# Crowd Development: The Interplay between Crowd Evaluation and Collaborative Dynamics in Wikipedia

ARK FANGZHOU ZHANG, University of Michigan
DANIELLE LIVNEH, University of Michigan
CEREN BUDAK, University of Michigan
LIONEL P. ROBERT JR., University of Michigan
DANIEL M. ROMERO, University of Michigan

Collaborative crowdsourcing is an increasingly common way of accomplishing work in our economy. Yet, we know very little about how the behavior of these crowds changes over time and how these dynamics impact their performance. In this paper, we take a group development approach that considers how the behavior of crowds change over time in anticipation and as a result of their evaluation and recognition. Towards this goal, this paper studies the collaborative behavior of groups comprised of editors of articles that have been recognized for their outstanding quality and given the Good Articles (GA) status and those that eventually become Featured Articles (FA) on Wikipedia. The results show that the collaborative behavior of GA groups radically changes just prior to their nomination. In particular, the GA groups experience increases in the level of activity, centralization of workload, and level of GA experience and decreases in conflict (i.e., reverts) among editors. After being promoted to GA, they converge back to their typical behavior and composition. This indicates that crowd behavior prior to their evaluation period is dramatically different than behavior before or after. In addition, the collaborative behaviors of crowds during their promotion to GA are predictive of whether they are eventually promoted to FA. Our findings shed new light on the importance of time in understanding the relationship between crowd performance and collaborative measures such as centralization, conflict and experience.

**ACM Reference format:**
Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel P. Robert Jr., and Daniel M. Romero. 2017. Crowd Development: The Interplay between Crowd Evaluation and Collaborative Dynamics in Wikipedia. *Proc. ACM Hum.-Comput. Interact.* 1, 2, Article 119 (November 2017), 21 pages.
https://doi.org/10.1145/3134754

## 1 INTRODUCTION

Collaborative crowds have become a staple for accomplishing work in our global economy [53, 54, 71, 72]. Thus it is crucial that we investigate how these crowds behave over time and how their behavior affects their performance. Group development theory, which seeks to understand the behavior of small groups over time [9, 11, 21, 38, 43, 62], can be useful in understanding the behavior of collaborative crowds. For instance, previous research on group development suggest that groups can fundamentally change their focus, work structure, and processes before an evaluation [22, 23]. In that sense, the factors identified as key predictors of crowd performance such as centralization

This research was partly supported by the National Science Foundation under Grant No. IIS-1617820.

2573-0142/2017/11-ART119
https://doi.org/10.1145/3134754

and conflict [5, 6, 36, 37, 53, 55] may not represent typical crowd behavior, but rather behavior that is distinct to pre-evaluation stages. Incorporating a group development perspective to the analysis of crowds would help identify other patterns in the behavior of collaborative crowds.

A secondary advantage of studying collaborative crowds through the lens of group development is the potential advancement of group development theory itself. Collaborative crowds pose unique challenges that have not been considered in models of group development. For example, crowds tend to work on multiple tasks together and they have unstable memberships with incoming and outgoing members. In contrast, theories of group development have focused on groups with stable membership performing a single task. Examining non-traditional groups embedded in rapid and dynamic environments can advance the literature in group development.

To advance both our understanding of both collaborative crowds and group development theory, here we studied how the behavior of collaborative groups changes over time in anticipation and as a result of promotion in Wikipedia – one of the earliest, most successful and richest examples of collaborative crowdsourcing sites.

There are a number of reasons for studying Wikipedia crowds. First, Wikipedia is an important domain that provides vast amount of knowledge and information. Second, there is a large amount of research on Wikipedia crowds with which we can compare our findings. Finally, Wikipedia also provides us with a very large set of groups that are working towards a common goal – creating a high-quality article [13, 20, 44]. The vast scale is crucial for attaining statistical power.

We examine in this paper the collaborative behavior of groups comprising of editors of Wikipedia articles awarded the Good Article (GA) status – one of the highest recognitions an article can get based on its quality. The English Wikipedia maintains an organized quality class structure for articles, with the requirements for GA explicitly defined. Examining the promotion toward GA serves as a unique opportunity to study collaborative dynamics as groups face different stages in their collaboration - preparation for evaluation, the evaluation, and post evaluation - and extrapolate our findings to other context (e.g., Wikipedia in other languages). While the GA status is assigned to articles of outstanding quality, it is also used by the editors who work on them to highlight the value of their contribution. We analyze how crowds change as they encounter these phases and whether their change in behavior is predictive of future performance.

We consider four types of collaborative measures among the group of editors: 1) level of activity such as the number of revisions, the number of editors and the size of revisions 2) centralization of contribution measured by the extent to which workloads are distributed in the group – for example, the least centralized group would have all its revisions equally distributed among all its editors, while the more centralized group would have one editor perform most of the workload 3) conflict among editors measured by the fraction of reverts and revert chains among all revisions and 4) experience editing GA among the editors in the group. To compare the behavior of GA crowds against the counterfactual of not being nominated or promoted, we pair each GA with a non-GA article with similar features prior to its nomination using the method of propensity score matching.

The results show that the collaborative behavior of the GA groups radically change prior to nomination and lend support to the group development theory. In particular, the GA groups experience increases in their levels of activity, centralization of group workload, levels of GA experience and decreases in conflict among editors. Moreover, after being promoted to GA, the groups converge back to their typical behavior and group composition. This indicates that crowd behavior prior to an evaluation period is dramatically different from the behavior before or after. We further consider whether and how the collaborative dynamics of GA can influence an article's promotion to FA. Specifically, we find that the extent to which an article grows after being recognized as GA has a sizable impact on its eventual promotion to FA. As a result, our findings shed new light

about the importance of time in understanding the relationship between crowd performance and collaborative measures such as centralization, conflict and experience.

Our study makes three contributions to the literature: (i) we identify how the evaluation of crowds can dramatically influence their behavior and composition; (ii) we show that some changes in behavior due to the evaluation are temporary and crowds go back to their pre-evaluation behaviors; and (iii) we demonstrate how changes after successfully going through one evaluation can impact the future success of crowds. Our results inform the design of future crowd platforms that need to account for the potential rapid changes in crowd composition and behavior.

## 2 RELATED WORK

### 2.1 Group Development

Numerous researches have been dedicated to understanding how groups behave over time and how their dynamics affect their performance [9, 11, 21, 38, 43, 62]. Among these, few recognize that the behavior of groups fundamentally changes just prior to an evaluation period [17, 38] or more generally over time [33, 59]. Most of these models are based on Gersick's punctuated equilibrium [22, 23], which suggests three stages of group development. In the first stage, groups attempt to define their goals but do little else in terms of accomplishing work. At this stage, group members are not concerned about deadlines because they believe they have more than enough time to accomplish the work. The second stage occurs at the midpoint approximately halfway between their start date and their project completion date. Groups assess their progress and evaluate what needs to be done followed by a sense of urgency. They are now concerned about the deadline and group members begin to focus and prioritize their work. It has been shown that this stage is characterized by conflict as groups struggle to ensure that every member takes the deadline and the groupâĂŹs work seriously [31]. The third stage is characterized by changes in the group behavior to facilitate the accomplishment of the project. This often involves the removal of unsuccessful approaches to work, the adoption of new approaches and further increases in effort to meet the deadline.

Our study differs from previous work in group development in three important ways. First, we study crowds whose compositions evolve over time as opposed to small groups whose membership are relatively stable. Specifically, we consider a number of collaborative characteristics relevant to how crowds form and work in online spaces. Second, while previous researches on group development commonly focus on group behavior before deadlines, we extend this scope toward group behavior after the achievement of a goal. Finally, our analysis is based on a large set of groups of varying sizes. The richness in the heterogeneity among groups allows us to provide more insights into the collaborative dynamics of group behavior.

