Private Peer Feedback as Engagement Driver in Humanitarian Mapping

MARTIN DITTUS, ICRI Cities, UCL
LICIA CAPRA, Dept. of Computer Science, UCL

Prior research suggests that public negative feedback on social knowledge sharing platforms can be powerfully demotivating to newcomers, particularly when it involves peer feedback mechanisms such as ratings and commenting systems. What is the impact on newcomer retention when feedback is private, and from a single peer reviewer? We study these effects using the example of the Humanitarian OpenStreetMap Team, a Wikipedia-style social mapping platform where the review process is closer to a teacher-learner model rather than a public peer review. We observe peer feedback for early contributions by 1,300 newcomers, and assess the impact of different classes of feedback, including performance feedback, corrective feedback, and verbal rewards. We find that verbal rewards and immediate feedback can have a powerful effect on newcomer retention. In order to better support such positive engagement effects, we recommend that system designers conceptually distinguish between mechanisms for quality control and for learner feedback.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; *Collaborative and social computing design and evaluation methods*; Collaborative and social computing systems and tools;

Additional Key Words and Phrases: Peer Feedback; Engagement; Retention; Rewards; Motivations; Crowdmapping; Crowdsourcing.

ACM Reference format:
Martin Dittus and Licia Capra. 2017. Private Peer Feedback as Engagement Driver in Humanitarian Mapping. *Proc. ACM Hum.-Comput. Interact.* 1, 1, Article 40 (January 2017), 18 pages.
https://doi.org/10.1145/3134675

1 INTRODUCTION

Studies of peer feedback in social media and online community platforms have shown that public feedback mechanisms can have unintended long-term consequences. Strongly negative contribution feedback may be discouraging to newcomers [15], and public peer ratings and commenting systems can yield harmful feedback effects affecting contributor behaviour [7, 23]. Such effects may arise in part because public peer feedback fosters a desire to build reputation and social standing [29, 32]. Do these outcomes differ on platforms where peer feedback is given in private?

We observe the impact of private peer feedback on newcomer engagement in the novel setting of the Humanitarian OpenStreetMap Team (HOT). HOT is a volunteer platform which provides maps for regions in humanitarian need, it seeks to grow a vast contributor community so it can better cover the vast scales of affected geographic regions. While participation is online and open to the public, volunteers need to learn specialist tools and workflows in order to contribute, and many newcomers drop out early [11].

HOT participation involves the tracing of roads and buildings that have been captured with satellite imagery. Contributions are coordinated with a Tasking Manager, and collected on the

social mapping platform OpenStreetMap (OSM), sometimes called the "Wikipedia of maps" [27, 31]. According to [12], one can conceptually distinguish three ways for participants to discover HOT projects: first, some initiatives are highly promoted in online and offline media, drawing public audiences; second, organisers may recruit participants directly, for example by establishing partnerships with other large organisations; finally, contributors may start by browsing the published list of available HOT projects on their own accord, with or without a particular intention. Contributors begin their work by choosing a free task in the tasking manager, and then contribute according to a set of instructions. This involves the use of OSM mapping tools to trace the imagery, and uploading the resulting maps to OSM. In addition to conventional OSM workflows, the mapping process with HOT further includes a peer review workflow in the form of a validation stage. During validation, contributors can self-nominate to review the work of other contributors. They can mark the work as accepted or rejected, and optionally send a private message to the contributor. (More detailed discussions of the HOT contribution process can be found in [11, 27, 31].) This validation stage can provide an important opportunity to receive early guidance and encouragement from a peer, possibly a more experienced contributor. This may have important engagement effects, for example by improving newcomer self-efficacy. Conversely, it is feasible that negative feedback and the rejection of contributions may be discouraging to newcomers.

We analyse the effects of private HOT validation feedback on newcomer engagement, and relate our findings to known effects of other platforms. Are HOT newcomers more likely to be retained when they are given certain kinds of feedback? Should validators be concerned about the engagement impact of their feedback, and spend effort on the content and tone of their messages? Are there opportunities to use validation feedback strategically to improve newcomer retention? How do these outcomes relate to platforms where feedback is given in private?

We find that verbal rewards and immediate feedback can have a powerful effect on newcomer retention, while the impact of other factors was inconclusive. Based on our findings, we recommend that system designers consider the role of peer feedback beyond quality assurance. In particular, we observe that peer feedback can also play a supporting function in early skills development, which is an important prerequisite for community capacity-building.

1.1 Summary of Contributions

We present the first large-scale quantitative study of the HOT validation process, with a focus on its effects on newcomer retention. The present study reproduces classic engagement effects of newcomer feedback in a novel setting. Much of the existing research of social knowledge sharing platforms is focused on feedback mechanisms that are public, such as peer ratings and commenting systems. In comparison, in HOT the review process is closer to a teacher-learner model: feedback is private and from a single reviewer, rather than socially negotiated. We assess whether this variation of the review model has an impact on engagement outcomes.

Our large-scale study contributes new knowledge which can support the analysis and design of social knowledge-sharing platforms. In particular, we provide an empirical expectation of the effects of private peer feedback on newcomer retention, and discuss how this can differ from public peer feedback. Better understanding of their effects can support the design of improved feedback mechanisms, for example by avoiding negative social feedback loops where they may be harmful to newcomer retention.

2 RELATED WORK

Prior work on the engagement impact of contribution feedback can be found in a wide range of settings. Research in behavioural psychology can provide general expectations for the impact

of different kinds of feedback. In the more specific context of social media and online platforms, further effects have been identified for social review sites, Q&A sites, and peer production platforms. Research on the role of feedback in education can provide additional expectations for settings where feedback is not socially negotiated.

2.1 Contributor Motivations as Engagement Drivers

Kraut et al. (2012) provide a comprehensive discussion of a wide range of empirically tested theories to help interpret engagement effects in the context of online community platforms. Among the many factors that are discussed, contributor motivations are one of the most important drivers for sustained participation. When they are appropriately considered in platform design, they can provide triggers for action, and support rewarding experiences that can sustain contributor engagement for long periods. However, when the contribution experience comes in conflict with existing contributor motivations, participants are unlikely to remain active.

