged-in (accidentally or deliberately).

While it is challenging to study anonymous work, it is not impossible, because interactions on the Internet come from an IP address and that address may be unique to the user. Panciera et al. [23] studied Cyclopath, a geographic wiki site where users share information about bicycle routes by editing a map of bicycle routes, annotating content, and adding bikeability ratings for trails and road segments. By analyzing user IP addresses, the authors were able to link 20% percent of anonymous events with known user accounts [23].

SETTING: ZOONIVERSE AND HIGGS HUNTERS

We carried out our study using data from the Zooniverse Higgs Hunters project. Zooniverse [26] is a citizen science platform with more than 1.5 million volunteers. Since the site's launch, it has hosted more than seventy citizen science projects in fields ranging from astronomy to literature. Professional scientists use Zooniverse to crowdsource the analysis of data objects to citizen scientists who volunteer on the platform. Each projects has two primary interfaces: a page for classifying data objects and a page for posting comments, in addition to less-used collections, tutorials, profile pages and so on.

Unlike Wikipedia, Zooniverse projects do not offer much opportunity to read before contributing. Typically, once volunteers arrive at a project page they complete a short tutorial covering the science behind the project and instructions about submitting a classification. To avoid the decisions of one volunteer biasing another, volunteers can not see others' annotation decisions, and are only shown the discussion about an object after submitting their own annotations. As a result, volunteers mostly learn by doing.

In all Zooniverse projects, classifications can be done without logging-in. If the volunteers is not logged in, after submitting three annotations, a pop-up window appears asking the volunteer if they want to sign in to an existing account or register for a new account. At this point, volunteers can choose to log-in or sign up (in which case the work is attributed to their user ID) or close the pop-up and continue contributing

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

anonymously. Log-in credentials are stored in persistent cookies on the volunteers' browsers so they can be automatically logged-in when they return to the site.

While an account is not required to access the Zooniverse platform, having an account is beneficial for both the volunteers and the system. For the user, activities such as building collections of data objects, private messaging and posting comments on the Talk and discussion boards are only available to logged-in users. Collections are similar to bookmarks: users can mark items and access them later. Many volunteers use collections to remember data they find interesting or to support independent science investigations using the data on the site. Past research has suggested that these features are important for learning and motivation [14]. Additionally, having an account allows volunteers to monitor their contributions to the system on profile pages that show the annotation count for each project.

Having users log-in is also important for the functionality of the system. Zooniverse projects rely on multiple annotators to increase the reliability of classifications. For this reason, it is imperative the data objects not be annotated by the same user more than once. Without users logging-in, it is difficult to know if the same user is annotating the same object. Furthermore, there is interest in developing task assignment algorithms to route more complex and challenging data to volunteers who are the most accurate annotators. Such algorithms depend on having an accurate record of a user's contribution history.

Higgs Hunters

The particular project we studied was Higgs Hunters (https://www.higgshunters.org), a particle physics citizen science project launched in 2014 that helps physicists searching for exotic particles in data from the Large Hadron Collider. It is hosted on the Zooniverse platform and so shares the social and technical arrangements described in the previous section. In the system, volunteers are shown an image of a collision in which charged particles are represented as lines. Volunteers are asked to mark off-center vertices, which are indications of new particles being created from the decay of other unseen particles. A screen shot of the classification interface is shown in Figure 1.

METHODS

To investigate the behaviours of users in Higgs Hunters we designed our research as a trace ethnographic study. Trace ethnography [11] is a research approach that combines participant-observation and data collected from system logs for the purpose of investigating the experiences of users as they participate on computer systems. The traces of participant activity serve as historical records of user interactions as they perform tasks on the project site. By using trace ethnography we reconstruct the "lived experience" of volunteers as they contribute on the site allowing us to view how the users "were" at different periods in their history.

Collecting System Log Data

Our dataset consists of all annotations submitted from between November 18th 2014 and June 20th 2015, a total of 204 days.

Figure 1. The Higgs Hunters annotation interface. In Higgs Hunters, users 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.

When users perform activities in the project, such as annotating data objects, posting comments on discussion forums, or liking comments other volunteers post, a log record is generated with a user name (blank for anonymous contributions), IP address, time stamp and other meta-data associated with the content, such as a unique identifier for the event. These database records are stored on a server hosted by Zooniverse. We were provided access to the database dumps for this study.

