Characterizing Opinion Dynamics and Group Decision Making in Wikipedia Content Discussions

Khandaker Tasnim Huq
University of South Florida
Tampa, FL, USA
kthuq@usf.edu

Giovanni Luca Ciampaglia
University of South Florida
Tampa, FL, USA
glc3@mail.usf.edu

ABSTRACT

Wikipedia, the online encyclopedia, is a trusted source of knowledge for millions of individuals worldwide. As everyone can start a new article, it is often necessary to decide whether certain entries meet the standards for inclusion set forth by the community. These decisions (which are known as “Article for Deletion”, or AfD) are taken by groups of editors in a deliberative fashion, and are known for displaying a number of common biases associated to group decision making. Here, we present an analysis of 1,967,768 AfD discussions between 2005 and 2018. We perform a signed network analysis to capture the dynamics of agreement and disagreement among editors. We measure the preference of each editor for voting toward either inclusion or deletion. We further describe the evolution of individual editors and their voting preferences over time, finding four major opinion groups. Finally, we develop a predictive model of discussion outcomes based on latent factors. Our results shed light on an important, yet overlooked, aspect of curation dynamics in peer production communities, and could inform the design of improved processes of collective deliberation on the web.

KEYWORDS

Wikipedia, Article for Deletion, Computational Social Science, Group decision-making, Discussion outcome prediction, Opinion dynamics

ACM Reference Format:
Khandaker Tasnim Huq and Giovanni Luca Ciampaglia. 2021. Characterizing Opinion Dynamics and Group Decision Making in Wikipedia Content Discussions. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3442442.3452354

1 INTRODUCTION

In the English Wikipedia, the Article For Deletion (AfD) process refers to the set of collective deliberations that wikipedians engage in when trying to decide whether problematic entries should be deleted from the encyclopedia. In particular, AfDs are one of multiple processes in place for deleting content from the English Wikipedia, happening only when the decision to delete some content will likely lead to some discussion. Other parts of the broader deletion process handle instead the more obvious cases, and simply require the posting of a template notice (like speedy deletion, or CSD, or proposed for deletion, or PROD). AfD deliberations, in contrast, take the form of semi-structured threaded discussions, in which the nominated entry is judged on whether it meets the notability standards set forth by the community, as well as any relevant editorial guidelines and policies. AfDs are therefore an instance of group decision-making, and as such have been the subject of much interest from the literature about peer production, like much of Wikipedia itself [19].

A typical AfD discussion involves a variable-size group of volunteer reviewers (typically registered Wikipedia editors with a minimum number of contributions), who are called to respond to an initial deletion proposal about a given entry. The nomination consists of a rationale for the proposed deletion — typically an explanation of why the entry violates or fails to meet some community standards. Users respond to the the nomination by providing their own recommendation, along with a justification. Typical recommendations include deleting the entry (i.e., a ‘Delete’ decision), integrating its contents into another entry (‘Merge’), deleting its contents but keeping a pointer that redirects readers to another entry (‘Redirect’), or simply keeping the page as it is (‘Keep’). Decisions are based on group consensus. If the reviewers fail to reach a consensus, no action is taken (i.e., the entry is kept). A page can be nominated multiple times, provided that the rationale of each nomination is different from the previous.

The AfD process has received some attention in the literature on online collaboration communities, in some instances to point out to its complexities [6, 7, 20]. The first detailed work on collective deliberation processes in AfD was presented by Taraborelli and Ciampaglia [24], who found evidence of herding among the participants, i.e. the phenomenon by which latecomers are influenced by the votes cast early on in the discussion. They also presented evidence that AfD participants cluster in two major groups, colloquially referred to as the ‘inclusionists’ (i.e., those with a strong preference for nominated entries to be kept, and more generally that Wikipedia, by virtue of being a digital encyclopedia, should include as much content as possible) and the ‘deletionists’ (i.e., those with a strong preference for nominated pages to be deleted, and who more generally advocate that, despite being a digital encyclopedia, Wikipedia should enforce clear content standards), which suggest the presence of substantial social bias in the AfD process overall.

Shortly after, Lam et al. observed that AfD decisions are sometimes overturned at a later stage, and used this observation to establish a connection between the composition of the deliberating group and decision quality [11]. They also observed that the

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452354

administrators who close the discussions do have an effect on the final outcome.

More recently, Mayfield and Black have proposed a number of predictive models, based on natural language processing, for detecting the stance of a user in an AfD, as well as the overall outcome of an AfD discussion [15, 16]. Maniu et al. [14] proposed a method to infer a signed network (or "web of trust") directly from user interactions in Wikipedia, which connects to sociological theories such as the theory of status and structural balance [13]. Sepehri Rad et al. [21] also used signed networks to infer the attitude of editors towards each other in the context of edit history and admin election. Finally, recently Lerner and Lomi [12] have shown that the structure of collaboration networks might have a significant impact on the quality of articles in Wikipedia.