In motivational theory, a distinction is made between intrinsic motivations by the participant, such as curiosity or a desire for social standing, and extrinsic motivations that are provided to the participant by others, for example in the form of rewards or punishment. The relationship between the two is complex, and subject to a number of moderating factors [6]. *Verbal rewards* such as expressions of appreciation and gratitude by peers have been shown to strengthen intrinsic motivations. Conversely, tangible rewards such as badges or access to higher social status do not always have the desired effect. For tangible rewards, *performance-contingent rewards* which are linked to performance are found to be more effective than task-contingent rewards, which are merely linked to the completion of a task [6, 16].

In volunteer communities, *constraints in the capacity and ability* of participants may pose further barriers to sustained participation. Newcomers can easily be demotivated by strongly negative feedback, compared to participants who are more experienced [15, 35, 36]. It may be possible to address this difference with targeted interventions to improve newcomer self-efficacy, for example by nurturing and educating new contributors, and by fostering a belief that participation is possible and will be welcomed [4].

2.2 Feedback and Ratings on Online Platforms

In social media and online community platforms, motivational aspects are often further moderated by social factors. This is particularly the case when feedback for contributions is made public, or when it is socially negotiated. As a result, contribution feedback in these settings can have unintended long-term consequences [7].

Empirical findings to date reveal a complex picture. On one hand, peer feedback can be a basis for important *social learning effects*. A study of Q&A sites finds that votes, favourites, and answers to questions provide a form of contributor feedback which can improve future question quality. The study further finds that effects are dependent on the content of such feedback, not merely the volume of feedback [2]. In an evaluation of Wikipedia socialisation tactics, it is observed that early user retention was increased by the use of welcome messages, assistance, and constructive criticism [8]. An effort to provide easy access to peer advice in the form of a Q&A forum yielded a significant increase in newcomer retention [25]. Other Wikipedia efforts to provide opportunities of social encounter yielded comparable outcomes [18, 26]. This suggests that to new participants, access to social encounter can make a significant difference in their decision to keep contributing.

On the other hand, organisers need to consider a *trade-off between participant motivations and outcome quality*. Positive and social feedback can increase participant motivation [36], however strongly negative feedback can harm motivation. On Wikipedia, reverts to article modifications can

improve article quality, but they are powerfully demotivating to newcomers [15]. As a consequence it may be advisable to provide more nuanced feedback, rather than an outright rejection of an entire contribution. When negative feedback is less strong and more directive, for example by providing recommendations for future improvement, it can improve task performance without harming motivation [36].

In some settings, negative feedback can have *long-term social effects relating to community regulation.* In a rating system for online news, early negative feedback was linked to significant behavioural changes that are detrimental to the community. Negatively reviewed authors became more likely to review others negatively, while positive reviews had no such effect. Authors who received no feedback were most likely to leave a community, suggesting that the social affirmation of positive feedback may be an important prerequisite for sustained engagement [7]. Similar feedback effects were found on an online travel site. Here, early negative reviews of a venue affected the tone of subsequent reviews, resulting in a herding effect [23].

Public feedback can further affect the *reputation and social standing* of participants. On Q&A sites, perceived reputation can be an important driver for participation, and reputation-building strategies are widespread across a range of participants [29, 32]. This is even more pronounced in online market places, competitive environments where feedback is given strategically in order to gain a market advantage [10, 30].

2.3 Feedback Models in Education

Hattie and Timperley (2007) provide an overview of contemporary theories of feedback in education. Of particular relevance to the present study are empirical expectations of learner engagement after receiving different types of feedback [17].

Performance feedback includes feedback about how well a task is being accomplished, for example by distinguishing correct and incorrect answers [16]. Positive performance feedback on a task of high interest to the learner can increase future motivation [9], or can increase self-efficacy for learners who showed low initial performance [17]. In contrast, negative performance feedback for participants with low self-efficacy or initial performance can reduce motivation [5, 24].

Corrective feedback includes specific and directed suggestions for further actions, for example by helping to distinguish correct from incorrect answers, and by providing more information that aids understanding and helps build procedural knowledge. When it corrects faulty interpretations, corrective feedback can improve self-efficacy and enhance existing motivation [22, 33].

The *timing of feedback* can play an important role, both in micro and macro scales. On the micro scale of individual tasks, immediate feedback was found to be a more effective driver for improved future task performance [17]. On the macro scale of a learner's journey, feedback was found to be more impactful at the beginning: newcomers are particularly receptive to external feedback, and more eager to adopt social norms and practices of their new environment [3].

In online education, certain types of *private peer feedback* are associated with higher attainment. In a study across multiple online education platforms, students reported an increase in course enjoyment from peer feedback, and stated that the resulting exposure to alternative problem-solving approaches improved their learning experience [1]. Among high school students interacting through an online platform, positive affective peer feedback such as socio-emotional support was shown to increase student motivation and self-efficacy, which in turn can increase their performance. However, peer grading and task feedback on the platform had no measurable effect on either motivation or future performance [21].

2.4 Knowledge Gaps

Much existing research of peer feedback in social media and social knowledge production platforms is focused on social feedback mechanisms that are public, such as peer ratings and commenting systems. It was shown that public feedback mechanisms are subject to social feedback effects including herding, and that they can foster a desire to build reputation and social standing.

In comparison, in HOT the review process is closer to a teacher-learner model: feedback is private and from a single reviewer, rather than socially negotiated. As a consequence, it is feasible that the absence of strong social feedback mechanisms changes the effects of the received feedback. To our knowledge, no existing empirical studies assess the relationship between peer feedback and subsequent newcomer retention in such a novel setting. Do outcomes differ from the online and education settings observed in previous studies?

Our large-scale study contributes new knowledge which can support the analysis and design of social knowledge-sharing platforms. In particular, we provide an empirical expectation of the effects of private peer feedback on newcomer retention, and discuss how this relates to public peer feedback. Better understanding of such effects can support the design of improved feedback mechanisms, for example by avoiding negative social feedback loops.

3 RESEARCH QUESTIONS

We regard validation feedback as an intervention that may have an effect on subsequent newcomer retention, and seek to determine under which circumstances the intervention becomes effective. We distinguish between the *validation verdict*, which is the decision whether a given contribution was accepted or rejected, and the *validation message* which can accompany a verdict with more detailed feedback. Together they constitute the validation feedback.