Generating Datasets and Variables

We first created two datasets for the study. Data set one including only activities associated with a user ID and another including activities with and without a user ID. For data set two, we used the procedure from Panceria et al. [23] to attribute some anonymous work to a user based on IP address. The procedure categorizes activities as logged-in, identified, anonymous, and ambiguous. Logged-in events are those associated with a user name. Panceria et al. define the other three types as follows:

1. "If an IP co-occurred with precisely one known user, classify the IP as *Identified*, and assume that all events from that IP are due to that user regardless of whether the event was *Logged-In*."
2. "If an IP co-occurred with more than one known user, classify the IP as *Ambiguous*."
3. "Otherwise, if the IP co-occurred with zero known users, classify the IP as *Anonymous*."

Second, we grouped events into what we called a work session, the set of events that seemed to be performed in a single sitting [10]. Sessions are interesting because volunteers returning to a project is an indication of their commitment, complementing the count of annotations. Session boundaries were determined by looking for larger gaps between sequential activities. The

intuition is that users 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.

In Higgs Hunters, working on a single annotation should not require a lot of time. However, the gap between annotations could potentially extend beyond a few minutes if users post or read comments about the data object. For a study of users of Galaxy Zoo—another citizen science project hosted on the Zooniverse platform—Mao et al., [18] suggested that a gap of more than 30 minutes between activities signalled the start of a new session. Therefore, following Mao et al. [18], we define a session as the sequence of events separated by less than 30 minutes. If the gap between two events is greater than 30 minutes, we mark the beginning of a new session.

In the second dataset, we identified sessions for identified users including both logged-in and identified events, and also for anonymous users using anonymous events grouped by IP address (i.e., assuming that all anonymous events from a particular IP address are from a single anonymous user).

Finally, we examined the time between adjacent contributions to estimate how long people spent working on the classification tasks. The time is estimated as the duration of the session (second to last event). Since we do not know when the user began the first classification, we excluded that time from the session duration.

Dataset Limitations. We acknowledge that the strategy for attributing anonymous work to users has limitations. First, we are interested in studying users, but the data are based on user IDs and IP addresses. For the first, we presume that most users use a single ID, but can not rule out users having multiple IDs or multiple users using a single ID, though we do not believe that either situation is common.

For the second, our strategy for assigning anonymous work assumes that users contribute regularly from a computer with a single IP. But a single user may have multiple IP addresses, e.g., someone contributing from multiple locations, from a mobile device or through a system such as Tor. In our dataset, 789 users had more than one IP associated with their account. If a user contributes from a unique IP while not logged in, there is no way to connect those, leading to an over-estimate of the number of anonymous users. Conversely, in the data we have 74 IP address that were used by multiple users, making it impossible to attribute anonymous work from those IP addresses. Finally, 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.

Analyzing the Data

Our research address how much work is done anonymously, in what pattern and if considering anonymous work makes a significant difference. We answer these questions first by providing descriptive statistics for the count of activities (overall and per session) in dataset two that are anonymous. Second, to put flesh on these numbers, we also provide more detailed descriptions of how anonymous works is distributed through the contribution histories of ten users selected randomly from the group of users who contribute anonymously.

To answer the third part of this question, we test whether the overall distribution of activity counts for users are significantly different including and not including anonymous work. While adding anonymous work certainly creates a different dataset, it might be that the differences between the two datasets are small compared to the high level of variation seen among different volunteers. To test for statistical significance of the difference between the two datasets, we use a non-matched sample test to compare them. Specifically, as the data come from a non-normal distribution, we used a non-parametric Wilcoxon rank sum test with continuity correction, which tests the hypothesis that the values in these two dataset come from the same distribution.

RESULTS

We first report on descriptive statistics for the full data set. The distributions of data in most on-line communities are highly skewed, therefore in addition to reporting averages and standard deviations, we report median values.

Descriptive Statistics

The stacked bar graph in Figure 2 shows the number of classifications submitted per day by logged-in and anonymous users. Users contributed 793,188 classifications, of which 684,087 (86.2%) were submitted while the user was logged-in and 109,101 (13.8%) were submitted anonymously. The contribution pattern in Figure 2 reveals a steady proportion of logged-in and anonymous classifications during the six month time period represented in our data.