### 2.2 Badges

Our study is also related to the literature that examines the impact of badges on the participation of individuals in online communities. Badges are commonly used as a mechanism for rewarding participants for their achievements in a variety of online communities, including collaborative crowdsourcing platforms such as Wikipedia [1, 40, 42, 51], social media sites such as Foursquare [4], question answering sites such as Stack Overflow [3, 49], news sites such as Huffington Post [32] and educational sites such as Khan Academy [45]. Despite the popularity of badges in various online communities, it has been documented that they can have unintended consequences. For example, participants typically reduce their efforts after receiving their badge [49]. The unintended consequences associated with badges have motivated the problem of optimal assignment of badges [3]. Our study contributes to this strand of literature in two ways. First, while previous researches

focus on badges awarded to individuals, we consider badges awarded to a *crowd* contributing to an article rather than an individual. Second, while badges are usually automatically assigned by the system, the recognition of GA at Wikipedia is assigned through a peer-review process and is more than a symbolic item – the GA status allows readers to better interpret the quality of an article.

### 2.3 Goal Setting

Our study is also related to a large body of literature that investigates the effectiveness of goal-setting. Some studies find subgoals to be useful in achieving the ultimate goal [8, 41], while others find that motivation to achieve the subgoal can at times distract from the ultimate goal [2, 18, 19, 28]. This line of work is of particular relevance to our analysis on whether GA groups ultimately receive the highest FA status. Yet, there are clear distinctions between the previous studies and the goal of this paper.

First, most studies consider goals and subgoals of individuals while we consider collaborative crowds. There is one notable exception, however. [74] studies the effect of group identification and direction setting on the amount of effort put in by the crowdsourcing platform volunteers in the context of Wikipedia's Collaborations of the Week. They focus on two types of direction setting—explicit direction based on publicized group goals and implicit direction based on role modeling. Unlike this study, which focuses on the impact of direction setting and group identity on the *individual*, in this paper we examine the behavior of *groups* as they work towards a goal.

Second, most studies focus on the mere setting of subgoals and how that affects achieving the ultimate goal. Although a handful of studies have investigated what types of behaviors result in success (e.g., [19]), their focus is still on how the subgoal is phrased and not on how the subgoal was approached or the relationship between the behavior change during or after the subgoal task and future success. In this paper, we tackle the latter problem.

## 3 BACKGROUND AND DATA

### 3.1 Wikipedia Grading Scheme

Wikipedia has a grading scheme that gives articles a label based on their quality. There are seven possible quality categories. In increasing order of quality, these categories are *Stub, Start, C, B, GA* (i.e., Good Article), *A*, and *FA* (i.e., Featured Article). The requirements to belong a category range from "a little more than a dictionary definition", for the Stub category, to "Professional, outstanding, and thorough; a definitive source for encyclopedic information", for the FA category.

The top three categories require both a nomination process and a peer review process. Specifically, GA and FA require a stricter and wider peer-review process, whereas A class allows peer reviews from editors within the same WikiProject. Under the policy of Wikipedia, any Wikipedia editor can nominate an article to be a GA and any registered user who has not edited the article but has the relevant knowledge and experience with the content policy of Wikipedia can serve as a reviewer for a nominated article. Once an article is nominated for GA, it enters a queue to be reviewed. During the review process, reviewers can suggest revisions and the editors of the article are encouraged to address them accordingly. Once the reviewers are satisfied with the quality of the article, the article is promoted to GA.

There are six basic criteria to become a GA article: (i) well-written (ii) verifiable with no original research (iii) broad in its coverage (iv) unbiased (v) stable in content over time and (vi) illustrated by images. While there is no official deadline for an article to become GA, it is likely that editors join efforts and plan a timeline to nominate an article. According to the instructions to nominate an article for GA, the nominator or the article must consult regular editors of the article, and editors must be available to respond to reviewers' comments in a timely manner. This suggests that some

planning among editors is required to nominate an article. Detailed descriptions of each criteria as well as an in-depth description of the nomination and review process are available on Wikpedia's Good Articles page.[1]

Although the standard for becoming a GA is fairly high, the highest possible distinction for a Wikipedia article is the FA category. FAs are considered the best articles available on Wikipedia and becoming FA requires a more extensive peer-review and higher quality[64]. Thus, while obtaining the GA status is a significant recognition for the editors of the article, the ultimate goal for any article is to eventually become FA. In this sense, the GA category, while significant, is only an intermediate goal for Wikipedia articles.

### 3.2 Data

Our data consist of all articles in English Wikipedia up to December 2016. Because GA and FA procedures were not consistently documented prior to January 2007, our analysis focus on group behavior after January 2007.

**Revision history**. We collect the metadata of every revision performed on every article. The metadata includes timestamp, editor ID, the change in number of bytes of the article after each revision, whether the revision is a revert of another one, and a short comment describing the revision if the editor chooses to add one. Our data on revision history consists of 12,099,465 articles, each with 36 revisions on average.

**Nomination and promotion**. We collect the history of articles that were nominated to and promoted to GA. Starting in 2008, Wikipedia maintains a bot account called Legobot[2]. Whenever an editor nominates or promotes an article as advised by GA instruction, Legobot records this information and updates the GA nomination history accordingly. We collect the time of nomination and the time of promotion for the GAs that are recorded by Legobot. Legobot starts archiving nomination and promotion to GA in 2008 and has consistently updated GA related information since September 2010. Our dataset contains information about 22,225 articles that have been promoted to GA up to January 2017. Among these, we exclude those that 1) do not have the date of nomination or the date of promotion recorded, or 2) are later retracted from GA, or 3) redirected to another page. Therefore, our data set contains 9,842 articles that are nominated for and promoted to GA. We also collect a list of articles that have been promoted to FA and the time they were promoted. Among all the GAs in our dataset, 820 have been promoted to the FA status.

**Predicted quality**. Many articles may meet all the criteria of GA articles, but they have not been nominated for GA and hence do not have the label. In our analysis, it is crucial to measure an article's quality level, regardless of whether it is promoted. To that end, we take an automated approach [65, 66] and use Objective Revision Evaluation Service (ORES), a service provided by the Wikimedia Foundation that automatically generates a score for the quality of any version of an article [27]. The score can be interpreted as the probability that an article meets the requirement to successfully go through the GA nomination and promotion process. For each article and each day on our data set, we use ORES to compute the quality score of the article. [3]

[1] https://en.wikipedia.org/wiki/Wikipedia:Good_articles

[2] https://en.wikipedia.org/wiki/User:Legobot

[3] Note that ORES is not the only attempt at automatically inferring the quality of Wikipedia pages. For instance, [69] uses metrics based on the editing intensity throughout the entire existence of an article. This method was not directly compared with ORES. We chose to use ORES quality assessment because (i) it performs evaluation based on a broader definition of quality while [69] only aims to classify Featured Articles (FAs) versus other articles, and (ii) the set of features used for classification is more comprehensive for ORES.

## 4 COLLABORATION AND TEAM COMPOSITION FEATURES

We consider four types of collaboration and team composition features: level of activity, centralization, conflict, and crowd experience. We focus on these measures because they have been found to vary in articles that undergo significant exogenous shocks [73] and in articles of different quality levels [53]. Next we describe how we measure each feature.

**Level of activity**. We measure the level of activity of an article along the extensive margin and the intensive margin, which capture whether to contribute and how much to contribute. On the extensive margin, we focus on the number of revisions and the number of editors. On the intensive margain, we focus on the size of revisions in bytes. For each article $a$, we let $E_t^a$ and $W_t^a$ be the number of revisions made on the article and the number of users who edited the article from time the article was created up to day $t$, respectively. We let $B_t^a$ be the size of the article in bytes on day $t$.

**Centralization**. In many crowd-work and online settings, including Wikipedia, a few people make a large percentage of the contributions [56], which we refer to as centralization. In the case of Wikipedia, centralization can be beneficial because it reduces the cost of coordination [35]. However, it might also reduce the diversity of the expertise and viewpoints of editors who contribute to the article, because fewer people are contributing most of the work.

We measure centralization using the Gini coefficient of the distribution of the number of revisions per editor. The Gini coefficient measures the level of inequality in a distribution [16]. Following the centralization measure used in [73], we normalize the Gini coefficient of an article at a given time by the maximum possible Gini coefficient given the number of editors and edits of the article. This normalization prevents the centralization measure to be driven by the volume of activity of the article. For each article $a$, we let $C_a^t$ be the centralization of article $a$ on day $t$, taking into account the full history of the article from its creation until day $t$.