3.1 RQ1. What is the Impact of the Validation Verdict?

How important is a positive validation outcome for newcomer engagement? Can an early rejection discourage further participation? The theory suggests that when intrinsic motivation is present, positive task feedback can foster repeat participation [9, 17]. However such feedback should be performance-contingent, not merely task-contingent: positive feedback is only effective if it is linked to contribution performance [6, 16]. In contrast, negative feedback can reduce motivation when initial self-efficacy is low [5, 24]. Are these effects observable for newcomers in HOT?

3.2 RQ2. What is the Impact of the Validation Message?

We consider the content of validation messages beyond the verdict alone, and distinguish between feedback directed at the task performance, and feedback directed at the self. In relation to task performance we observe uses of positive and negative performance feedback [9, 17], and of corrective feedback [22, 33]. We further observe uses of verbal rewards [6], a kind of affective feedback which includes signs of appreciation and encouragement. Do these different types of feedback have an impact on subsequent newcomer engagement?

3.3 RQ3. What is the Impact of Delayed Feedback?

Does the timing of the verdict matter? According to the theory, task-specific feedback is more likely to be effective if it is given early [17]. Is overall retention affected when feedback is delayed?

4 METHOD

4.1 Dataset

Our evidence is derived from two data sources. A primary source is the HOT Tasking Manager, it provides a public list of all HOT projects.[1] Each project seeks to map specific features in a particular region of the world, typically in preparation for humanitarian field work, for example in response to an environmental disaster. The Tasking Manager divides the mapping work for each project into manageable tasks, and provides instructions and a participation history. It also stores a private record of the subsequent validation process, including validation verdicts and validation messages.

A second data source is the OSM edit history, a large data set which captures map contributions over time. All map contributions by HOT volunteers are immediately made public on OSM and captured by this edit history, however they can then be amended and reverted by validators and other contributors. The full data set is freely available for download.[2]

In an initial stage we link these two data sets, identifying the map contributions made by HOT contributors while participating in particular HOT projects. Since summer 2015, OSM editing tools automatically annotate map contributions with a HOT project identifier, which makes such an identification straightforward. For earlier records, map contributions can be identified and linked based on their location, date, and contributor ID. The resulting connected data set allows us to identify the specific map contributions for every contributor during a project participation, including any map edits by validators.

This yields an annotated participation history which forms the basis of our analysis. In an initial stage, we identify a set of basic features for every contributor. The feature *join_day* records the date of a newcomer's first HOT contribution, in days since 1st January 2014. It can serve as control variable to capture changes in newcomer retention over time.

4.2 Study Population

We observe first-time HOT contributors with no prior OSM experience. We start our observation period in October 2014, the first month when HOT contributors were shown a notification for all validation messages in the Tasking Manager interface. We only include newcomers who join at least 90 days before the end of the available edit history at the time of study. This allows us to observe their retention over time. In particular, we select first-time contributors to HOT who:

- first contributed between 1st October 2014 and 20th June 2016 (both inclusive),
- have no more than one prior day of OSM contribution experience outside of HOT,
- have a task validated within the first 45 days of joining,
- were sent a validation message along with their verdict.

Almost 1,300 contributors match these criteria. For each newcomer, we identify the first validation message they received, these form the basis of our study. For each of the 1,300 messages we determine *delay*, the delay between submission and feedback message in days, and the binary indicator *low_delay* which marks message that were received before the median delay of 28 hours.

For these initial cases, the average task acceptance rate is 34%. We do not consider any subsequent messages newcomers may have received: our study is focused on the impact of early interventions during a newcomer's participation history, in part to avoid confounding factors in later participation stages, but also to ensure consistency in our analysis. Furthermore, it is comparatively rare to receive multiple messages: only 15% of participants received more than one message throughout their participation lifetime, and only 3% received more than 5.

[1] http://tasks.hotosm.org
[2] http://planet.osm.org/planet/full-history/

4.3 Message Classification

We characterise the content of the 1,300 validation messages by means of a thematic analysis. Informed by a review of the literature we distinguish multiple complementary aspects of learner feedback: positive and negative performance feedback [9, 17], corrective feedback [22, 33], task-contingent rewards [6, 16], and the use of verbal rewards [6]. They are not mutually exclusive: each message can contain combinations of these.

Validation messages tend to be short, at an average of 100 characters, which makes it feasible to manually label the full corpus. Each message is reviewed by the first author to determine whether it satisfies the specified criteria for one or more of these feedback techniques. This labelling process is only concerned with the message content. We further discard 74 messages which were written in languages other than English.

To support this process and ensure labelling consistency, in an initial labelling stage we iteratively develop a codebook that is then used as a reference. This codebook includes a definition and example phrases for every code. To build it, we first label a random sample of 100 and then 500 messages, collecting example phrases for each code, and clarifying our definition statements when necessary. We repeat the process until we reach a saturation point where the definition statements and example phrases allow an unambiguous distinction between all codes across the samples. We then manually label each message in the full corpus, recording the presence or absence of each code. To increase confidence, a second rater labels a randomly sampled subset of 500 messages, after receiving one hour of training. We measure rater agreement with Cohen's Kappa for each individual code. Across the labelled codes, we find an agreement between 0.82 and 0.97 among raters, these Kappa scores fall within the range of almost perfect agreement [20].

These codes will be discussed in turn. Three codes describe *feedback that is focused on the task*:

- Positive Performance Feedback (PPF). Phrases that express a *positive assessment of the merit* of the current contribution. Example phrases include: good job, great work, looks good, this is high quality, good interpretation of the imagery, nice mapping, etc.
- Negative Performance Feedback (NPF). Phrases that express a *negative assessment of the merit* of the current contribution, for example by highlighting omissions or mistakes. Example phrases include: doesn't look complete, missing, needs improvement, not done yet, not mapped, not very accurate, still not traced, untagged, needed squaring, etc.
- Corrective Feedback (CF). Phrases that provide *procedural guidance* about the contribution process, typically to improve future work. This can include explicit requests for specific acts, including detailed instructions, or references to external documentation. It can include implicit request for action, for example a negative merit assessment. It can also include invitations to consider a particular concern, clarify community expectations, or other means of fostering deeper understanding. Example phrases include: could you, can you correct, don't use, double check that, read the instructions, learn how to, look for, make sure, please complete, missing, untagged, please do not, please trace, should be drawn, etc.