A common theme to all prior work on AfDs is that the outcomes of AfD deliberations are affected by situational factors, like herding [24], or social status [13]. However, there could be also more long-term processes at play, such as adaptation and social learning. These factors could influence deliberations that take place over multiple discussions and the tenure of multiple individual contributors, and could have strong implication on the overall quality of the decisions taken. We are inspired in particular by recent work on Wikipedia editors [22] showing that diversity of opinion is associated with increased content quality in the dynamic of content production. Therefore, in this research we aim to answer the following questions:

Q1. Are there any biasing factors that can explain the voting patterns of editors in a given discussion (i.e., including the degree of agreement / disagreement with her peers)?
Q2. How do preferences on content inclusion / deletion form among AfD reviewers, and what is their evolution over the tenure of an AfD reviewer?
Q3. Can the estimation of the votes of the editors be regarded as predictive factor of AfD final decision outcomes?

To address these questions, we perform a signed network analysis to capture the dynamics of the discussions in the group decision-making process over time, in a manner reminiscent of prior work [10, 23]. In our case, we quantify the preference of each reviewer toward any particular decision, by measuring their votes toward inclusion or deletion — the two most frequent recommendations. We analyze the structure of this network, discovering a strong core-periphery structure corresponding to different cohorts of reviewers, and quantify the level of agreement and disagreement across the structure of the network.

Our analysis reveals that there are strongly polarized groups in the AfD community, and that the evolution of group structure in different cohorts of editors reflects different historical periods of the broader Wikipedia project.

We then study the evolution of individual reviewers and of their voting preferences, discovering, despite the relative stability, substantial longitudinal variation across groups especially in early periods, which could suggest the presence of social learning phenomena among reviewers. In particular, we find that one group (strong deletionists) is much less susceptible to change than others. This is reminiscent of the phenomenon of the 'committed minority' from the study of opinion dynamics [4].

Based on this observation, we apply a latent factor model to characterize the evolution of individual-level preferences in the AfD community. We compare this model in two tasks: stance detection (i.e. given a discussion, predict the vote of an individual editor) and outcome prediction (i.e. given a discussion, predict its final outcome). We evaluate our approach on a recent dataset spanning 10+ years worth of AfD discussions. To preview our findings, we compare our model predictions against the state of the art, which was obtained by Mayfield and Black using language models such as Bert to learn distributed representations of the textual recommendations written by editors during the conversation [15]. Our model, which makes use of only the information about *who* votes on *which* conversation, obtains an AUC of 82%, which is comparable to the state of the art based on NLP.

To facilitate replication of our finding, all our data and code are available on Github at the following address: https://github.com/CSDL-USF/wiki-workshop-2021-huq-ciampaglia.

The rest of the paper is organized as follows. Section 2 describes the main methods of the paper, including the data collection, the measurements of preference or bias of reviewers, the procedure for building the signed network to analyze the structure in AfD decisions, the clustering technique used to mine the patterns of evolution of reviewer preference over time, the prediction model, and the performance of the predictive model. Section 3 discusses the results. Section 4 presents discussions and future work.

2 DATA AND METHODS

In this section we describe the main sources of our data and the methods used to analyze it. In describing the data, we provide some exploratory results, like the cluster analysis of voting preferences, that motivate other parts of our methodology, like the signed network analysis. Therefore, for ease of exposition, we include them here. Later, in Section 3, we build upon these exploratory findings to report the main results of the paper.

2.1 Data Collection

For our empirical analysis we leverage the corpus and metadata recently published by Mayfield and Black [16]. This data set contains records for all the AfD discussions in the English Wikipedia that took place between January 2005 and December 2018. From this data set, we select the information on timestamps, outcomes, nominations, votes (i.e., the recommendations), users, and policy citations. After parsing the data, we obtain a total of $1,967,768$ recommendations. To reduce noise in our estimates, we remove from the dataset all editors with occasional contributions. In prior work, Taraborelli and Ciampaglia recommended, for statistical reasons, to retain editors using a minimum threshold of the 5 AfD recommendations [24]. After filtering out the users who cast less than 5 recommendations, we end up with $1,495,963$ recommendations made by $20,153$ reviewers over $355,505$ discussions. In the left panel of Figure 1 we plot the frequency of the four main vote options: Delete, Keep, Merge, and Redirect. We see that the vast majority of recommendations are cast for keeping or deleting the nominated entry. In the right panel of Figure 1 we show instead the frequency distribution of the 5 most typical AfD outcomes. Here, a major portion of discussion outcomes are 'Delete' and 'Keep' and

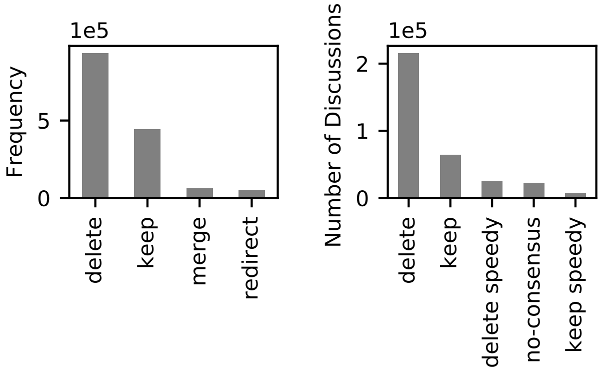

Figure 1: Frequency of the most typical recommendations (left) and of the most typical outcomes (right) in the AfD dataset.