**Conflict**. Reverts are revisions that undo the changes that some other editors previously made. Although reverts are initially intended to eliminate vandalism such as deleting the entire article, they also reflect conflicting viewpoints and disagreement among editors. We follow prior work in using reverts to measure the level of conflict among editors of an article [60, 61, 63]. For each article $a$, we let $R_t^a$ be the fraction of reverts in the article from the creation of the article until day $t$.

While fraction of reverts has been primarily employed as a measure for conflict, it might not reflect precisely the extent to which editors disagree with one another. For example, some reverts might only clean up revisions that are spams. We further measure conflict by the fraction of revert chains. The rationale behind this is that reverts dedicated to spam cleanups are unlikely to be reverted again and if a revision is reverted back and forth, that might be largely due to conflict of opinions rather than cleaning up spams. Procedurally, we focus on revert chains of at least length 2, i.e., reverts that revert the other reverts or reverts that are themselves reverted.[4]

**Level of experience**. An important part of being successful in promoting an article to GA is to be familiar with Wikipedia's reviewing procedures and content policies. We measure the *GA experience* of an article's editors through the extent to which the editors have had experience editing articles that eventually are promoted to GA. That is, experience is measured not by only the quantity of work editors have contributed, but also by the quality of their work, which has been found to be predictive of future article quality [59]. Because we know that not all editors contribute equally to an article, we weight the fraction of experienced editors by their contribution to the focal article as well as their contributions to other articles that are eventually promoted to GA.

For each editor $e$, we let the *editor GA experience* of $e$ at time $t$ be the fraction of edits that $e$ contributed to articles that eventually were promoted to GA before their promotion, over the total

[4] Among all the reverts, 62.2% belong to a revert chain of length 1 (never reverted again), 19.5% belong to a revert chain of length 2 (reverted once), 7.0% belong to a revert chain of length 3, and the rest belong to a longer revert chain.

number of edits $e$ has contributed to Wikipedia. In essence, editor GA experience measures what fraction of an editors efforts have gone into preparing articles to become GA. We then let the *article GA experience* of an article $a$ at time $t$, $X_t^a$, be the average editor GA experience among all editors of $a$ weighted by the fraction of edits that they contributed to $a$ before time $t$.

We use these four features – level of activity, centralization, conflict, and level of experience – to describe the collaboration and crowd composition dynamics of an article. In order to identify a control group, we further consider machine predicted quality and sentiment in edit comments.

**Machine predicted quality**. For each article $a$, we let $Q_t^a$ be the predicted quality of the article on day $t$ using the Objective Revision Evaluation Service (ORES) as described.

**Sentiment in edit comments**. As an additional way to measure potential conflict and emotion in the communication of the editors, we apply sentiment analysis to the comments that editors make when they edit the article. We use the Linguistic Inquiry and Word Count (LIWC2015) text analysis tool to measure the level of positive and negative sentiment in each comment [50]. LIWC essentially identifies words that carry positive and negative sentiment and takes the proportion of words of each type in the comment. For each article $a$, we let $P_t^a$ and $N_t^a$ be the average positive and negative sentiment in the comments made on edits up to day $t$, respectively.

While there are many techniques to measure sentiment in text [7, 10, 26, 30, 52, 67], we choose LIWC because it has been well-validated and frequently used [14, 39], particularly for measuring sentiment in short text similar to edit comments such as small segments of blogs and instant messages [25, 57]. However, we acknowledge that LIWC and other lexicon based approaches have drawback including their inability to detect ambiguous use of words and their non-exhaustive lexicons. When used in specific domains such as Wikipedia, lexicon based approaches can miss specialized words that indicate sentiment within the domain, but typically not outside.

We use these collaborations and composition features for two purposes. First, we use them to create a matched sample to be use as control group to compare the collaboration dynamics of GAs with non-GAs that have similar features. Second, we use them to track the dynamics of collaboration before, during, and after the review process for GAs. We use all features for the matching step, but based on prior findings on crowd dynamics on Wikipedia [73], we only use level of activity, centralization, conflict, and experience for the second step.

## 5 PROPENSITY SCORE MATCHING

Evaluating the impact of reaching GA status on an article's collaborative dynamics requires imputing its unobserved counterfactual condition from the outcomes of other observed articles. To that end, we employ the method of propensity score matching to control for pre-existing differences between GAs and non-GAs and potential selection effects. For each article nominated for and promoted to GA, we identify a non-GA - which exhibits similar features that might influence the nomination and promotion process - as its control.

Propensity score matching procedure involves two steps. First, we estimate the propensity score for each article using a set of covariates. Given these covariates, the estimated propensity score captures the probability of an article being promoted to GA. Then, we implement the nearest-neighbor matching by pairing each GA with the non-GA that has the closest propensity score. To the extent that the conditioning covariates capture the selection of GAs, matching on propensity score allows us to control for multiple covariates in a one-dimensional space. The paired article serves as a control to impute the unobserved counterfactual and evaluate the impact of being GA [5].

[5]It is worth noting that while propensity score matching allow us to control for possible counterfactuals to some extent, it does not provide definite evidence of causality [34]. Thus, further experimental evidence is required to establish causal relationships between our variables.

To estimate the propensity score, we specify the probability of an article being GA in week $t$, $p_t^a$, as a logistic function of a set of covariates. The rationale behind conditioning on these covariates is two-fold. First, it controls for the differences in the collaborative behavior that exist prior to nomination. Hence, we include in the conditioning covariates the measures that we aim to study: the number of revisions, the number of editors, bytes, fraction of reverts, Gini ratio and level of GA experience. Second, we need to control for potential selection effects that might influence whether an article is nominated for GA. For example, the criteria for GA requires broadness in coverage and stableness in coverage, which further lends support for including level of activity and fraction of reverts. To further improve the estimation of propensity score and quality of matching, we include in the covariates the ORES score and LIWC score, both of which provide measures for the quality and controls for the potential bias that articles of higher quality are more likely to be nominated. Overall, the propensity score is specified as:

$$p_t^a = \text{logit}(E_t^a, W_t^a, B_t^a, C_t^a, R_t^a, X_t^a, Q_t^a, P_t^a, N_t^a)$$

Our original dataset contains more than 170,000,000 observations and provides an extremely unbalanced sample. Specifically, the GAs represent no more than 0.01% of all observations. Applying logistic regressions and matching on such a large and unbalanced sample will be computationally untenable and have a negative impact on the estimated propensity score. To mitigate the negative consequence of class unbalance, we apply random downsampling [12] to the non-GAs before performing propensity score matching. We select a random 1% sample from articles which are not GAs and retain with the entire sample of GA articles. Combining these two samples yields our final set with 1,679,912 observations. We then perform the matching procedure on this set.

Table 1 provides the results of logistic regressions, including the estimated coefficients, standard errors and p-values. The level of activity is, in general, positively correlated with an article's GA status. In particular, articles that receive more revisions and revisions with larger sizes are more likely to be nominated for and promoted to GA. Moreover, those with a larger fraction of reverts have a lower propensity score, which is consistent with the criteria on stableness for GA. A likelihood ratio test rejects the null model (p-value $< 0.1\%$), indicating that our specification provides a good characterization on the probability of GA.

We then calculate the propensity scores using the estimated parameters. We use the nearest neighbor matching by pairing each GA to the non-GA with the closest predicted propensity score. Table 2 summarizes the balance check between the GAs and the non-GAs matched to them. For each collaborative measure, we provide the mean and standard deviation for the GAs and the non-GAs in the first four columns. The fifth column gives the normalized difference between the two groups and indicates how GAs are different from their matched non-GAs on average. The last column gives the log of the ratio between the standard deviation, which measures whether one group exhibits a larger dispersion than another. The results show that the normalized differences for all relevant features are not statistically different from zero, and the log ratio of standard error are comparable for the GAs and their matched non-GAs, indicating that the common support condition is satisfied.