During early iterations, we further observed a significant number of instances where validators accepted contributions which contained errors or were otherwise incomplete. Instead of rejecting them, they corrected the mistakes themselves, marked the task as 'accepted', and attached a message where they summarised their refinements. This can be seen as an example of a *task-contingent reward* [6], in an attempt to avoid rejecting a newcomer's early work. We introduced an additional code for such cases to assess their impact:

- Generous Acceptance (GA). Tasks that have been marked as accepted, but where the validator message includes either CF or NPF.

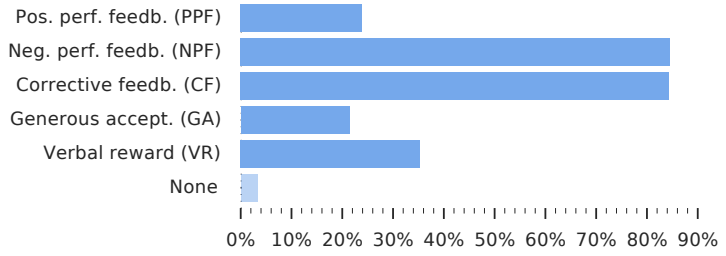

Fig. 1. Frequency distribution of feedback techniques throughout the message corpus.

A correlation analysis confirmed that GA places additional burden on the validator: newcomers whose validation messages is labelled with GA had a larger share of their contributions removed by a validator (Spearman correlation coefficient $\rho_S = 0.32$, $p < 0.001$). In comparison, "strict" acceptance without CF or NPF also incurred such deletions, but to a lesser degree ($\rho_S = 0.10$, $p < 0.001$).

Finally, we include *verbal rewards* as a feedback category that is more broadly focused on the contributor, and which can help strengthen self-efficacy in early learners. For the purpose of this study, we consider verbal rewards to be any feedback that shows appreciation and encouragement, including PPF:

- Verbal Rewards (VR). Phrases that express *positive assessment of merit, gratitude, or encouragement.* This includes any instance of PPF. Further example phrases include: thank you, thank you for mapping, thanks for your contribution, keep mapping, keep up the good work, please don't hesitate, please keep mapping, etc.

Figure 1 shows a frequency distribution of all feedback techniques for the full message corpus. The distributions show that most messages are corrective: CF and NPF are the most widely used feedback types, they are included in approximately 80% of all messages. NPF almost always is used as part of a correction (i.e., along with CF), rather than merely by itself. This suggests an overall constructive tone of the peer feedback. GA is the least widely used technique, at 28%. Only 4% of messages include none of the feedback techniques and remain unlabelled, many of these are instances where submitters accidentally marked a task as completed.

4.4 Newcomer Retention

Newcomer retention is measured by means of a survival analysis, capturing activity during an observation period of 45 days. We record the relative date of the last observed contribution as feature *last_day*, measured in days since the initial contribution. We confirm death events (censure) by ensuring that there is no activity within the next 45 days after this initial observation period, this is recorded as binary feature *has_died*.

We fit a Kaplan Meier survival model to establish base survival rates, these provide a basic expectation of overall newcomer retention. The resulting survival plot is shown in Figure 2. The base expectation of survival is low: only 14% of newcomers are projected to still be active 10 days after their initial contribution, 9% after 20 days, and 6% after 30 days. The 95% confidence interval for this model is within ±3% throughout the observed period, suggesting a high model fit. In other words, the base expectation is that newcomers will not be retained, regardless of any intervention they may experience.

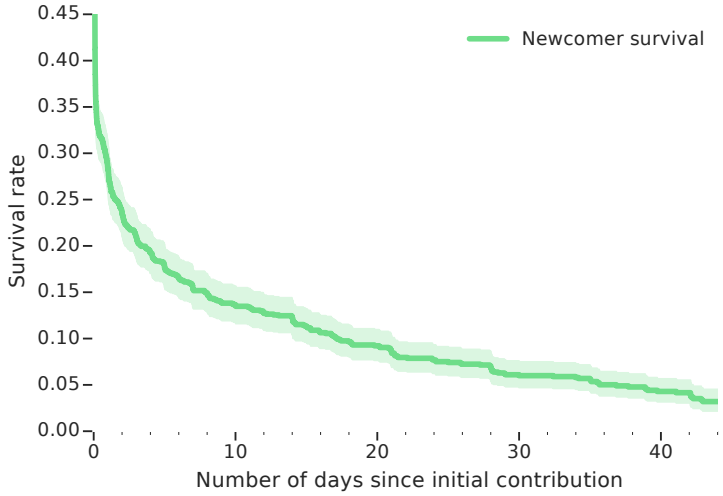

Fig. 2. Base survival function for all study participants, with 95% confidence interval.

4.5 Contextual Factors

While we primarily seek to observe the impact of validation feedback, we need to account for any other factors that are likely to influence newcomer retention in a significant manner. This includes contributor-specific factors such as intrinsic motivation, but also any project-specific factors which may have an effect on subsequent retention.

Much prior work on online communities has shown that intrinsic motivation and commitment to goals is a strong predictor of future engagement [19]. We cannot easily capture contributor motivation at large scale, however we can estimate it based on initial contribution behaviour. To this purpose we measure the *initial effort* as *initial_hours*, the number of hours for which each newcomer contributes to HOT during their first 24 hours of activity. A correlation analysis confirms that this measure is highly correlated with the subsequent survival period *last_day* (Spearman correlation coefficient $\rho_S = 0.50, p < 0.001$). This makes it a suitable proxy measure for intrinsic motivation. We further determine *high_initial_hours*, a binary indicator to identify newcomers whose initial effort was above the median of 74 minutes.

Prior work has shown that urgent disaster-related HOT campaigns can be significant recruiting events, but tend to have lower average retention rates than other more sustained campaigns [12]. To account for this, we include a final control variable *disaster_campaign* to capture whether a newcomer's first HOT contributions were during such a campaign.

4.6 Analysis

We regard validation feedback as an intervention that may have an effect on newcomer retention, and seek to determine under which circumstances the intervention becomes effective.