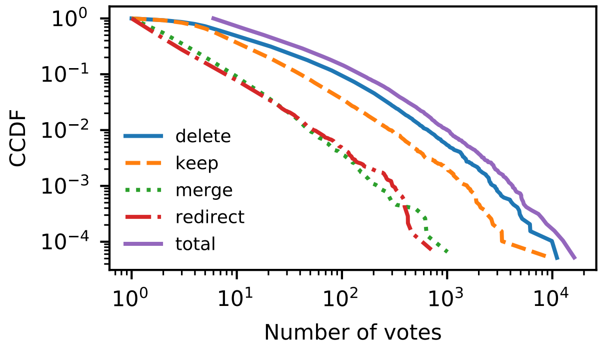

Figure 2: Complementary cumulative distribution of number of recommendations per user in the AfD dataset.

their variants. We note that the proportion of Delete to Keep votes is roughly 68%, while for outcomes it is 77%. This suggests that delete votes are more decisive.

In Figure 2 we show that the relative popularity of these two options is preserved even if we consider the individual level of activity of an editor, as measured by the total number of recommendations by users across all four categories. Hence in the following we focus only on the recommendations cast to recommend either a 'Delete' or 'Keep' action, as they are the two most frequent recommendations among all. Likewise, to simplify the rest of our analysis, we combine 'Delete speedy' and 'Keep speedy' outcomes with 'Delete' and 'Keep', respectively, since these simply represent situations in which the discussion was closed after a very brief deliberation.

2.2 Measurement of Preference

We quantify the preference of each reviewer for recommending either 'Keep' or 'Delete' as the proportion of times she recommended 'Keep' out of the total number of discussions she took place in. This score is a number between 0 (full deletionist) and 1 (full inclusionist). We fit a Gaussian mixture model (GMM) to identify groups of users with similar preferences based on their score. We use the implementation provided by Scikit-learn [3], which uses the Expectation-Maximization algorithm (EM). However, we found that EM was prone to instabilities due singularity, so used an alternative implementation, also from Scikit-learn, which uses the Variational Bayes algorithm, which did not suffer from singularity issues. To choose the number of components, we perform model selection. In the right panel of Figure 3 we plot the Bayesian Information Criterion (BIC) of each model as a function of the number of components k, and observe a minimum at $k = 6$. We compared this fitted model with the one obtained using $k = 4$ components, and even though the clusters are quantitatively different, for ease of exposition, in the following we choose to use this more parsimonious model. In the left panel of Figure 3 we show the mixture distribution for $k = 4$, where we can see that, across the full spectrum, it is possible to identify 4 main groups, roughly corresponding to the following classes of users: a) strong deletionists, b) moderate deletionists, c) moderate inclusionists and d) strong inclusionists. Thus, unlike prior work, which showed evidence for a division in two main factions [24], we find that the data support a division into more that 2 groups.

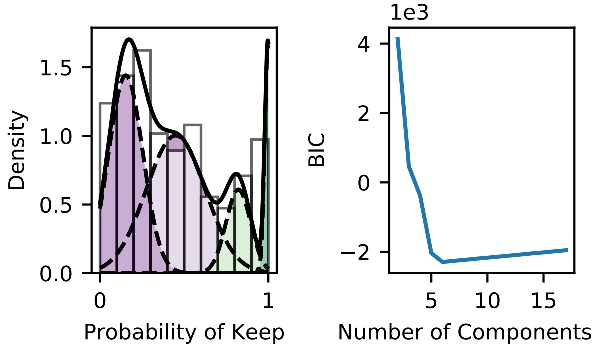

Figure 3: Right: The Bayesian Information Criterion for models of different complexity. Left: The Gaussian mixture model with $k = 4$ components. For ease of exposition we use this more parsimonious model instead of the one with $k = 6$ components.

2.3 Signed Network Analysis

Figure 4 gives a schematic description of our network analysis methodology. We defined a signed network as follows: if two users u, v make the same recommendation in the same discussion p then there is a positive edge (with weight $\omega(u, v, p) = +1$) between them, otherwise a negative one ($\omega(u, v, p) = -1$). Note that this definition yields a graph with parallel edges. We obtain a signed network $G = (V, E)$ by collapsing all parallel edges and computing the sign of the total weight between two users:

$$w(u,v) = \text{sign}\left(\sum_p \omega(u,v,p)\right) \quad u,v \in V \tag{1}$$

Where $\text{sign}(x) = 1$ if $x \geq 0$, else 0.