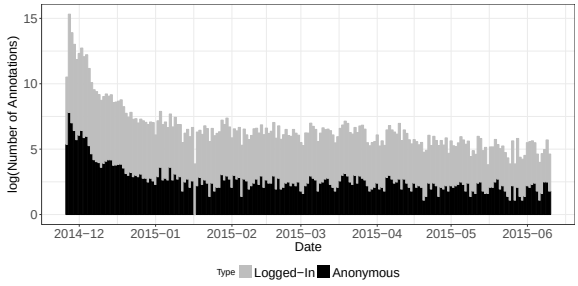

Figure 2. Number of classifications submitted daily by anonymous and logged-in users.

RQ1a: Describing Volunteer Activities

Below, we describe the contribution statistics when we examine the activites of users considering all activities, just logged-in activities and anonymous activities.

All Activities

We found 22,507 unique IP addresses in our dataset. Users submitted 793,188 classifications in 32,972 sessions and spent

an estimated 7,512 hours annotating images and interacting on the website. On average, users contributed 24.05 (SD = 56.13) classifications, in 1.46 (SD = 3.77) sessions, and spent approximately 13 (SD = 26) minutes classifying data. The large standard deviation in classifications, sessions, and time indicate data distributions that are highly (right) skewed—a typical characteristic of user participation data in online communities. As an example of the skewness, one user classified more than 16,000 classifications, while 6,529 (29%) users contributed only one classification. Since the data are skewed, reporting median values give a better indication of the typical work. The median values for classification is 4, session is 1, and time spent annotating is 2 minutes.

Logged In Work

The subset of data containing only logged-in classifications (classification records where the user name is not blank) included 6,354 unique user names. The logged-in work includes 684,087 classifications submitted across 17,349 sessions, spanning 6,194 hours. On average, logged-in users contributed 107.66 (SD = 699.3) classifications across 2.73 (SD = 11.5) sessions. The median work for logged-in users is 16 classifications in 1 session lasting 10 minutes.

Anonymous Work

Finally, we examined the subset of classifications that were anonymous (N = 109,101). Anonymous classifications were submitted from 17,213 unique IP addresses in 18,712 sessions. The difference between the number of IPs and sessions (1,499) reveals that activities were performed from the same IP address anonymously in multiple sessions. Anonymous user IPs contributed in the system for a total of 1,216 hours and contributed on average 5.83 (SD = 15.83) classifications in 1.24 (SD = 1.9) sessions, and lasting approximately 53 (SD = 114) seconds. The median values were 3 classifications in 1 session and lasting 24 seconds.

Attributing Anonymous Work

As described above, we examined the IP addresses associated with classification records to try to attribute non-logged-in contributions to a known user. This attribution was possible because 668,000 (84%) of logged-in classification came from an IP address that was used by only a single user account. For 16,087 (2%) classifications, the IP address was used by multiple user names, making it impossible to distinguish which user was responsible for anonymous classifications from those IP addresses.

RQ1a: Describing anonymous work

The bar chart in Figure 3 show the results of applying Panciera's et al.'s categorization to our dataset. The three bars at the bottom are all types of anonymous work, but were attributed to a more specific one of Panciera's et al.'s categories. We found 28,744 (26%) anonymous classifications that came from an IP address that was otherwise used by only one logged-in user. These classifications represented 4% of the total classifications in the project (fewer than the 10.8% Panciera et al. [23] identified). 80,357 (10%) classifications could not be linked to a user account because the IP address was not used by known user and 2% (N = 16,087) of classifications had IPs that co-occurred with more than one known user account.

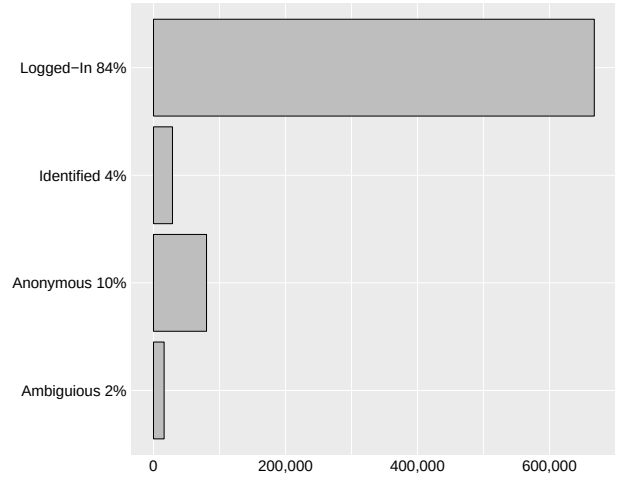

Figure 3. Event IP groupings showing the number of classifications in Higgs Hunters that were from Ambiguous (2%), Identified (4%), Logged-In (84%), and Anonymous (10%) classifications.