## 6 RESULTS

We present the results of our analysis on the change in collaborative dynamics - activity, centralization, conflict and experience - resulting from the transition toward GA status. For all subsequent analysis, we consider three relevant periods: 1) 100 days before nomination, 2) review period, and 3) 100 days after promotion. Our decision to study the 100-day period is made to balance the length of collaborative dynamics to be examined and the number of observations included. Among the 9,842 GAs, 13% are nominated 50 days after creation, 4% in 50-100 days after creation, 2% in 100-150

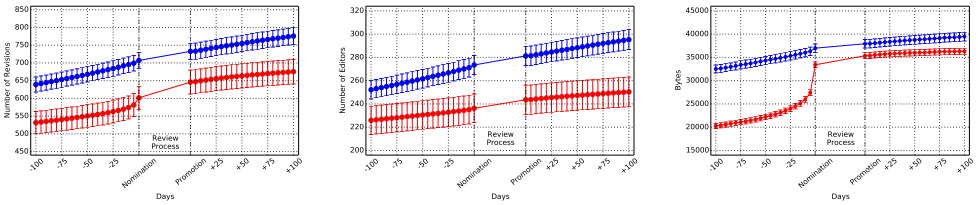
| | coefficient | s.e | p-value |
|---|---|---|---|
| Number of revisions | 0.0003 | 0.0002 | 0.056 |
| Number of editors | -0.0008 | 0.0001 | 0.000 |
| Bytes | 0.00001 | 0.0000 | 0.000 |
| Fraction of reverts | -16.2439 | 0.5360 | 0.000 |
| Centralization | -0.3156 | 0.0540 | 0.000 |
| Level of experience | 16.7381 | 0.1170 | 0.000 |
| ORES score | 6.3026 | 0.0470 | 0.000 |
| Positive sentiment | 0.5980 | 0.1910 | 0.002 |
| Negative sentiment | -1.2076 | 0.1641 | 0.462 |
| pseudo $R^2$: 0.519 | | | |
| log likelihood: -29038 | | | |
| p-value for likelihood ratio test: 0.000 | | | |

The intercept term is suppressed.

Table 1. Logistic regressions of GA status over collaborative features.

| | GA | | non-GA | | | |
|---|---|---|---|---|---|---|
| | Mean | Std. | Mean | Std. | Nor. Dif | Log Ratio of Std. |
| ln(number of revisions) | 5.286 | 1.968 | 5.896 | 1.516 | -0.462 | 0.261 |
| ln(number of editors) | 4.200 | 2.391 | 4.872 | 1.759 | -0.467 | 0.307 |
| ln(bytes) | 10.160 | 0.509 | 10.199 | 0.614 | -0.052 | -0.188 |
| Fraction of reverts | 0.027 | 0.001 | 0.029 | 0.001 | -0.061 | 0.494 |
| Centralization | 0.424 | 0.014 | 0.432 | 0.010 | -0.067 | 0.337 |
| Level of Experience | 0.166 | 0.006 | 0.151 | 0.013 | 0.157 | -0.712 |

Table 2. Balanceness of propensity score matching.




(a) Number of revisions (b) Number of editors (c) Number of bytes

Fig. 1. Dynamics of level of activity over the three period. The red line and the blue line represent the GA and the matched non-GA, respectively.

days after creation, and the rest beyond 150 days. Though extending the 100-day period to a longer one could allow us to study collaborative dynamics over a longer period, it would exclude more

| | non-GA | | | GA | | |
|---|---|---|---|---|---|---|
| | Before nomination | Review process | After promotion | Before nomination | Review process | After promotion |
| Number of revisions | 0.642 | 0.623 | 0.480 | 0.466 | 2.468 | 0.299 |
| Number of editors | 0.204 | 0.169 | 0.155 | 0.089 | 0.313 | 0.070 |
| Bytes | 41.697 | 35.706 | 17.149 | 64.107 | 120.289 | 9.258 |
| Fraction of reverts($10e-4$) | 0.090 | 0.070 | 0.070 | -0.100 | -1.410 | 0.050 |
| Centralization($10e-4$) | 1.360 | 3.650 | 0.350 | 3.790 | 87.480 | 0.710 |
| Level of Experience | 0.270 | -0.260 | -0.430 | 0.670 | 9.600 | -0.520 |

Table 3. Weighted average of slopes from the regression parameters.

than 20% observations. In addition, we also perform analysis on 50-day and 150-day window, and the results are not qualitatively different.

Among the articles studied in our sample, there exhibits a large amount of variation in the number of days it takes a GA article to be promoted after nomination. Analysis without controlling for the number of days in the review period will bias for the results. To mitigate the potential biases introduced by variations in the number of days in the review period, we perform our regression analysis by allowing the relevant coefficients to vary with the length of review period. For each collaborative measure, we employ the following regression framework:

$$y_{it} = \sum_{l} \mathbb{I}\{g_i = l\} f(\mathbb{I}_t\{\text{pre}\}, \mathbb{I}_t\{\text{post}\}, \mathbb{I}_i\{\text{GA}\}, t)$$

where $y_{it}$ is the outcome measure for article $i$ on day $t$, $g_i$ denotes the number of days in the review period for article $i$, $\mathbb{I}_t\{\text{pre}\}$ is an indicator function that equals to 1 if day $t$ is before nomination for article $t$, $\mathbb{I}_t\{\text{pre}\}$ is an indicator function that equals to 1 if day $t$ is before nomination, $\mathbb{I}_t\{\text{post}\}$ is an indicator function that equals to 1 if day $t$ is after promotion. $f(\mathbb{I}_t\{\text{pre}\}, \mathbb{I}_t\{\text{post}\}, \mathbb{I}_i\{\text{GA}\}, t)$ is a function that includes the all possible interaction terms. Therefore, our regression framework nests all possible length of review period and allow articles with review period of different length to have different slopes and intercepts. Note that because of the variation in the length of review, we do not include the dynamics at a daily level within this period. Moreover, we report the estimated slope weighted from the back-of-the-envelope calculation from the regression results in table 3. We illustrate the dynamics for the collaborative measures over the three periods in figure 1 through 4. In each figure, we plot the corresponding measure as well as the error bars denoting 1.96 standard error (i.e., 95% confidence interval) for the GAs (red) and the non-GAs (blue).

*Activity.* We measure the level of activity of an article both on the extensive margin - as measured by the number of revisions and the number of editors, and the intensive margain - as measured by the number of bytes. Figure 1(a) through 1(c) illustrate the evolutionary dynamics of article activity before nomination and promotion. Prior to nomination, the GAs exhibit diversified patterns between the intensive margin and the extensive margin, compared with their matched non-GAs. On the intensive margin, we find that articles promoted to GA status increases at a faster rate on the intensive margin - 64 bytes per day as opposed to 42 bytes per day. On the extensive margin, however, the GAs appear to grow slower in the number of editors (0.09 per day versus 0.20 per day) and the number of revisions (0.47 per day versus 0.64 per day) than their non-GA counterparts. During the review period, the GAs are consistently more active than their matched non-GAs. Specifically, the GAs gain 0.31 new editors, 2.47 new revisions and 120.29 bytes on average per day.

On the other hand, the matched non-GAs gain 0.17 new editors, 0.62 new revisions, and 35.71 bytes per day. After the articles are promoted to the GA status, they grow at a slower speed than the non-GAs.

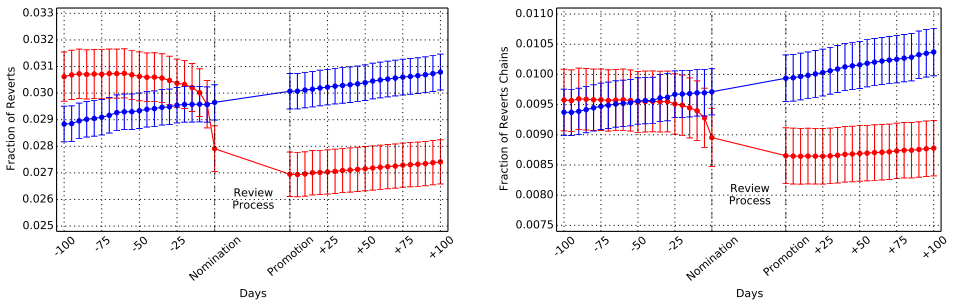
One notable pattern in the dynamics of the GAs is that as the articles approach the date of nomination, there is a drastic increase in the number of bytes. Among all the bytes the GAs gain in the 100 days prior to nomination on average, more than 50% are made in the last week and more than 25% are made in the last two days. In contrast, the increase in the average bytes of the matched non-GAs in the last week prior to the relative nomination day only represent 15% of the entire progress. Such a pattern indicates that to make the articles qualified for GA status, the editors are involved in intensive editing work as the nomination day is approaching.