A Cox proportional hazards model is used to explain the survival rates we observed, based on a set of features and control variables. This is comparable to a regression analysis, but specifically intended to model participant survival. In the context of this study, the term 'hazard' is a synonym for the risk of abandoning HOT participation. A hazards model yields a rate of risk for each covariate, denoting the relative increase in hazard per unit increment. The model fit for such a

Aspect	Variable	Description
Project	*disaster_campaign*	Disaster event or more sustained campaign?
Contributor	*join_day*	Join date, in days since 1st January 2014
	initial_hours	Initial effort: hours spent mapping in initial 24h
	high_initial_hours	Was the initial effort above the median (74m)?
Validation outcome	*accepted*	Was the contribution accepted?
	delay	Delay between submission and feedback, in days
	low_delay	Was feedback delay below the median (28h)?
Message content	*PPF*	Positive performance feedback?
	NPF	Negative performance feedback?
	CF	Corrective feedback?
	GA	Generous acceptance?
	VR	Verbal rewards?
Contributor retention	*last_day*	Last active day, in days after joining
	has_died	No further activity for at least 45 more days?

Table 1. Features collected per first-time contributor.

model is described by its concordance, where 0.5 is equivalent to the performance of a random predictor, and 1.0 denotes perfect prediction accuracy. We further compute 95% confidence intervals for all covariates. In cases when an effect cannot be found, we employ a power calculation method proposed by [14] to determine whether our models have sufficient statistical power to identify an effect. A Cox model relies on two basic assumptions: first, that all covariates are multiplicatively related to the hazard, capturing an effect relative to an initial baseline hazard. Second, that all observed effects are time invariant: all covariates are constant throughout the observation period. In the present study, this assumption can be justified based on a conceptual observation: we seek to model the impact of a specific intervention at the start of the observation period.

We derive a feature vector per first-time contributor using the available features and control variables. Table 1 shows the full feature vector per newcomer which is used for our analysis; different subsets of these features are used to test specific effects. All models include the covariate *disaster_campaign* to control for project-specific factors, and *join_day* and *initial_hours* to control for engagement factors related to the contributor. These variables have been discussed in previous sections. This basic model has a condition index of 2.5, indicating low multicollinearity, and a concordance of 0.68. A 10-fold cross-validation shows that the model is stable when presented with different subsets of the data: the median concordance across iterations is 0.68, with a low standard deviation of 0.02.

To set a basic expectation of potential outcomes, the Q-Q plots in Figure 3 visually compare the survival distributions associated with different message classes. These figures show that acceptance of a task, generous acceptance, and verbal rwards are associated with longer survival periods, while the remaining features are not clearly associated with a single outcome. These are mere associations, hazard models in later sections will confirm whether they indicate real relationships.

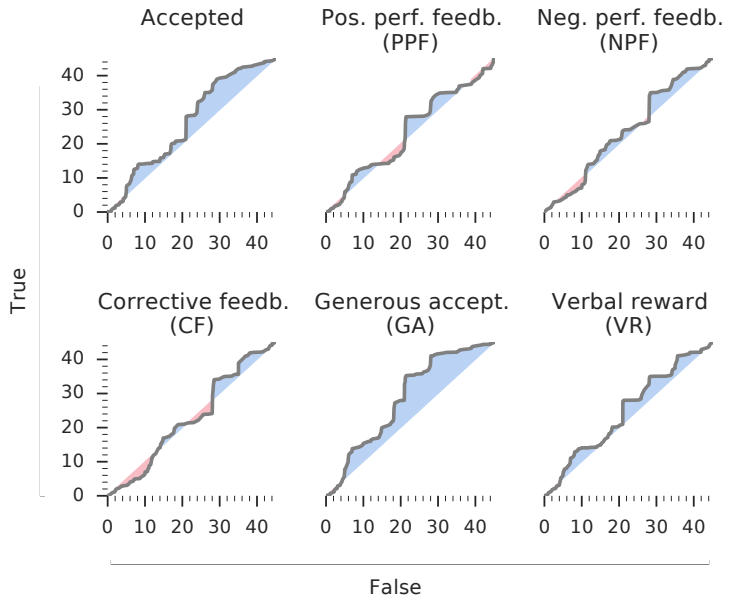

Fig. 3. Q-Q plots of newcomer survival for different feedback types. Each plot compares the distribution of those who have received a particular type of feedback (*True*), and those who have not (*False*). Survival is measured in days after joining.

4.7 Limitations

Although we make an effort to reduce the chance of confounding effects, this observational study can only reveal associative rather than causal relationships. We are further limited to an assessment of existing practices, rather than an exhaustive assessment of all possible practices.

Our study is specifically focused on contributor engagement, and an analysis of the impact on task performance is out of scope, including an assessment of the impact on future contribution quality. Analyses of such effects are presented with methodological challenges which exceed the scope of the present study. In particular, it is not evident how contribution quality can easily be measured at the scale of this study. A fundamental challenge is the absence of a reliable ground truth: by definition, the maps produced by HOT are the first of their kind, and there typically are no other maps to compare against. As a consequence, the few early studies of HOT data quality either rely on specific contextual knowledge of particular geographies, or they assess quality by reproducing the volunteer work with other means, both of which limits their scale [13, 34].

Due to a limitation in the setting, we are not able to determine whether certain messages would reactivate contributors who have previously been inactive. This is because messages sent during the validation process do not result in an email notification, instead they are displayed prominently on the Tasking Manager the next time a contributor logs in. In the current form, HOT messaging requires contributors to actively log in in order to receive a new message. As a consequence, we can only assess the impact of validation messages on contributors who return at least once.

Aspect	Variable	Hazard	p
Project	*disaster_campaign*	133.6%	< 0.001
Contributor	*join_day*	100.0%	< 0.001
	initial_hours	82.8%	< 0.001
Feedback	*accepted*	(94.6%)	0.43

Table 2. Proportional hazards model for the effect of task acceptance on newcomer retention, when compared to task rejection.

5 FINDINGS

5.1 RQ1. Validation Verdict

In order to assess the engagement impact of the validation verdict, we build multiple Cox hazard models to compare the effects of three different validation verdicts: acceptance, generous acceptance, and rejection.

An initial hazard model compares the impact of acceptance and rejection outcomes across the full study population. The feature vector includes all control variables, and *accepted* as a predictor.