Even with the lower percentage of identified anonymous work, we were able to connect anonymous contributions to 3,099 (50.3%) of users, that is, half of users anonymously at some point during their tenure in the project. When we added the anonymous classifications to registered classifications, the average number of classifications per user increased from 132.17 (SD = 759.68) to 140.85 (SD = 761.5), an average increase in classification count of 8.67 (SD = 20.06), with a median value of 5. One user's classification history increased by 616 classifications.

RQ1b: User Stories

To paint a richer picture of anonymous work, we identify patterns of behaviors that emerged when users contributed anonymously. To highlight the patterns of anonymous work, we describe in more detail the contribution histories of ten users selected randomly from the population of users who contributed anonymously. Their classification contributions are shown in Figure 4. On the x-axis are sessions and on the y-axis, users. The points in the figures represent classifications, with the size of the points representing the number of classifications a user submitted in the session. The top chart displays only logged-in classifications and the bottom chart shows the merged logged-in and identified classifications. The points coloured in black are sessions in which the user contributed anonymously. Note that the smallest black dots in the top chart represent sessions in which the user only contributed anonymously (i.e., there are no logged-in contributions).

The chart reveals that these users contributed anonymously in a variety of ways. There are three interesting characteristics of anonymous classifications that emerge. First, anonymous classification is a particularly important part of work in the first session. Of the 8,980 total first sessions in dataset, 3,069 (34.5%) included at least one anonymous classification.

In most cases, the size of the points are noticeably larger in the second chart, indicating a large fraction of first session work was anonymous. User 9 for example (the top row), contributed 275 total classifications in 2 sessions, 240 in her first session and 35 in the second session; 66 (27.5%) of the classifications in the first session were anonymous. Some users do not log in at all during their first session. User 10 (7th row), who contributed across 15 sessions in total, contributed 2 classifications in his first session, both anonymously. Users 8, 4, 6, 7, and 3 also contributed only anonymously during their first session. In the total dataset, the number of users never logging in during their first session was 396 or 4% of total users.

Second, some users contributed anonymously even beyond their first session. Some users begin the project and contribute anonymously for some time while others might intermittently contribute anonymously. An extreme example of the former is User 3 (the bottom row), who contributed in 18 sessions in total, anonymously in the first 9 sessions. After session 9, User 3 only contributed classifications while logged-in. Surprisingly, there was a four month time-span from the first session to the 9th session. In contrast, User 5 (third from the bottom) contributed across 16 sessions in total, anonymously in three: 6, 10, and 11.

Finally, although not represented in Figure 4, we noticed an ongoing sequence of anonymous contributions from the same IP address (i.e., potentially but not certainly from the same user). We saw 367 classifications over a period of two hours and 43 minutes during the first session from that address, followed later by seven more sessions, for a total of 1,127 classifications without a known user name.

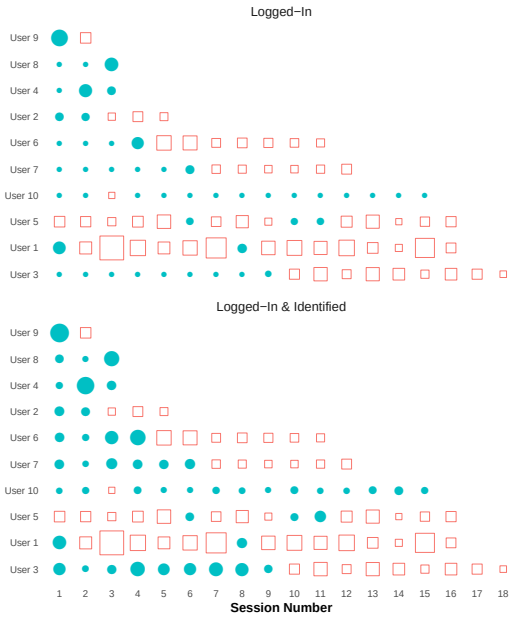

Figure 4. The x-axis shows sessions and the y-axis, unique users. The top chart includes only classifications submitted when users were logged-in. The bottom chart includes logged-in and identified classifications (i.e., anonymous, but attributed to a user by IP address). The circles indicate sessions with anonymous classifications and the squares, sessions with none. The size of the points represents the number of classifications in the session.