*Conflict.* We analyze the conflict in the collaborative crowds by examining the fraction of reverts. As a benchmark, the matched non-GAs experience a gradual increase in the fraction of reverts over the time frame. The estimated slopes in the three periods are not significantly different from one another, and the estimated discontinuities on the nomination day and promotion day are not significantly different from zero.

Figure 2(a) shows that the GAs present a prominently different pattern in dynamics of conflict as compared to the matched non-GAs. Over the 100-day window prior to nomination, the GAs experience an overall decrease in the fraction of reverts. In particular, we find that the level of conflict is relatively stable in the first 50-day window and starts decreasing when it comes to the second 50-day window. Specifically, the fraction of reverts drops from 3.0% to 2.8% in the last 10 days prior to nomination.

Moving from the preparation period to the review period, we find that the GA articles continues to reduce conflict - the fraction of reverts drops from 2.8% to 2.7%. After the articles are promoted to the GA status, the fraction of reverts starts to increase.

We also provide the dynamics of conflict measured by revert chains in figure 2(b). We see that both measures exhibit a similar pattern. Prior to nomination, the fraction of revert chains experience an overall decrease. Specifically, as the date of nomination approaches, fraction of revert chains declines drastically. During the review period, fraction of revert chains continues to decrease from 0.9% to 0.85%. After the article is promoted to GA, fraction of revert chains steadily increase.




(a) Fraction of Reverts

(b) Fraction of Revert Chains

Fig. 2. Dynamics of conflict over the three period. The red line and the blue line represent the GA and the matched non-GA, respectively.

Overall, the dynamics of conflict measured by both the fraction of reverts and that of revert chains generally indicate that the collaborative group tends to decrease conflict prior to promotion,

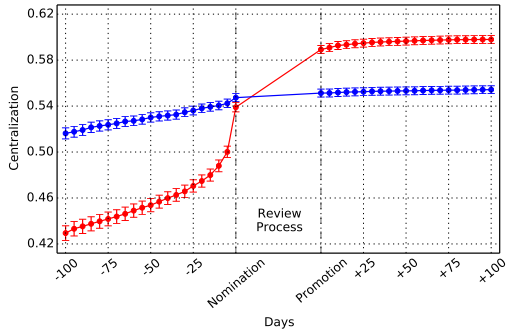



Fig. 3. Dynamics of centralization over the three period. The red line and the blue line represent the GA and the matched non-GA, respectively.

perhaps hoping to help the article go through the review process. After promotion, the level of conflict increases.

*Centralization.* We next study how editors distribute the works among themselves. Figure 3 illustrate the dynamics of centralization for the GAs and their matched non-GAs. As time goes on, articles of both types becomes more centralized - though in a different rate from each other. The matched non-GAs starts with a normalized Gini coefficient of 0.53 at the start of the time frame we study and this reaches 0.55 eventually on average. Compared with the centralization in the review period, the slopes for centralization before nomination and after promotion are not significantly different.

The GAs experience a more drastic increase in centralization than their matched non-GAs. Over the time frame we study, centralization of the GAs increases from 0.47 to 0.60 on average. Our regression results indicate that the slopes for centralization is significantly larger for the treated articles than those for the non-GAs for all three periods.

Notably, the centralization of the GAs experience a prominent increase in the last 5 days prior to nomination from 0.51 to 0.54. One explanation for such a pattern is that to qualify the article for GA nomination, a small group of core editors are intensively involved with the editing work and therefore contributes to a drastic increase in centralization.

*Experience.* Our last measure for collaborative dynamics is the editors' experience for GAs. Figure 4 illustrates the evolution of GA experience over the time frame we study. The matched non-GAs start with 0.17 at 100 days before nomination and this increase steadily. After the nomination, the non-GAs experience gradual decreases. A comparison of the slope in the review period to those before nomination and after promotion indicate that the trends of GA experience are not significantly different from each other.

The articles promoted to GA status appear to behave differently from the non-GAs matched to them. Firstly, prior to nomination, the GAs experience fast increase in the editors experience of GAs. This is consistent with our previous findings on the measure of bytes and centralization - to qualify the articles for GA nomination, a group of experienced editors work more intensively.

During the review period, we find that the editors' GA experience continues to increase, though at a slower speed than it does before nomination. After the article is promoted, the editors' experience for GAs starts decreasing.

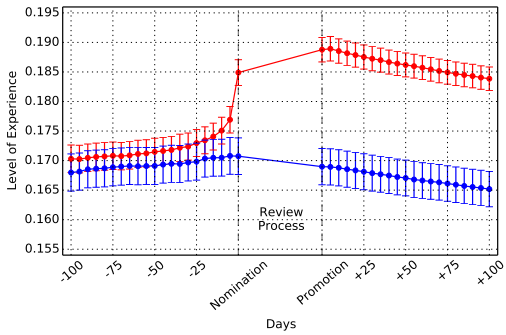



Fig. 4. Dynamics of level of experience over the three period. The red line and the blue line represent the GA and the matched non-GA, respectively.

## 7 FEATURED ARTICLE PREDICTION

Until now our analysis has focused on distinct patterns in the collaborative dynamics of articles that receive GA recognition compared to articles that do not. We have yet to look, however, at how these collaborative measures following the production period affect an article's likelihood of reaching FA, the highest label of recognition at Wikipedia. We use logistic regression analysis to identify which of the collaboration metrics and their respective patterns post production period are most predictive of an article's future success in becoming FA. Our model ultimately uses features associated with group dynamics measured in different stages to predict the binary outcome of whether or not an article that received the Good Article recognition at time $t$ will receive the Feature Article recognition in $(t, t + \Delta_t]$. Because we are most interested in the time period directly after Good Article promotion and its subsequent effect, we focus our analysis on predicting FA status in three time periods ($\Delta_t$): three months, six months, and one year.

*Feature Selection.* To determine which features should be used to build our model, we first consider all collaboration metrics at the time of GA promotion and their respective slopes between promotion and 100 days after promotion. Three measures (number of revisions, number of bytes, and number of editors) presented high risk of multicollinearity. To avoid this issue, we include only the number of revisions and the most informative of the three as measured by variance explained, and we remove the other two from our model. The final list of independent variables can be found in the first column of Table 4.

*Training and testing.* To gain a more robust understanding of the predictive power of our model, we use ten-fold cross-validation. Given the imbalance of classes (only a small fraction of Good Articles end up being featured), we construct a balanced dataset by randomly down-sampling GAs that do not make it to FA. This technique has been shown to increase classifier performance for imbalanced datasets [12]. Note that, unlike the earlier analysis where the dataset used include all Wikipedia articles, the model here is estimated using only articles that attain GA status and are of high quality. Therefore, confounding factors relating to quality prior to treatment is less of a concern. Hence, we do not rely on propensity score matching in this analysis. The accuracy of the model is tested through F1-score, precision and recall.

*Results.* Based on our model, we identify a number of features highly predictive of an article's future FA success in the 3 months, 6 month, and 1 year period its GA promotion. We find that we can predict with an F1-score of 85.34%, 78.64%, and 73.83%, precision scores of 86.72%, 82.10%, and

| | 3 months | | 6 months | | 12 months | |
|---|---|---|---|---|---|---|
| | **Estimate** | **Std. Error** | **Estimate** | **Std. Error** | **Estimate** | **Std. Error** |
| **Intercept** | -3.0645 | 0.616 | -2.7866 | 0.494 | -2.6814 | 0.456 |
| **Gap** | -0.0057*** | 0.002 | -0.0053*** | 0.002 | -0.0025* | 0.001 |
| **Revisions at Promotion** | -0.0004*** | 0.000 | -0.0002* | 7.51e-05 | -0.0003*** | 8.32e-05 |
| **Reverts at Promotion** | 4.4294 | 3.759 | 4.9824 | 2.955 | 6.0329* | 2.564 |
| **Centralization at Promotion** | 3.7282*** | 0.813 | 3.4153*** | 0.644 | 3.2287*** | 0.586 |
| **Experience at Promotion** | -2.9883* | 1.375 | -1.2952 | 1.095 | -1.2908 | 1.011 |
| **Revisions Slope** | 3.1146*** | 0.370 | 1.8400*** | 0.248 | 2.2841*** | 0.261 |
| **Reverts Slope** | 445.7330 | 1377.598 | 1174.1174 | 1246.238 | 1570.4206 | 1280.819 |
| **Centralization Slope** | 1932.0713*** | 469.274 | 1949.9771*** | 385.368 | 1400.7422*** | 347.432 |
| **Level of Experience Slope** | 1402.9088 | 948.627 | 1271.1011 | 751.487 | 1596.2255* | 733.946 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4. Logistic regression results for predicting whether an article will become FA after 3, 6, and 12 months of promotion to GA.