The impact of GA is tested with two separate models. GA is first compared with 'strict' acceptance, these are instances where the validation message did not include instances of NPF or CF. This test is computed across those in the study population whose contributions have been accepted, and *GA* is used as the predictor. A second test compares GA with rejection, here the test is computed across those who have received a GA or rejection verdict, and *GA* is the predictor.

For each of these three basic models, we further derive three variants: the full model, and versions of the model that only include high-effort or only low-effort contributors, as determined by *high_initial_hours*. This allows us to identify effects that are dependent on intrinsic contributor motivation, as expressed by their initial contribution effort.

We find that none of the models show an effect associated with the respective predictor. Power analysis suggests that we may simply not have sufficient data to establish an effect: in each model, the statistical power of the feature of interest is below 0.2 ($p<0.05$). Instead, the strongest effects involve contributor-related factors: in all models, the control variable *initial_hours* is significant. In some models, the control variable *disaster_campaign* is significant, indicating engagement effects relating to contributor recruitment.

Table 2 shows an example of such an outcome, in this case for a model comparing acceptance and rejection, with a concordance of 0.679. In this model, the intervention covariate *accepted* is not found to be significant. This covariate only has a statistical power of 0.14 ($p<0.05$), which suggests there is insufficient evidence available to identify a significant effect.

In other words, we find no evidence that the task verdict has an impact on newcomer retention. In particular, we find no evidence that rejection of work harms retention, or that task-contingent rewards through GA improve retention.

5.2 RQ2. Validation Message

To assess the impact of the validation message, we determine the engagement effect of each feedback type. We build a separate model for each PPF, NPF, CF, and VR, using the respective feedback type as a predictor. To identify interactions for sub-populations, we further build variants of these models to only include those who have been accepted, those who have been rejected, and those who showed high or low initial effort.

Aspect	Variable	Hazard	p
Project	*disaster_campaign*	168.9%	< 0.001
Contributor	*join_day*	(100.0%)	0.14
	initial_hours	68.1%	< 0.01
Feedback	*VR*	78.0%	< 0.05

Table 3. Proportional hazards model for the effect of verbal rewards (VR) on newcomers who contributed for less than 75 minutes on their first day.

Aspect	Variable	Hazard	p
Project	*disaster_campaign*	133.3%	< 0.001
Contributor	*join_day*	100.0%	< 0.01
	initial_hours	82.4%	< 0.001
Feedback	*low_delay*	78.6%	< 0.001

Table 4. Proportional hazards model for the effect of early feedback on all newcomers.

We find a significant effect for VR, at a concordance of 0.644. The full model is shown in Table 3. According to this model, for newcomers whose initial contribution effort is below the median, VR reduces the hazard rate to 78%. In other words, among contributors who contributed for less than 75 minutes on the first day, those who received a message with a verbal reward were more likely to remain active in HOT. Compared to the baseline, their likelihood to stop contributing drops to 78%. The 95% confidence interval for this feature provides further support for the presence of an effect (lower threshold at 62.4% hazard rate, upper threshold at 97.6% hazard rate).

For models involving PPF, NPF, and CF, the predictors are not significant. In all other cases, power analysis for the feature of interest never exceeds 0.3 ($p<0.05$), suggesting a lack of sufficient evidence to identify a significant effect.

Models for pairwise combinations of all features are similarly inconclusive, with two exceptions: for newcomers whose initial contribution effort is below the median, the pairwise combinations of VR/CF and VR/NPF both reduced the hazard rate to 77% and 76%, respectively ($p<0.05$, both at concordance 0.646). They can be regarded as more specific extensions of the previously observed effect involving VR.

5.3 RQ3. Timing

To assess the impact of feedback timing, we first test whether early feedback was associated with improved retention. Across the validation messages analysed for the study, feedback was sent relatively quickly. 40% of messages were sent within 12 hours after task submission, the median was at 28 hours. The 25% slowest validation responses were sent a week or more after submission.

A hazards model for early feedback is shown in Table 4. It includes *low_delay* as a predictor, this feature denotes whether a newcomer received their first feedback within 28 hours after submitting. The model shows that early feedback matters: it yields a reduction of the hazard rate to 79%, at a model concordance of 0.672. The 95% confidence interval for this feature further supports this (lower threshold at 69.9%, upper threshold at 88.3%).

A separate model for *delay* as predictor showed that each additional day of delay incurs a proportional hazard increase of 1.4% (with a 95% confidence interval of 0.7% to 2.1%).

6 DISCUSSION

Across all models we computed, the most consistently predictive factor for high contributor retention was a contributor's initial contribution effort, as measured by *initial_hours*. The second-most frequent predictive factor was *disaster_campaign*, controlling for the particular recruitment effects during event-centric disaster campaigns which can yield lower average newcomer retention [12]. Both support the general expectation in the literature that intrinsic motivation is one of the strongest predictors for future participation [6, 9, 11, 28]. Two further aspects related to peer feedback were found to have a powerful effect on newcomer retention: the use of verbal rewards in validation messages, and feedback that was sent without much delay.

Verbal rewards lead to a significant increase in retention (a reduction in hazard rate to 80%) among newcomers who had contributed for less than the median 75 minutes during their first day, possibly indicating low intrinsic motivation or self efficacy. The finding suggests that verbal rewards may increase self-efficacy, which may in turn increase retention. In comparison, newcomers who already start with a high degree of self-efficacy may not require such affective-supportive feedback to remain engaged.

The literature offers a range of possible interpretations for this outcome: the importance of fostering a belief that participation is possible and will be welcomed [4], the importance of positive affective feedback such as socio-emotional support [21], and the importance of positive interactions as a baseline behaviour instead of harmful silence [7], all suggesting that the social affirmation of positive feedback may be an important prerequisite for future participation. The process of contributing to HOT can be considered a depersonalised form of interaction: it is often focused on the task, rather than the learner. In the absence of other prominent social cues, small phrases of support may have a large effect.