RQ1c: Does It Make A Difference?

To address the final question, we test statistically whether adding anonymous work makes a significant difference in the resulting data. To do so, we compared our two datasets, data set one that includes only logged-in classifications, and data set two that combined logged-in classifications with identified classifications. We used the Wilcoxon rank sum test to test whether the two data sets were statistically significantly different. The results $Z = 5422300$, $p < 0.001$, indicating that the two datasets are drawn from different distributions.

We also compared the number of sessions in the two data sets. If a user classified anonymously in a full session (e.g., forgetting to log in), grouping by user name would miss this activity. Linking known users with IP addresses increased the number of sessions for 485 volunteers (8%). The session count increased by 1.32 (SD = 0.92) on average for the 485 users, with a median increase of 1. One user's history increased by 14 sessions. We examined the distributions of sessions from a dataset that included anonymous sessions and one excluding anonymous sessions. The results of the Wilcoxon rank sum test again indicate that the difference is significant, $Z = 171160$, $p < 0.001$. In other words, omitting anonymous work leads to a statistically significantly different data set and significantly different users.

DISCUSSION

Many Facets of Anonymity

Anonymous work plays a significant role in the Higgs Hunters project, as 13% of the total contribution are made anonymously. Replicating the IP classification categorization from [23], we were able to associate 26% of anonymous classifications with known users. Our analysis revealed three important findings in this regard (1) anonymous work is not limited to first session activities (users contributed approximately 13K classifications after their first session), (2) a portion of users appear to deliberately remain anonymous contributors, lastly (3) users' data appear to come from different distributions when their anonymous classifications are taking into account. These findings highlight the complexities of deriving complete and accurate descriptions of user behaviors and accounts of user activities.

Analysis of the behaviours of ten users point to a complex relationship between the logged-in and anonymous classifications. Some users contribute a lot of anonymous classifications in initial sessions and then always contribute logged-in, while other users seem to contribute a few anonymous classifications in their first session and sporadically contribute anonymously thereafter. Our analysis revealed this was the case for some users. Users 3, 6, 7, 10, 4, and 8 all of whom only contributed anonymously in their first sessions and some beyond the first session (in the case of User 3, sessions 1–5 would not have been a part of her history if we left out anonymous classifications).

To the second finding, that some users seem content to remain anonymous, points to what Muller [19] describe as situational disposition, where users intentionally maintain a state of anonymity online. In Zooniverse, the technical architecture might provide some additional clues. In a recent survey, 64.6 percent of respondents never posted a message in the forums or "Talk" pages, activities that require a user to login. In another Zooniverse survey[2] users reported participating "Only when I have spare time", which might limit the desire to go through the hassle of logging-in. Some users reported that they only participate in the project to help scientist filter data, comparable to Anthony et al.'s [1] second group of good Samaritans, those who contribute without need for recognition or tracking contributions (which is what registering provides in Zooniverse).

We should also note, while the cases above show many users contributed anonymously in their first session, there were 708 users who did not contribute anonymously during their first session and 636 (90%) of those users never contributed anonymously. This behaviour points to a dedicated effort to track work in the project.

Finally, the point that users look different when anonymous classifications are included was clear though visualizing the chart in Figure 4 and the statistical analysis we performed that showed in fact the number of classifications and time have impacts on how users "look" in the system (50% of users got bigger in terms of contributions and 8% in terms of sessions). This research points to the need for additional investigations examining the role and characteristics of the anonymous user. Given the existence of different user typologies, designing systems to support users with different participation intentions seems important to retaining volunteers.

Researching User Behaviors Online

The difference in how users look might be more significant for some users than it is for others. Imagine the Wikipedia editor who applied for administrator status and contributed 100 anonymous edits that are not included in their history. These edits might push a user beyond an edit threshold. Similarly, another citizen science project (Gravity Spy) offers multiple workflows and uses a crowd classifier to determine whether users are promoted to advanced levels [28]. In such a project, excluding anonymous classification might delay a user's promotion. It may be that including anonymous activities is most necessary for computational models that impact how users achieve status or ascend roles in online production communities.