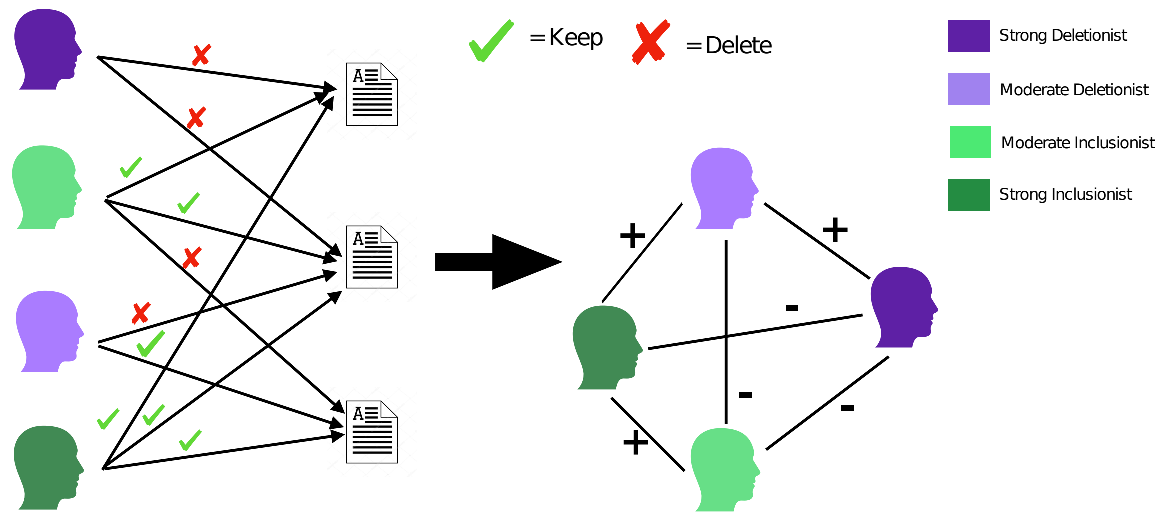

Figure 4: Signed network of AfD editors. We begin from a bipartite network between editors and discussions. In this network, each edge is a vote to either keep or delete an article. We then consider the signed projection of this network, where there is an edge among any two editors if they co-voted on the same discussion(s), and the sign indicates whether the relative majority of co-votes (across all common discussions) agreed on the preference (+) or not (−). Colors represent individual voting preferences and correspond to cluster labels from the GMM model (see Fig. 3).

2.4 Preference Evolution

To characterize how individual preferences evolves over time (Q2), we consider how the likelihood of recommending 'Keep' changes over the tenure of an AfD reviewer. Here the tenure of a user is given by the sequence of all the discussions they took part in, in chronological order. However, since some users participate more often to AfDs than others, and over different periods of time, we split each sequence into 10 equally-sized groups, and compute a score for each group, obtaining a voting score trajectory as a function of tenure deciles, instead of as a function of time. We then cluster these 10-dimensional points with a Gaussian Mixture Model.

2.5 Outcome Prediction Task

We apply the latent factor model [1] that can measure the preference of the editors and can 'recommend' the more likely vote to take in a given AfD. This model can identify the hidden factors of both users and items that can impact on the user preferences. A positive rating (+1) can be interpreted as a vote to 'Keep' and a negative rating (−1) as a 'Delete'. Given the past recommendations cast by editors, we estimate the rating of each article and the associated latent factors. This latent factor model predicts the rating $\hat{r}_{ui}$ that the u-th editor would give to the i-th article using the following formula:

$$\hat{r}_{ui} = \alpha + b_u + b_i + q_i^T p_u \tag{2}$$

Here, α is the mean of all ratings, b_u and b_i are the biases of editor u and article i, and q_i and p_u are the latent factors of the editor and article. Note that our data consist of user–item votes, which can be arranged in a matrix. In practice, we split the ratings in training and testing. The right-hand side of Eq. 2 is computed using only training data, which are needed to perform the matrix factorization. These yield the latent factors q_i and p_u, and allow to compute the user/item biases b_u and b_i. Therefore, the estimated rating $\hat{r}_{ui}$ is computed only using information from the training data.

Table 1: Features used in the Outcome prediction task.

Feature	Range
Mean predicted rating	$[-1, +1]$
Variance of predicted rating	$\mathbb{R}^+$
# of predicted positive preferences (binarized)	$\mathbb{Z}^+$
# of predicted negative preferences (binarized)	$\mathbb{Z}^+$

We train a Singular Value Decomposition model with 400 hidden factors implemented using Surprise [9], which is a python package for modeling recommender systems. From this model, we obtain the predicted ratings given by editors to articles in a range of −1 to +1. We compute the average of both observed and predicted rating of the articles. To predict outcomes, we label discussions using three possible labels: 'Keep' and 'Speedy Keep' (labeled as +1), 'Delete' and 'Speedy Delete' (−1), and 'No consensus' (0).

To predict discussion outcomes we adopt a model stacking approach, see Figure 5. We build a stacked model in which we first compute the predicted rating of articles using the latent factor model, then we use the estimated rating to build features for the AfD outcome prediction. Our model comprises two stages: at the first stage, we use the latent factor model to predict the rating of each discussion participant. At the second stage, we use Logistic regression to predict the overall outcome of the discussion. We use the implementation provided in Scikit-Learn [3], with L2 regularization and the LIBLINEAR solver [5]. Classification features are listed in Table 1, where preferences are obtained by binarizing the predicted rating using a threshold.

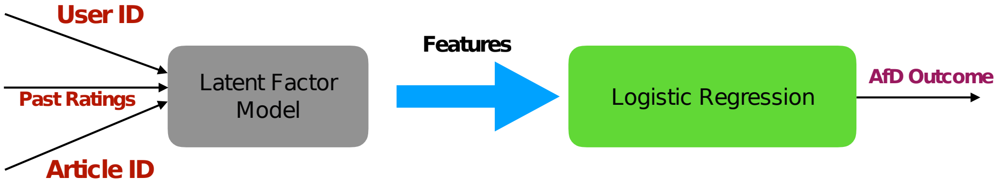

Figure 5: Architecture of the stacked model used for predicting AfD discussion outcomes. The output of the latent factor model (article and user bias terms) are fed as features to a logistic regression classifier for predicting the outcome of an AfD discussion.