This interpretation is further supported by the second significant effect, the importance of early feedback. Peer feedback that is sent a week after a contribution is significantly less likely to still have a motivational impact. In comparison, feedback that is sent within 28 hours or less yielded a reduction of the hazard rate to 78%. Any additional day of delay increased the hazard rate. These findings suggest that the absence of any feedback message likely also increases hazard, supporting a prior observation that contributors who receive no feedback are more likely to leave [7]. However, this does not necessarily mean that delayed feedback cannot also be effective. Rather, this is in part a limitation in the setting: in the current form, HOT messaging requires contributors to actively log in in order to receive a new message. It is feasible that delayed feedback can still be effective when it is coupled with an email notification mechanism, so that contributors receive feedback messages even after they have stopped actively contributing.

In contrast to prior studies involving public feedback, we could not confirm that negative feedback has a negative impact on retention. This includes both the rejection of contributions, and the use of negative performance feedback. These forms of negative feedback in HOT are comparatively mild compared to other platforms: for example they do not necessarily result in a discarding of all contributions. Furthermore, because the feedback is private, it does not incur a risk of reputational damage or other social feedback effects.

We can confirm that task-contingent rewards do not improve future engagement, which confirms prior expectations in the motivational literature [6, 16]. These were issued by some validators in the form of generous acceptance. Additionally, task acceptance in itself appears to have no effect. We further cannot confirm an engagement effect for corrective feedback, possibly because such

feedback is targeted at task performance rather than contributor motivation. However, the evidence basis for these tests is low, which suggests we may simply not have sufficient data to establish an effect. In this respect, it should be highlighted that an initial survival analysis of all participants established the base expectation that newcomers will not be retained, regardless of any intervention. It is further worth pointing out that we encountered a highly imbalanced distribution for certain kinds of feedback: both NPF and CF are used in the vast majority of messages (>80%), suggesting there may not be sufficient counter-examples to capture their effect.

These findings also raise the question whether we successfully captured the most important contributing factors. Our model captures existing theory of the effects of peer feedback well, which is reflected in our detailed labelling process. However, many factors relating to the more general concern of participant retention are not included, in large part because they are unobservable to us. For example, we cannot currently capture prior familiarity with mapping practice outside of OSM, nor the extent to which participants experience peer interactions outside of the observed setting, be it online or in person.

6.1 Implications

We believe that our findings are relevant for a wide range of social knowledge sharing platforms. HOT participation as observed here is open to a lay audience, and the observed feedback effects are independent of the specific mechanics of the HOT contribution workflow. Further, our study places a focus on a transferrable concern: the extent to which private peer interactions can help foster sustained newcomer engagement.

Based on our findings, we recommend that system designers consider the role of peer feedback beyond quality assurance: verbal rewards in peer feedback can have important effects on newcomer retention. To better support such effects, system designers need to distinguish between two separate concerns: on one hand, assessing contribution quality, which is a concern of global task coordination; on the other, guiding newcomers in early skills development, which is a concern of community capacity-building and newcomer training.

Many online platforms conflate these two concepts, so that a single mechanism is responsible for both quality assurance and learner feedback, for example in the form of peer ratings and commenting systems. While the resulting public recognition may be motivating to some, a wide range of studies has shown that negative feedback in such public settings can be demotivating [7, 10, 15, 23, 29, 30, 32, 35, 36]. Our findings demonstrate that private peer feedback is an alternative design option to effectively support community capacity-building while avoiding this particular risk of newcomer discouragement.

In the specific case of HOT, we find that current validation workflows conflate these two concerns, which in turn affects the tone of validation feedback. In the current form, more HOT validation messages include negative performance feedback than verbal rewards, suggesting that validators focus on regulating contribution quality, and that they may not be aware of the motivational impact of their feedback. To address this, we recommend that improved validation interfaces should regard quality control and learner feedback as separate concerns. Organisers may further seek to provide guidance on how best to articulate impactful feedback. For example, they could emphasise that the specific wording of a feedback message matters, and that the recipient of a validation message may appreciate encouragement by a peer. Furthermore, we see potential to introduce further workflows to accompany mappers throughout their early experiences, and to help identify contributors who might benefit from early feedback, ensuring that newcomers have access to peer contact and mentoring.

Based on the available evidence, we have not found an engagement effect related to generous task acceptance. This validation strategy may be beneficial when timely completion of a project is of importance, however it places a burden on validators to finish the mapping work themselves. In comparison, other strategies may be similarly acceptable when timely completion is not a concern, including the rejection of low-quality contributions when accompanied with a supportive message. This is an area that warrants further research and experimentation.

From a theoretical perspective, we see opportunities for further research in this novel setting. Our findings suggest that private peer feedback on HOT may not suffer from the negative social feedback loops observed on platforms where peer feedback is publicly negotiated. For example, HOT contributors may be less concerned about building a public reputation. However the full implications of this difference are not yet evident, and warrant further research.

7 CONCLUSION

Contributing to HOT can be considered a depersonalised form of interaction: it is often focused on the task, rather than the learner. In light of this, it is not surprising to find that the social affirmation of peer feedback can be an important factor in future participation. In the absence of other prominent social cues, small phrases of support can have a large effect. Through this study, we find evidence that verbal rewards and a timely response can significantly improve newcomer retention. On the other hand, we cannot find evidence that feedback related to task performance has an effect on newcomer retention. In particular, we do not find that negative feedback harms retention, as is found on platforms where feedback is publicly negotiated, and where negative feedback can harm one's reputation.

As a consequence of these findings, we propose that public and private peer feedback can be considered complementary mechanisms which can foster different kinds of outcomes. Better understanding of such effects can support the design of improved feedback mechanisms, for example by avoiding negative social feedback loops where they may be harmful to newcomer retention.

7.1 Future Work

There is much scope to further identify specific effects associated with private peer feedback on social knowledge sharing platforms, and relate them to the more well-known effects of public peer feedback. To complement the findings to date, further research may seek to identify the role of social messaging in peer feedback, in particular the use of personal and impersonal language. In addition, the role of the peer reviewer deserves closer scrutiny. What is the effect of reviewer experience level, feedback style, or their standing in the community? For example, do more experienced contributors write different reviews than one-off or first-time reviewers? Does this have an impact on the motivation of the reviewed? There is further research potential to assess the impact of peer feedback on task performance, including contribution quality. What kind of peer feedback is most beneficial for future task performance? There is further scope to observe the longer-term development of those that remain active. Are participants receptive to different feedback throughout their learning trajectory? For example, are newcomers more receptive to being taught by unknown peers, compared to more experienced participants? Finally, there are opportunities to study the extent to which peer feedback supports newcomer socialisation. Is early messaging an effective means of propagating social norms, for example by encouraging polite and constructive exchange? Do norms that are acquired in this manner propagate to future contributor generations, as newcomers become experts and review the contributions of others?