78.21%, and recall scores of 85.47%, 79.15%, 74.67% for FA success in three months, six months, and one year respectively. Table 4 provides the regression results. Below we provide insights gained from feature weights estimated:

(1) Centralization: Out of all the features (and across all time periods), centralization (as measured by the Gini score) and change in centralization (as measured by the slope of the Gini score between GA promotion and 100 days out) is most predictive of future FA success. This suggests that increasing centralization is important for feature article success; with few people distributing the work, less coordination is required, which gives articles a distinct advantage in achieving FA status.
(2) Activity: For the three 3-month and 6-month periods we observe that the effect of the number of revisions is negative and significant, where the effect size of the slope of the number of revisions is positive and significant across all time periods. These results imply that an article is more likely to achieve FA when the number of revisions, or activity of the article, is at first small, but increasing across the specified time period. This suggests that an article does in fact require extensive additional work following GA promotion to ultimately reach FA recognition, even though the activity subsided at the time of GA promotion for most articles.
(3) Conflict: Although the fraction of edits that are reverts is positively and significantly predictive of FA status for the one year period, the slope of the fraction of reverts across all time periods, and the fraction of reverts at the time of promotion for the 3-month and 6-month periods are not significantly correlated with becoming a FA. Thus we conclude that conflict is not a consistent indicator of future article success—the predictive power is observed only for early-on conflict (at promotion) when determining success in distant future (1 year out).
(4) Gap: We also find that the longer the gap between GA nomination and GA promotion, the less likely an article is to be awarded FA across all three time periods. We posit that a longer gap might be indicative of lack of interest in a given article which in turn might lower the likelihood of achieving FA.

## 8 DISCUSSION AND CONCLUSION

Drawing on theories of group development, we study how collaborative crowds operate when working towards an evaluation period, how the dynamics shift after the evaluation has passed and the crowds have been recognized, and how crowds' behavior after recognition is associated

with further success. Overall, our results highlight that collaborative crowds rapidly change their behavior before the evaluation period as suggested by models of punctuated equilibrium [22, 23]. Next we discuss our results followed by their implications for research on collaborative crowds and group development research.

*Before promotion to GA.* As a whole, these results are consistent with models of punctuated equilibrium and theories of group development, which would predict that groups fundamentally change before the evaluation period. The finding is also in agreement with [69] who claim that shift to high quality happens in a burst for Wikipedia articles, suggesting the shift requires a coordinated effort. We see evidence of this in the stage leading up to the evaluation when crowds going through promotion to GA increase faster in size (number of bytes), but slower in number of edits and editors, suggesting that editors are making larger edits. They also exhibit a faster increase in centralization and GA experience, and a sharper decrease in conflict, as compared to their non-GA counterparts. Interestingly, a recent study [73] shows that this decrease in conflict and increase in centralization is also observed at times of exogenous shocks and is in agreement with threat rigidity literature [58]. While nomination for GA status is not a *threat*, we believe that the uncertainty of the process might be revealing similar patterns. These findings suggest that the group dynamics during times of preparation for a significant evaluation are similar to those observed during perceived exogenous threats.

*Review period.* During the evaluation period, when crowds are going through the review stage and addressing concerns raised by the reviewers, we see an attenuated version of the same patterns observed during the preparation period. Groups produce fast, keep reducing conflict and increasing centralization – though at small rates compared to the preparation phase. We note that the dynamics during the review period depend not only on the group of editors, but also on the reviewers. Indeed, the amount and type of revisions suggested by the reviewers will impact the editing dynamics of the editors. However, because the direction of the changes match the direction before the nomination, it is likely that the momentum in dynamics from before nomination continues to have an effect during the review period.

*After promotion to GA.* Punctuated equilibrium models and other theories of group development do not hypothesize what would happen to a group after its evaluation. In fact, groups in such studies typically disband after the evaluation. Research on badges has shown that *individuals* do change their behavior after being recognized, but less is known about how groups respond to recognition. The question remains: how do crowds or groups behave after the evaluation period when they are recognized for good performance?

There are at least three alternatives. One, crowds could begin to return to pre-evaluation levels, making anything that happened prior to the evaluation temporary. Two, crowds could maintain their current rates, making pre-evaluation changes permanent, but stable after the evaluation. Finally, crowds could continue to increase or decrease their rate further away from their pre-evaluation levels, such that pre-evaluation changes permanent and growing.

After articles are promoted to GA, we observe that the crowds start looking similar to their counterparts. In fact, they become slightly less active and grow at smaller rates. This finding provides clear parallels to research about the effects of badges on individual participation in online communities (e.g. [3, 49]). We observe that, in addition to reducing activity, in line with past work on badges, groups also change their collaboration mechanisms (e.g. centralization and conflict) and resemble their counterparts that do not go through the GA process. In other words, while the centralization and conflict dynamics of articles that go through the GA promotion process are very different from their control counterparts during the preparation stage, they become very similar

after they obtain recognition as GA. This suggests that going through the GA process does not have a lasting impact on conflict and centralization.

*Promotion to FA.* Punctuated equilibrium and other theories of group development do not consider how a group's past behavior and characteristics influences their future performance. Results related to the promotion to FA, in Table 3, highlight the importance of past crowd behaviors and characteristics on future crowd performance. First, the gap between GA nomination and GA promotion is a significant factor—the longer it takes a crowd to get promoted to GA relative to the time it is nominated the less likely it is to get promoted to FA. Second, prior centralization and current levels of centralization were positive and significant predictors of promotion to FA—centralized efforts with lower synchronization costs are more likely to be promoted to FA. Third, prior revisions and current levels of revisions were also significant predictors of promotion to FA. Our analysis reveals that there is indeed more extensive work to be done to achieve FA and groups that stop growing are less likely to attain it. Finally, results for GA experience and conflict were not consistently significant.

The findings from FA promotion have important implications for the incentives research. Studies, as noted before in the related work section, present mixed findings as to the value of subgoals (see [8, 41] for positive and [2, 18, 19, 28] for negative findings). Here, we observed that the ability of achieving the ultimate goal of FA promotion is not only a function of having a GA subgoal but also a function of how groups reacted to the subgoal (GA) success. This perhaps rectifies the conflicting studies on the value of intermediate goal setting–it is not only about subgoal setting but how groups, or individuals respond to it.

### 8.1 Implications for Research on Crowds

*Time of measurement.* Our results regarding timing have several implications for research on crowds. One, if measures are taken too early they may not be representative of how a crowd actually behaves as it approaches evaluation. For example, in the case of Wikipedia, measures taken 1 month before GA nomination would be much higher in conflict and much lower in centralization and level of activity than they would be immediately before nomination. Two, measures taken during or after the evaluation period would be driven by the evaluation itself. In other words, measures taken of variables like centralization may not necessarily be predictors of better crowd performance but instead be the outcome of the evaluation itself. This introduces endogeneity, which occurs when predictors are correlated with the error term and are often caused when there are issues of mutual causality between the predictor and outcome variable [70]. When this happens, one of the assumptions of ordinary least squares regression (OLS) is violated and another statistical technique has to be used. Finally, studies should clearly indicate when measures were taken relative to the date of article evaluation. This information is vital in order to accurately interpret the results.

*Centralization.* Our study also highlights the very important role of centralization in Wikipedia collaborative crowds, which is consistent with other work [5, 35–37, 53, 55]. We have observed that crowds increase their centralization as they approach evaluation and that centralization is linked to better future performance. This is consistent with the assertion that centralization reduces coordination costs, and thus increases performance, which has been observed in other studies [36, 37]. However, because we also observed that centralization increases before entering an evaluation period, questions remain regarding whether centralization causes performance or whether the evaluation process causes centralization in Wikipedia crowds remains.