To evaluate the model we use two different levels of cross-validation, one for the rating prediction and one for the outcome. Both layers use a 5-fold split. In the first layer, we use the latent factor model from Eq. 2 to predict individual user ratings. In the second layer, we combine rating predictions together to predict the outcome. To avoid leakage, we arrange the layers in such a way that all the instances from a particular discussion stay in only one fold and there is never crossover from the same discussion among training and testing datasets across different layers.

As an upper bound on model accuracy, we train another stacked model, but we use the observed rating, as opposed to the predicted ones, to build the feature set for the outcome prediction step. This is motivated by the observation that English Wikipedia administrators routinely follow group consensus, and follows a similar choice by Mayfield and Black [15].

3 RESULTS

Having uncovered a strong division among AfD users based on their voting preferences (Figure 3), we ask to what extent the structure of the signed network G also reflects these divisions. Is the network "polarized" into homogeneous groups of like-minded individuals? One possibility could be that the network is shaped by homophily: people who vote in the same discussions tend to share the same preferences, which would explain the presence of the separate groups shown in Figure 3. However, the decision to express a preference in a particular discussion may depend on different factors, not on the preferences of the other participants.

To answer this question, we compute the correlation between edge signs and co-membership to the same preference cluster, as inferred by the GMM. To make sure our analysis is not affected by our choice of reducing weights into binary signs, we also compute the same correlation but using the original edge weights instead (i.e. the summation term in Eq. 1). Finally, to better understand the structure of the network, we first decompose the network into its k-cores, and repeat the analysis for each value of k. The k-core of a network G is the maximal subgraph whose nodes all have degree $\geq k$, thus forming a highly cohesive sub-group.

In Figure 6 we plot, as a function of the core number k, the Spearman correlation coefficient between the sign of an edge and the co-membership information. As a baseline, we also plot the same correlation, but instead of using the GMM co-membership information, we consider co-membership in network communities identified by a community detection algorithm (we used the Louvain method [2] to identify communities).

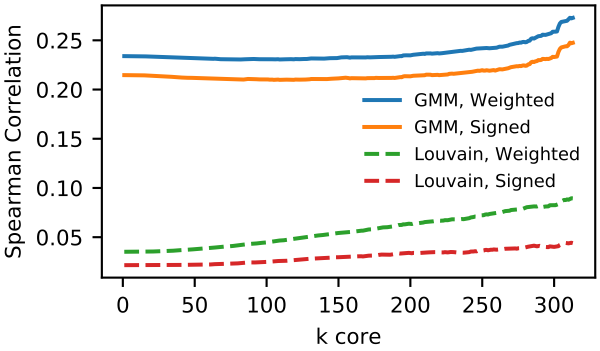

Figure 6: Correlation between edge sign / weight and class co-membership for Gaussian mixture model (GMM) and network communities (Louvain).

We observe that the correlation is higher for co-membership in GMM clusters than Louvain communities, suggesting that individual voting preference is more important in explaining patterns of agreement among editors than the choice of the particular discussion to take part in. This observation does not depend on whether we choose to use edge weights as opposed to signs. Furthermore, the correlation increases with k, suggesting that the core of the network is a dense sub-network formed by sub-groups of like-minded editors who tend to express the same preference when co-voting. These sub-groups are still connected to each other, but by edges of negative sign, suggesting that they correspond to different factions voting in the same AfD discussions.

Focusing more in detail on the origin of the core of the network, in Figure 7 (left panel) we plot the distribution of k-shell numbers for editors who participated in their first AfD discussion in the same two-year period. The k-shell of a network is the set of nodes of G that belong to its k-core but not to its $(k + 1)$-core. Therefore, it provides a partition of the node set V into mutually-exclusive groups that get closer and closer to the core of the network as k increases.

The left panel of Figure 7 shows that more recent cohorts of editors tend to belong to shells with a smaller k, and that there is a sudden drop in k around year 2007. Editors who joined before 2007 tend to overwhelmingly belong to the more central parts of the network. These earlier cohorts are not only formed by more

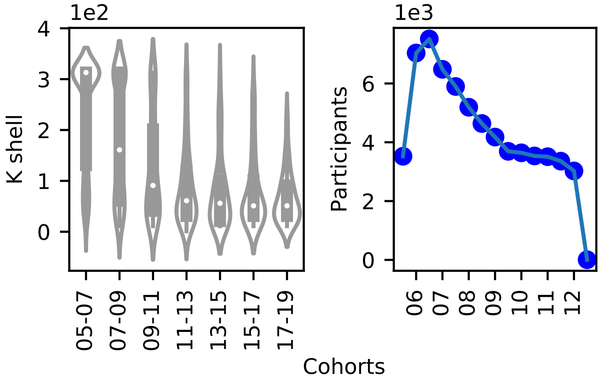

Figure 7: Left: distribution of k-shell values for bi-yearly cohorts of editors who started taking part to AfD discussions in the same period. Right: cohort size over time. Each dot corresponds to the number of editors who started taking part to AfD discussions in that year.

experienced and more active editors, but they are also the largest, as shown in the right panel of Figure 7, where we plot the cohort size as a function of time.