8 ACKNOWLEDGMENTS

We extend warm regards to HOT for their invaluable support. We further thank the HOT validator community for their guidance, in particular Nick Allen and Ralph Aytoun.

REFERENCES

[1] Panagiotis Adamopoulos. 2013. What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. In *ICIS*.
[2] June Ahn, Brian S Butler, Cindy Weng, and Sarah Webster. 2013. Learning to be a better Q'er in social Q&A sites: social norms and information artifacts. *Proceedings of the American Society for Information Science and Technology* 50, 1 (2013), 1–10.
[3] Blake K Ashforth and Alan M Saks. 1996. Socialization tactics: Longitudinal effects on newcomer adjustment. *Academy of management Journal* 39, 1 (1996), 149–178.
[4] Jonathan Bishop. 2007. Increasing participation in online communities: A framework for human–computer interaction. *Computers in human behavior* 23, 4 (2007), 1881–1893.
[5] Joel Brockner, William R Derr, and Wesley N Laing. 1987. Self-esteem and reactions to negative feedback: Toward greater generalizability. *Journal of Research in Personality* 21, 3 (1987), 318–333.
[6] Judy Cameron, Katherine M Banko, and W David Pierce. 2001. Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *The Behavior Analyst* 24, 1 (2001), 1.
[7] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How Community Feedback Shapes User Behavior. In *Proc. ICWSM '14*.
[8] Boreum Choi, Kira Alexander, Robert E Kraut, and John M Levine. 2010. Socialization tactics in Wikipedia and their effects. In *Proc. CSCW '10*. ACM, 107–116.
[9] Edward L Deci, Richard Koestner, and Richard M Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin* 125, 6 (1999), 627.
[10] Chrysanthos Dellarocas. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science* 49, 10 (2003), 1407–1424.
[11] Martin Dittus, Giovanni Quattrone, and Licia Capra. 2016. Analysing volunteer engagement in humanitarian mapping: building contributor communities at large scale. In *Proc. CSCW '16*.
[12] Martin Dittus, Giovanni Quattrone, and Licia Capra. 2017. Mass participation during emergency response: event-centric crowd-sourcing in humanitarian mapping. In *Proc. CSCW '17*.
[13] Melanie Eckle and João Porto de Albuquerque. 2015. Quality assessment of remote mapping in OpenStreetMap for disaster management purposes. In *Proc. ISCRAM '15*.
[14] Laurence S Freedman. 1982. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in medicine* 1, 2 (1982), 121–129.
[15] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proc. WikiSym '11*. ACM, 163–172.
[16] Judith M Harackiewicz. 1979. The effects of reward contingency and performance feedback on intrinsic motivation. *Journal of personality and social psychology* 37, 8 (1979), 1352.
[17] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.
[18] Aniket Kittur, Bryan Pendleton, and Robert E Kraut. 2009. Herding the cats: the influence of groups in coordinating peer production. In *Proc. WikiSym '09*. ACM, 7.
[19] Robert E Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. 2012. *Building successful online communities: Evidence-based social design.* Mit Press.
[20] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
[21] Jingyan Lu and Nancy Law. 2012. Online peer assessment: effects of cognitive and affective feedback. *Instructional Science* 40, 2 (2012), 257–275.
[22] Richard S Lysakowski and Herbert J Walberg. 1982. Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal* 19, 4 (1982), 559–572.
[23] Loizos Michael and Jahna Otterbacher. 2014. Write Like I Write: Herding in the Language of Online Reviews.. In *Proc. ICWSM '14*.
[24] Richard L Moreland and Paul D Sweeney. 1984. Self-expectancies and reactions to evaluations of personal performance. *Journal of Personality* 52, 2 (1984), 156–176.
[25] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proc. CSCW '13*. ACM, 839–848.

[26] David R Musicant, Yuqing Ren, James A Johnson, and John Riedl. 2011. Mentoring in Wikipedia: a clash of cultures. In *Proc. WikiSym '11*. ACM, 173–182.
[27] Leysia Palen, Robert Soden, T Jennings Anderson, and Mario Barrenechea. 2015. Success & scale in a data-producing organization: the socio-technical evolution of OpenStreetMap in response to humanitarian events. In *Proc. SIGCHI '15*. 4113–4122.
[28] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proc. GROUP '09*. ACM, 51–60.
[29] Laurence D Parnell, Pierre Lindenbaum, Khader Shameer, Giovanni Marco Dall'Olio, Daniel C Swan, Lars Juhl Jensen, Simon J Cockell, Brent S Pedersen, Mary E Mangan, Christopher A Miller, et al. 2011. BioStar: an online question & answer resource for the bioinformatics community. *PLoS Comput Biol* 7, 10 (2011), e1002216.
[30] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.
[31] Robert Soden and Leysia Palen. 2014. From crowdsourced mapping to community mapping: the post-earthquake work of OpenStreetMap Haiti. (2014).
[32] Yla R Tausczik and James W Pennebaker. 2012. Participation in an online mathematics community: differentiating motivations to add. In *Proc. CSCW '12*. ACM, 207–216.
[33] Gershon Tenenbaum and Ellen Goldring. 1989. A meta-analysis of the effect of enhanced instruction: Cues, participation, reinforcement and feedback and correctives on motor skill learning. *Journal of Research & Development in Education* (1989).
[34] Clay Westrope, Robert Banick, and Mitch Levine. 2014. Groundtruthing OpenStreetMap building damage assessment. *Procedia engineering* 78 (2014), 29–39.
[35] Donghee Yvette Wohn. 2015. The Effects of Feedback and Habit on Content Posting in an Online Community. *iConference 2015 Proceedings* (2015).
[36] Haiyi Zhu, Amy Zhang, Jiping He, Robert E Kraut, and Aniket Kittur. 2013. Effects of peer feedback on contribution: a field experiment in Wikipedia. In *Proc. SIGCHI '13*. ACM, 2253–2262.

Received June 2017; revised July 2017; accepted August 2017