Future Work

The work presented above has some limitations. First, We were not able to link 12% of the classification work with a user: 10% was anonymous from an IP address without a known user account and 2% was ambiguous, from an address associated with more than one account. Assigning the 12% of unassociated work to known users would be challenging simply because of the workings of IP addresses. Future research might use alternative approaches to assign anonymous work to a user. For example, the authors in [27] found that browser characteristics (e.g., browser version, plug-ins, screen size, etc.) of individuals consistent across sessions. We suspect these browser characteristics could be used to assign anonymous activities in other online settings.

[2]https://blog.zooniverse.org/tag/survey/

ACKNOWLEDGEMENTS

We thanks to the volunteers and Zooniverse team for access to data. This material is based on work supported by the National Science Foundation under Grant No. IIS xx-xxxxx.

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. Judd Antin and Coye Cheshire. 2010. Readers are not free-riders: reading as a form of participation on wikipedia. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 127–130.
3. Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. 2015. Functional Roles and Career Paths in Wikipedia. In *CSCW '15*. ACM Press, New York, New York, USA, 1092–1105.
4. Susan L Bryant, Andrea Forte, and Amy S Bruckman. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*. ACM, 1–10.
5. 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.
6. Moira Burke and Robert E Kraut. 2008a. Mopping up: modeling wikipedia promotion decisions. In *CSCW '08: Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 27–36.
7. Moira Burke and Robert E Kraut. 2008b. Taking up the mop: identifying future wikipedia administrators. In *CHI'08 extended abstracts on Human factors in computing systems*. ACM, 3441–3446.
8. Boreum Choi, Kira Alexander, Robert E Kraut, and John M Levine. 2010. Socialization tactics in wikipedia and their effects. In *CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM Request Permissions, New York, New York, USA, 107.
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 *CHI '14*. ACM Press, New York, New York, USA, 2985–2994.
10. R Stuart Geiger and Aaron Halfaker. 2013. Using edit sessions to measure participation in wikipedia. In *CSCW*

'13: Proceedings of the 2013 conference on Computer supported cooperative work. ACM Request Permissions, New York, New York, USA, 861.

11. R Stuart Geiger and David Ribes. 2011. Trace ethnography: Following coordination through documentary practices. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 1–10.
12. Aaron Halfaker, Oliver Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. 2015. User Session Identification Based on Strong Regularities in Inter-activity Time. In *WWW '15*. ACM Press, New York, New York, USA, 410–418.
13. Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*. ACM, 163–172.
14. Corey Jackson, Carsten Østerlund, Veronica Maidel, Kevin Crowston, and Gabriel Mugar. 2016. Which Way Did They Go?: Newcomer Movement through the Zooniverse. In *CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, New York, New York, USA, 624–635.
15. Corey Brian Jackson, Carsten Østerlund, Gabriel Mugar, Katie DeVries Hassman, and Kevin Crowston. 2014. Motivations for Sustained Participation in Crowdsourcing: Case Studies of Citizen Science on the Role of Talk. In *2015 48th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 1624–1634.
16. 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 *CHI '16:*. San Jose, CA.
17. Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. 2016. Early Activity Diversity: Assessing Newcomer Retention from First-Session Activity. In *CSCW '16*. ACM Press, New York, New York, USA, 594–607.
18. 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*.
19. Michael Muller. 2012. Lurking as personal trait or situational disposition: lurking and contributing in enterprise social media. In *CSCW '12: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. IBM Research, ACM, New York, New York, USA, 253–256.
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*. ACM Request Permissions, New York, New York, USA, 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 G Terveen. 2009. Wikipedians are born, not made. In *GROUP '09*. ACM Press, New York, New York, USA, 51.
23. Katherine Panciera, Reid Priedhorsky, Thomas Erickson, and Loren G Terveen. 2010. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. In *CHI '16:*. ACM Request Permissions, New York, New York, USA, 1917–1926.
24. 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.
25. Jennifer Preece and Ben Shneiderman. 2009. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction* 1, 1 (March 2009), 13–32.
26. 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*.
27. 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*. International World Wide Web Conferences Steering Committee, 1261–1269.
28. 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 Østerlund, Joshua Smith, Laura Trouille, and Vicky Kalogera. 2016. Gravity Spy: Integrating Advanced LIGO Detector Characterization, Machine Learning, and Citizen Science. *arXiv. gr-qc* (Oct. 2016), 1–27.