We now turn to the question of whether individual voting preferences change over time. In Figure 8 we plot the evolution of voting preference for editors in the most central shell ($k = 313$). These are the most central editors in the signed network, for which we expect a stronger signal. We fit a GMM on the preference trajectory deciles using the EM algorithm. We perform model selection and find that a model with $k = 4$ components fits the data better than other alternatives. We also fit another GMM (again using the EM algorithm) on the same data but without splitting by deciles. Comparing these clusters with those obtained from the trajectories, we find qualitatively similar groups, suggesting that user preferences are relatively stable over time for these more central editors. However, despite the overall stability of trajectories, we also observe a substantial narrowing of opinions in the early period of an AfD reviewer tenure. This could be evidence of social learning due to imitation. Strong deletionists exhibit the least amount of change, suggesting the possibility of lower susceptibility, or higher resistance, to opinion change in this group, which is reminiscent of models of opinion dynamics [4].

Finally, we turn to the task of predicting discussion outcomes. To motivate our choice of the latent factor model, we start by fitting the latent factor model to the data, to compute the predicted rating of each participant in all AfD discussions in the dataset. For each discussion we compute the average predicted rating, where the average is taken on all participants in the discussion. We then bin discussions based on this average and compute, for each bin, the average discussion outcome of each bin (see Methods). Finally, we repeat the same procedure but instead of using the predicted rating, we group by the average observed rating, i.e. directly from the data. Figure 9 shows that, in both cases, we find a positive correlation between ratings and outcomes, suggesting that predicted ratings of individual users are informative features for predicting the outcome of a discussion.

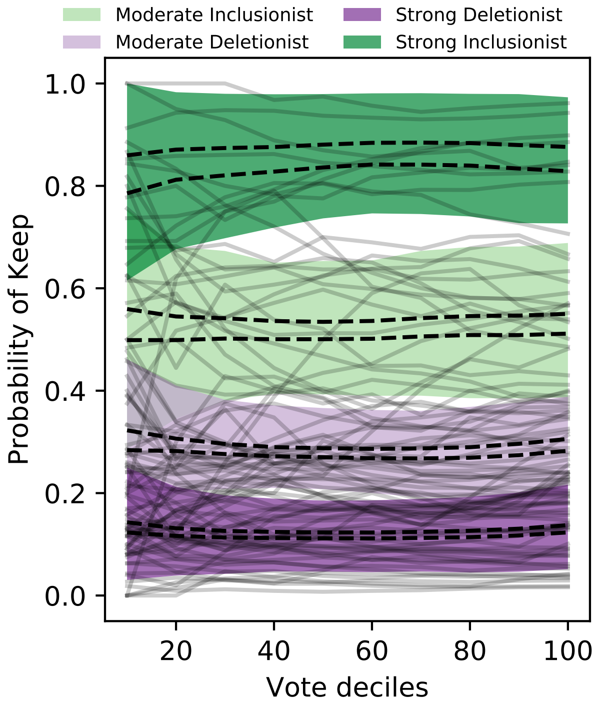

Figure 8: Evolution of voting preferences over time. The four shaded areas correspond to the 4 components of a GMM fitted on individual voting trajectory deciles. The gray solid lines show the trajectories of a sample of randomly-selected individuals. The black dashed lines correspond to the 95% confidence intervals of the average trajectory of each cluster.

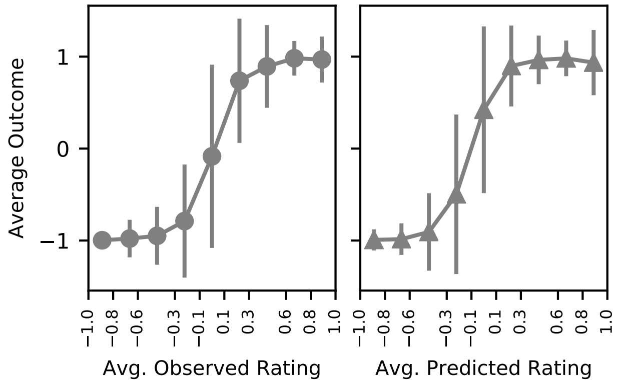

Figure 9: Average discussion outcome as a function of the observed (left) and predicted (right) rating. Error bars represent ±1 standard deviations from the mean.

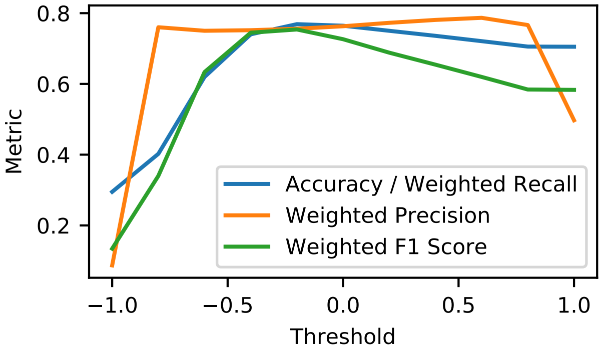

Figure 10: Performance of the latent factor model for rating prediction, as a function of the binarization threshold. Precision and recall are weighted by the number of true instances for each label (positive and negative).