In fact, our findings related to centralization might be less applicable to other peer production communities. Every study we find that examine article grade as an outcome has concluded that centralization is a significant predictor of crowd performance. But this could be caused by the

nature of the evaluation process used by the Wikipedia community. For example, studies examining the role of centralization in other collaborative crowds like SourceForge.net has found that the benefits of centralization vary greatly by the stage of the project [15] and that can actually hurt crowd performance [47]. Future research is needed to disentangle the impacts of centralization and the evaluation process of collaborative crowds in the Wikipedia community.

*Conflict.* In our study, we observe that crowds reduce conflict as they approach the evaluation period. However, we also observed that while conflict is typically not related to future performance, there is one exception. Conflict at the time of promotion to GA was positively related to promotion to FA 1 year out. This suggests that, while conflict is typically low when crowds are close to the evaluation period, it could be healthy when the evaluation is far away. Indeed, theories on group development suggest that early conflict could be positively related to performance because it typically deals with decisions about the work that needs to be done in the future and can be resolved easily. However, conflict that occurs later in the process may be the result of lingering problems that were not resolved [22, 43, 48]. Sharp increases in conflict that occur later might be expected to be associated with decreases in performance [23, 31, 46, 68].

### 8.2 Implications for Group Development Theory

Prior theories of group development such as punctuated equilibrium have assumed that groups have stable memberships and clear boundaries and either do not account for leadership or assume a very static and formal type of leadership. Yet, our results indicate that both a turnover in membership and emergent informal leadership, such as members with experience, are vital to understanding changes in crowd behavior prior to an evaluation period. In doing so, this study offers several contributions to the group development theory literature.

One, the impact of membership turnover depends on the composition of the members being added. We found that the addition of more experienced members seems to be a positive. Prior research on group development has not considered membership turnover. However, other studies that have examined membership turnover in groups have found negative effects [24]. The reason for the negative effects given is the disruption caused by the influx of new members. However, our results highlight that the impact of turnover might be largely determined by the composition of the influx of new members. The addition of more experienced members might provide benefits that exceed any potential drawbacks caused by disruptions. Yet, the influx of experienced editors might have come at the expense of less experienced editors. Such influxes can also lead to higher member attrition or the inability to control member-quality. Both point to the challenge of sustaining development in crowds with open-membership, a challenge that may or may not apply to more stable organizational groups. Future research is needed to fully explore how the impacts of membership turnover differs between groups and crowds.

Two, emergent informal leadership appears to be one of the primary factors that appears to be driving behavior change. We see that the concentration of informal leadership (centralization) precedes the behavioral changes in crowds. It is likely that this concentration of informal leadership is needed to effect such rapid behavior changes. If this holds for groups as well as it does for crowds, then it is likely that the concentration of informal leadership can be used to better understand when groups are likely to be successful in changing their behaviors to meet the demands of the evaluation—the more groups can concentrate their informal leadership, the more likely they are to be successful at rapidly changing their groupâĂŹs behavior.

### 8.3 Design Implications

Wikipedia's platform includes features that aim to improve the quality of articles when they face specific circumstances, such as protecting controversial articles by only allowing administrators to edit them [29]. Our study shows that collaboration dynamics change significantly before, during, and after times of evaluation. Furthermore, we show that such changes have implications for future success. This suggests that design choices or new platform features that make these collaboration dynamics more salient to peer-production groups can help them achieve better results and produce better articles faster.

For instance, a dashboard that shows groups their current centralization levels, how such levels have changed over time, and how their centralization compares to other, more successful groups, can help groups strike the right balance and achieve their goals. Similar information can be provided about the other collaboration measures. This dashboard can be presented in a separate tab, much like edit history, on the corresponding article.

The exact design of the collaboration dynamics dashboard will require extensive research. We believe that future work and user studies that engages the Wikipedia community, as well as other collaborative crowdsourcing platforms, can determine effective designs that generate the right nudges and ultimately improve the production quality of collaborative crowdsourcing platforms.

The second important design implication revealed as a result of this study relates to how articles are reviewed. We have observed that there was high variance in the number of days articles spend under review (between nomination and promotion) and articles that spend more time between GA nomination and promotion are less likely to attain FA status later on. A possible explanation for this finding is that editors who have to wait longer for reviewed and promotion are discouraged by the wait and do not continue working towards FA status. Alternatively, it is possible that editors who nominate articles that are not ready for promotion, and hence take longer in the evaluation period, are less likely to write articles of high enough quality for FA status.

Both of these explanations suggests that it would be worth investing effort to make the GA review process more equitable. Dashboards that remind reviewers how long a given article has been under review, how frequently they communicate with the editors of the article, and how that compares to other articles can help improve the review process. Additionally, tools that enable editors to assess whether an article is likely to be ready for nomination, such as ORES [27], can also help avoid long waits for review and promotion.

## ACKNOWLEDGMENTS

This research was partly supported by the National Science Foundation under Grant No. IIS-1617820. We thank Karthik Ramanathan and Shailesh Vedula for their valuable assistance in data collection and Aaron Halfaker for helpful conversations and for his assistance with ORES.

## REFERENCES

[1] B Thomas Adler and Luca De Alfaro. 2007. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 261–270.

[2] On Amir and Dan Ariely. 2008. Resting on laurels: the effects of discrete progress markers as subgoals on task performance and preferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34, 5 (2008), 1158.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 95–106.

[4] Judd Antin and Elizabeth F Churchill. 2011. Badges in social media: A social psychological perspective. In *CHI 2011 Gamification Workshop Proceedings*. ACM New York, NY, 1–4.

[5] Ofer Arazy and Oded Nov. 2010. Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 233–236.

[6] Ofer Arazy, Oded Nov, Raymond Patterson, and Lisa Yeo. 2011. Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems* 27, 4 (2011), 71–98.

[7] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.. In *LREC*, Vol. 10. 2200–2204.

[8] Albert Bandura and Karen M Simon. 1977. The role of proximal intentions in self-regulation of refractory behavior. *Cognitive therapy and research* 1, 3 (1977), 177–193.

[9] Denise A Bonebright. 2010. 40 years of storming: a historical review of Tuckman's model of small group development. *Human Resource Development International* 13, 1 (2010), 111–120.

[10] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*.

[11] Artemis Chang, Julie Duck, and Prashant Bordia. 2006. Understanding the multidimensionality of group development. *Small Group Research* 37, 4 (2006), 327–350.

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[13] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 32–41.

[14] Lorenzo Coviello, Yunkyu Sohn, Adam DI Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A Christakis, and James H Fowler. 2014. Detecting emotional contagion in massive social networks. *PloS one* 9, 3 (2014), e90315.

[15] Sherae Daniel, Ritu Agarwal, and Katherine J Stewart. 2013. The effects of diversity in global, distributed collectives: A study of open source project success. *Information Systems Research* 24, 2 (2013), 312–333.

[16] Robert Dorfman. 1979. A Formula for the Gini Coefficient. *The Review of Economics and Statistics* (1979), 146–149.

[17] Jeff Ericksen and Lee Dyer. 2004. Right from the start: Exploring the effects of early team events on subsequent project team development and performance. *Administrative Science Quarterly* 49, 3 (2004), 438–471.

[18] Ayelet Fishbach and Ravi Dhar. 2005. Goals as excuses or guides: The liberating effect of perceived goal progress on choice. *Journal of Consumer Research* 32, 3 (2005), 370–377.

[19] Ayelet Fishbach, Ravi Dhar, and Ying Zhang. 2006. Subgoals as substitutes or complements: the role of goal accessibility. *Journal of personality and social psychology* 91, 2 (2006), 232.

[20] Fabian Flöck and Maribel Acosta. 2014. WikiWho: Precise and efficient attribution of authorship of revisioned content. In *Proceedings of the 23rd international conference on World wide web*. ACM, 843–854.

[21] Monica J Garfield and Alan R Dennis. 2012. Toward an integrated model of group development: disruption of routines by technology-induced change. *Journal of Management Information Systems* 29, 3 (2012), 43–86.

[22] Connie JG Gersick. 1988. Time and transition in work teams: Toward a new model of group development. *Academy of Management journal* 31, 1 (1988), 9–41.