Figure 9 suggests that above a certain threshold of the predicted rating, discussions tend to end in keep, while below in delete. In fact, from the error bars in the right panel of Figure 9, one can notice that there is substantial variability in the outcome even for small negative values of the predicted rating. When building features for the second layer of our stacked model, we considered several thresholds for binarizing the predicted rating (see Figure 10), and found that values of $\theta = -0.2$ and $\theta = -0.4$ worked best.

Figure 11 shows the ROC curve using either the predicted and, as an upper bound, the observed ratings. The area under the curve for the model using a threshold of $\theta = -0.4$ is 0.82, which is slightly better than that of the model with $\theta = -0.2$. We also compute weighted precision, recall, and F1 score of the model, shown in the figure as dashed lines.

4 DISCUSSION AND FUTURE WORK

Our study provides an updated characterization of the potential biases in AfD group discussions. In particular, we find evidence for the existence of a larger number of groups than previously reported in the literature [24]. This suggests that the differences between inclusionists and deletionists are more nuanced than previously thought. In response to our first research question (Q1), to better characterize the extent of these divisions, we built signed networks of AfD editors, in which the sign of the edges capture the level of agreement (or disagreement) among users. A k-core decomposition shows that the structure of the network is highly influenced by a group of highly active, and more experienced users. The observation of a strong drop in size after 2007 suggests that the users from the core of the network joined the English Wikipedia in earlier phases of the project. This is compatible with the fact that participation in Wikipedia peaked around 2007, and steadily decreased after [8]. What is interesting, however, is that these editors remained committed to participating in content curation efforts.

Our study also characterizes the evolution of individual editors preferences over time (Q2). Editors involved in AfD discussions adapt to a particular voting tendency early during their tenure in the AfD process. This is reminiscent of results from prior work, that found that highly active contributors are active from a very early stage [18]. In the context of AfD discussions, this finding could potentially suggest the presence of social learning mechanisms, for example due to imitation. Also, strong deletionists seem more resistant to changing their opinions compared to other groups. More generally, an interesting open question is to determine which stable user characteristics in peer production systems are due to learning phenomena or to the presence of inherent individual traits.

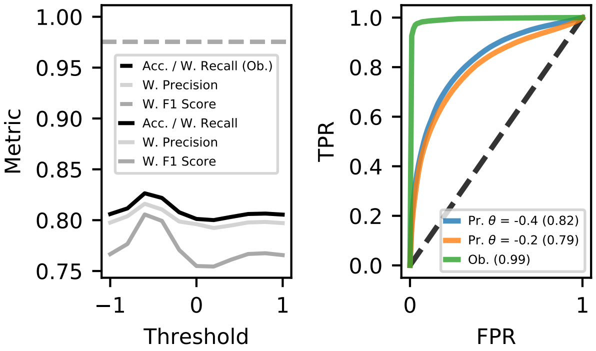

Figure 11: Left: Performance of the outcome prediction model as a function of the binarization threshold. Precision and recall are weighted by the number of true instances for each label (positive and negative). The dashed lines (all overlapping) represent the metric of the observed ratings. Right: ROC curve of the stacked model for the outcome prediction task for different rating features (Pr. = predicted, Ob. = observed). The number in parenthesis is the AUC.

Finally, we observe that information of the individual 'ratings' predicted from a latent factor model can help us predict the outcome of the overall conversation. The estimated ratings carry information about the composition of the group taking part in the discussion, such as the bias or preference of individual editors. Note that our goal with this prediction task is to estimate the overall outcome of a discussion without observing the individual comments and ratings. Even though the majority opinion is generally accepted, the model can be helpful in those instances where this is not the case, or even to suggest when consensus is likely to be reached. It is interesting to note that even though our model is relatively simple and makes use of a limited number of features, its performance is in line with that of the state-of-the-art based on NLP [15], which rely on a much larger amount of information that can be gleaned from text. This work deals with the same task we tackled here (outcome prediction). However, unlike our model, which learns latent factors and bias terms from the vote information, it uses language models such as Bert to learn representations of the full textual comments (i.e., not just the vote information) left by the AfD reviewers.

In future work we would like to explore more the evolution of reviewer expertise. In many recommender systems, users with the same experience level usually have similar preferences, and different experience levels indicate different rating patterns. This observation could be used in the future to estimate the level of editor experience from traces of user activity, in a manner reminiscent of prior work [17]. For example, using a latent factor model with experience levels, we could infer the experience level of AfD users. This metric could shed more light about how preference of reviewers evolves over time.

Another aspect we would like to study using AfD discussions is the quality of content production as a function of curation decisions. Previous work has found a connection between group composition and quality [11]. Our hypothesis is that there is also a correlation between the probability that an article is kept and its quality. As our analysis shows, there are important historical biases affecting the AfD process, and so it is reasonable to assume that other factors, such as the amount of ideological polarization among deletionists and inclusionists may have changed over time. Ideological polarization in the AfD community could potentially influence the quality of inclusion decisions, which has potential impact on the quality of the articles in the English Wikipedia. An assessment of the quality of an editorial decision could be made based on a number of criteria, like the quality and notability of the articles, the reliability and fact-check worthiness of the information, the degree to which relevant policies are met or violated, the level of systematic, gender, racial, and political bias in the discussions etc. Such an assessment could help shed more light on Q3, about how to quantify the effect of polarization in the AfD decision-making process.