[23] Connie JG Gersick. 1989. Marking time: Predictable transitions in task groups. *Academy of Management journal* 32, 2 (1989), 274–309.

[24] Cristina B Gibson and Jennifer L Gibbs. 2006. Unpacking the concept of virtuality: The effects of geographic dispersion, electronic dependence, dynamic structure, and national diversity on team innovation. *Administrative Science Quarterly* 51, 3 (2006), 451–495.

[25] Alastair J Gill, Robert M French, Darren Gergle, and Jon Oberlander. 2008. The language of emotion in short blog texts. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 299–302.

[26] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1, 2009 (2009), 12.

[27] Aaron Halfaker and Amir Sarabadani. 2016. Monthly Wikipedia article quality predictions. (10 2016). https://doi.org/10.6084/m9.figshare.3859800.v3

[28] Chip Heath, Richard P Larrick, and George Wu. 1999. Goals as reference points. *Cognitive psychology* 38, 1 (1999), 79–109.

[29] Benjamin Mako Hill and Aaron Shaw. 2015. Page protection: another missing dimension of wikipedia research. In *Proceedings of the 11th International Symposium on Open Collaboration*. ACM, 15.

[30] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[31] Karen A Jehn and Elizabeth A Mannix. 2001. The dynamic nature of conflict: A longitudinal study of intragroup conflict and group performance. *Academy of management journal* 44, 2 (2001), 238–251.

[32] Julie Jones and Nathan Altadonna. 2012. We Don'T Need No Stinkin' Badges: Examining the Social Role of Badges in the Huffington Post. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 249–252. https://doi.org/10.1145/2145204.2145244

[33] Gerald C Kane and Sam Ransbotham. 2016. Research NoteâĂŤContent and collaboration: An affiliation network approach to information quality in online peer production communities. *Information Systems Research* 27, 2 (2016), 424–439.

[34] Gary King and Richard Nielsen. 2016. Why propensity scores should not be used for matching. *Copy at http://j.mp/1sexgVw Download Citation BibTex Tagged XML Download Paper* 378 (2016).

[35] Aniket Kittur and Robert E Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination. In *Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work*. ACM, 37–46.

[36] Aniket Kittur and Robert E Kraut. 2010. Beyond Wikipedia: coordination and Conflict in Online Production Groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. ACM, 215–224.

[37] Aniket Kittur, Bryant Lee, and Robert E Kraut. 2009. Coordination in collective intelligence: the role of team structure and task interdependence. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1495–1504.

[38] Andrew P Knight. 2013. Mood at the midpoint: Affect and change in exploratory search over time in teams that face a deadline. *Organization Science* 26, 1 (2013), 99–118.

[39] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[40] Travis Kriplean, Ivan Beschastnikh, and David W McDonald. 2008. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 47–56.

[41] Gary P Latham and Gerard H Seijts. 1999. The effects of proximal and distal goals on performance on a moderately complex task. *Journal of Organizational Behavior* (1999), 421–429.

[42] David W McDonald, Sara Javanmardi, and Mark Zachry. 2011. Finding patterns in behavioral observations by automatically labeling forms of wikiwork in barnstars. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM, 15–24.

[43] Joseph E McGrath. 1990. Time matters in groups. *Intellectual teamwork: Social and technological foundations of cooperative work* 23 (1990), 61.

[44] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. âĂIJThe sum of all human knowledgeâĂİ: A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 219–245.

[45] Briana B. Morrison and Betsy DiSalvo. 2014. Khan Academy Gamifies Computer Science. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education (SIGCSE '14)*. ACM, New York, NY, USA, 39–44. https://doi.org/10.1145/2538862.2538946

[46] Sean A Munson, Karina Kervin, and Lionel P Robert Jr. 2014. Monitoring email to indicate project team performance and mutual attraction. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 542–549.

[47] Ning Nan and Sanjeev Kumar. 2013. Joint effect of team structure and software architecture in open source software development. *IEEE Transactions on Engineering Management* 60, 3 (2013), 592–603.

[48] Jamie Newell, Likoebe Maruping, Cynthia Riemenschneider, and Lionel Robert. 2008. Leveraging e-identities: The impact of percieved diversity on team social integration and performance. *ICIS 2008 Proceedings* (2008), 46.

[49] Hüseyin Oktay, Brian J Taylor, and David D Jensen. 2010. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*. ACM, 1–9.

[50] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[51] Michael Restivo and Arnout van de Rijt. 2014. No praise without effort: experimental evidence on how rewards affect Wikipedia's contributor community. *Information, Communication & Society* 17, 4 (2014), 451–462.

[52] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (2016), 23.

[53] Lionel Robert and Daniel M Romero. 2015. Crowd size, diversity and performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1379–1382.

[54] Lionel P Robert and Daniel M Romero. 2016. The influence of diversity and experience on the effects of crowd size. *Journal of the Association for Information Science and Technology* (2016).

[55] Lionel P Robert and Daniel M Romero. 2017. The influence of diversity and experience on the effects of crowd size. *Journal of the Association for Information Science and Technology* 68, 2 (2017), 321–332.

[56] Daniel M Romero, Dan Huttenlocher, and Jon Kleinberg. 2015. Coordination and Efficiency in Decentralized Collaboration. In *Ninth International AAAI Conference on Web and Social Media*.

[57] Daniel M Romero, Brian Uzzi, and Jon Kleinberg. 2016. Social networks under stress. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 9–20.
[58] Barry M Staw, Lance E Sandelands, and Jane E Dutton. 1981. Threat rigidity effects in organizational behavior: A multilevel analysis. *Administrative Science Quarterly* (1981), 501–524.
[59] Michail Tsikerdekis. 2017. Cumulative Experience and Recent Behavior and their Relation to Content Quality on Wikipedia. *Interacting with Computers* 29, 5 (2017), 737–754.
[60] Milena Tsvetkova, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri. 2017. Even good bots fight: The case of Wikipedia. *PloS one* 12, 2 (2017), e0171774.
[61] Milena Tsvetkova, Ruth García-Gavilanes, and Taha Yasseri. 2016. Dynamics of Disagreement: Large-Scale Temporal Network Analysis Reveals Negative Interactions in Online Collaboration. *Scientific reports* 6 (2016).
[62] Bruce W Tuckman and Mary Ann C Jensen. 1977. Stages of small-group development revisited. *Group & Organization Studies* 2, 4 (1977), 419–427.
[63] Fernanda B Viegas, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. 2007. Talk before You type: Coordination in Wikipedia. In *Proceedings of the 40th Hawaii International Conference on System Sciences*. IEEE, 78–78.
[64] Fernanda B Viégas, Martin Wattenberg, and Matthew M McKeon. 2007. The hidden order of Wikipedia. In *International conference on Online communities and social computing*. Springer, 445–454.
[65] Morten Warncke-Wang, Vladislav R Ayukaev, Brent Hecht, and Loren G Terveen. 2015. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 743–756.
[66] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell me more: an actionable quality model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*. ACM, 8.
[67] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 347–354.
[68] Jaime B Windeler, Likoebe M Maruping, Lionel P Robert, and Cynthia K Riemenschneider. 2015. E-profiles, conflict, and shared understanding in distributed teams. *Journal of the Association for Information Systems* 16, 7 (2015), 608.
[69] Thomas Wöhner and Ralf Peters. 2009. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM, 16.
[70] Jeffrey M Wooldridge. 2015. *Introductory econometrics: A modern approach*. Nelson Education.
[71] Teng Ye, Katharina Reinecke, and Lionel P Robert Jr. 2017. Personalized Feedback Versus Money: The Effect on Reliability of Subjective Data in Online Experimental Platforms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 343–346.
[72] Teng Ye, Sangseok You, and Lionel Robert Jr. 2017. When Does More Money Work? Examining the Role of Perceived Fairness in Pay on the Performance Quality of Crowdworkers. In *Eleventh International AAAI Conference on Web and Social Media*. AAAI. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15601
[73] Ark Fangzhou Zhang, Livneh Danielle, Ceren Budak, Lionel P. Robert Jr., and Daniel M. Romero. 2017. Shocking the Crowd: The Effect of Censorship Shocks on Chinese Wikipedia. In *Eleventh International AAAI Conference on Web and Social Media*.
[74] Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 935–944.