Finally, future work could also extend our network analysis. While building the signed network, the first step was to build a bipartite network between editors and AfD discussions / articles. This approach has some limitations. Even though two editors have not co-voted in the same AfDs, it would be helpful if there was a way to encode the level of agreement (or disagreement) between them. To resolve this issue, we would like to build a bipartite network between editors and the categories of the articles nominated for deletion. The idea is that some AfD categories might be more or less biased toward certain outcomes. This in turn could help to infer the agreement sign among editors, even in those cases in which two editors never took part in the same discussions.

REFERENCES

[1] Robert M. Bell and Yehuda Koren. 2007. Lessons from the Netflix Prize Challenge. *SIGKDD Explor. Newsl.* 9, 2 (Dec. 2007), 75–79. https://doi.org/10.1145/1345448.1345465

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[3] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. Springer, Prague, Czech Republic, 108–122.

[4] Guillaume Deffuant, Frédéric Amblard, Gérard Weisbuch, and Thierry Faure. 2002. How can extremism prevail? A study based on the relative agreement interaction model. *Journal of artificial societies and social simulation* 5, 4 (2002), n.a. pages.

[5] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* 9 (June 2008), 1871–1874.

[6] Heather Ford and R. Stuart Geiger. 2012. "Writing up Rather than Writing down": Becoming Wikipedia Literate. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (Linz, Austria) *(WikiSym '12)*. Association for Computing Machinery, New York, NY, USA, Article 16, 4 pages. https://doi.org/10.1145/2462932.2462954

[7] R. Stuart Geiger and Heather Ford. 2011. Participation in Wikipedia's Article Deletion Processes. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (Mountain View, California) *(WikiSym '11)*. Association for Computing Machinery, New York, NY, USA, 201–202. https://doi.org/10.1145/2038558.2038593

[8] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57, 5 (2013), 664–688. https://doi.org/10.1177/0002764212469365

[9] Nicolas Hug. 2020. Surprise: A Python library for recommender systems. *Journal of Open Source Software* 5, 52 (2020), 2174. https://doi.org/10.21105/joss.02174

[10] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 453—-462. https://doi.org/10.1145/1240624.1240698

[11] Shyong K. Lam, Jawed Karim, and John Riedl. 2010. The Effects of Group Composition on Decision Quality in a Social Production Community. In *Proceedings of the 16th ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '10)*. Association for Computing Machinery, New York, NY, USA, 55—-64. https://doi.org/10.1145/1880071.1880083

[12] Jürgen Lerner and Alessandro Lomi. 2019. The Network Structure of Successful Collaboration in Wikipedia. In *Proceedings of the 52nd Annual Hawaii International Conference on System Sciences*. University of Hawai'i at Manoa, Honolulu, HI, 2622–2631.

[13] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed Networks in Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1361–1370. https://doi.org/10.1145/1753326.1753532

[14] Silviu Maniu, Bogdan Cautis, and Talel Abdessalem. 2011. Building a Signed Network from Interactions in Wikipedia. In *Databases and Social Networks* (Athens, Greece) *(DBSocial '11)*. Association for Computing Machinery, New York, NY, USA, 19–24. https://doi.org/10.1145/1996413.1996417

[15] Elijah Mayfield and Alan Black. 2019. Stance Classification, Outcome Prediction, and Impact Assessment: NLP Tasks for Studying Group Decision-Making. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics, Minneapolis, Minnesota, 65–77. https://doi.org/10.18653/v1/W19-2108

[16] Elijah Mayfield and Alan W. Black. 2019. Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 206 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359308

[17] Julian John McAuley and Jure Leskovec. 2013. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 897–908. https://doi.org/10.1145/2488388.2488466

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

[19] Joseph Michael Reagle. 2010. *Good Faith Collaboration: The Culture of Wikipedia*. The MIT Press, Cambridge, MA, USA.

[20] Jodi Schneider, Alexandre Passant, and Stefan Decker. 2012. Deletion Discussions in Wikipedia: Decision Factors and Outcomes. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (Linz, Austria) *(WikiSym '12)*. Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. https://doi.org/10.1145/2462932.2462955

[21] Hoda Sepehri Rad, Aibek Makazhanov, Davood Rafiei, and Denilson Barbosa. 2012. Leveraging Editor Collaboration Patterns in Wikipedia. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media* (Milwaukee, Wisconsin, USA) *(HT '12)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2309996.2310001

[22] Feng Shi, Misha Teplitskiy, Eamon Duede, and James A Evans. 2019. The wisdom of polarized crowds. *Nature human behaviour* 3, 4 (2019), 329–336.

[23] B. Suh, E. H. Chi, B. A. Pendleton, and A. Kittur. 2007. Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations. In *2007 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, Piscataway, NJ, USA, 163–170. https://doi.org/10.1109/VAST.2007.4389010

[24] D. Taraborelli and G.L. Ciampaglia. 2010. Beyond Notability. Collective Deliberation on Content Inclusion in Wikipedia. In *Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop (SASOW)*. IEEE, Piscataway, NJ, USA, 122–125. https://doi.org/10.1109/SASOW.2010.26