

Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments

JANE IM, University of Michigan School of Information*, USA

AMY X. ZHANG, MIT CSAIL, USA

CHRISTOPHER J. SCHILLING, Wikimedia Foundation, USA

DAVID KARGER, MIT CSAIL, USA

Resolving disputes in a timely manner is crucial for any online production group. We present an analysis of Requests for Comments (RfCs), one of the main vehicles on Wikipedia for formally resolving a policy or content dispute. We collected an exhaustive dataset of 7,316 RfCs on English Wikipedia over the course of 7 years and conducted a qualitative and quantitative analysis into what issues affect the RFC process. Our analysis was informed by 10 interviews with frequent RfC closers. We found that a major issue affecting the RfC process is the prevalence of RfCs that could have benefited from formal closure but that linger indefinitely without one, with factors including participants' interest and expertise impacting the likelihood of resolution. From these findings, we developed a model that predicts whether an RfC will go stale with 75.3% accuracy, a level that is approached as early as one week after dispute initiation.

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

Keywords: dataset; collaboration; online communities; deliberation; Wikipedia; online discussion; comments; consensus; dispute resolution

ACM Reference Format:

Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. 2018. Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2, CSCW, Article 74 (November 2018). ACM, New York, NY. 24 pages. <https://doi.org/10.1145/3274343>

1 INTRODUCTION

The internet has enabled large-scale collaboration on tasks of a grand scale, from building the world's largest encyclopedia to solving open mathematics problems [10]. However, given the scale of interaction between diverse participants, it is no surprise that disputes often occur while working together. Thus, resolving disputes in a timely manner is of fundamental importance in any workgroup towards maintaining productivity and a healthy community. Understanding and improving such online processes for deliberation and resolution can have impact in areas including open democratic initiatives and civic participation [30], as well as virtual teams [19], open source

*The work was done during the author's undergraduate years at Korea University.

Authors' addresses: Jane Im, University of Michigan School of Information*, 105 S State Street, Ann Arbor, Michigan, 48104, USA, imjane@umich.edu; Amy X. Zhang, MIT CSAIL, 32 Vassar St. Cambridge, Massachusetts, USA, axz@mit.edu; Christopher J. Schilling, Wikimedia Foundation, 1 Montgomery Street, Suite 1600, San Francisco, California, USA, cschilling@wikimedia.org; David Karger, MIT CSAIL, 32 Vassar St. Cambridge, Massachusetts, USA, karger@mit.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2018/11-ART74 \$15.00

<https://doi.org/10.1145/3274343>

development [27], and online community maintenance [32]. Nowhere is this more clear than on Wikipedia, a place where almost all conflict is resolved through online deliberation. The stakes for deliberation can be high—for instance, the addition of two paragraphs about a city on its Wikipedia page can lead to significant changes in tourism [17]. As a result, conflicts arise on the platform regularly [23, 47], mirroring conflicts around contested information in the world. Prior research has often focused on “edit wars”, or back-and-forth edits on Wikipedia articles, as well as on article talk pages [39], where editors go to informally resolve an issue, as signals of conflict and resolution. However, there are also various formal resolution processes for disputes that cannot be resolved informally, with differing layers of escalation. The study of these formal processes can reveal insights about factors leading to resolution as well as areas of friction, towards the design of better processes and systems for online deliberation and resolution.

To better understand online deliberation, we investigated one of the primary formal processes on English Wikipedia for deliberation and resolution of content and policy disputes—the Request for Comment (RfC) process. Using RfCs, editors who cannot resolve a dispute may publicize their deliberation to the broader Wikipedia community to invite participation, sometimes culminating in a *closing statement* by a neutral editor that summarizes the discussion and makes a resolution.

We created a novel, comprehensive dataset of 7,316 RfCs from English Wikipedia dating from 2011 to 2017, parsed to separate out closing statements, authors, and reply structure. This dataset is released publicly for the research community.¹ We employed a mixed-methods approach by analyzing this data quantitatively as a whole as well as qualitatively by selecting a random subset of 40 RfCs to manually inspect. To inform our analysis, we interviewed 10 of the most frequent RfC closers to understand their motivations and considerations when deciding whether to close an RfC.

From the complementary sources of data, we examined what major factors in the RfC process result in failure to come to a resolution. Not all RfCs require a formal resolution by a closer; instead, some may informally end due to overwhelming agreement by participants or withdrawal of the RfC by the initiator. In our dataset, we found that 57.65% of RfCs end up getting formally closed through the addition of a summary statement resolving the dispute. However, of the 42.35% of RfCs with no formal resolution, we found that 78% had no participant activity to informally end the RfC—in other words, that *a full one third of all RfCs in our dataset were left stale*. A prevalence of stale and unresolved disputes may mean that effort put into discussion is wasted and time is lost waiting for resolution.

From interviews and qualitative analysis of our dataset, we uncovered reasons for why these RfCs do not get formally closed, including factors such as poorly articulated initial statements by inexperienced discussion initiators, lack of interest from third-party experienced Wikipedia editors, and excessive bickering or contentiousness during the discussion.

Using these factors to inform a series of features, we developed a model to predict whether an RfC will go stale based on information about the page before the RfC initiation as well as what transpired over the course of participation in the RfC. When trained and tested on our entire dataset, the best model achieved 75.3% accuracy, an improvement of 8.1% over a baseline of simply predicting that it will not go stale. We find that the most informative features as to whether an RfC will go stale are the interest and expertise level of participants, followed by features related to the size and shape of the discussion. Furthermore, we consider how well such a model performs as an RfC progresses in time after its initiation. At their start point with just an initial statement, prediction of the outcome of RfCs is little better than the baseline of predicting closure for all RfCs. However, even after just one week of participation, we can predict the likelihood of going stale at

¹https://figshare.com/articles/rfc_sql/7038575

above 70% accuracy. Using this model, participants and the initiator of an ongoing RfC can assess the likelihood of an RfC going stale which can inform future actions.

Finally, we revisit the major goals of a deliberative process and how novel tools such as our model and new designs can help make the deliberations and resolutions on Wikipedia more effective. We consider how tools for publicizing RfCs or connecting editors with different levels of expertise could improve consensus-building. We also consider how tools for better organization and sensemaking of discussion can be of use to initiators, participants, and closers within Wikipedia, as well as in other communities conducting deliberation.

2 BACKGROUND AND RELATED WORK

2.1 Self-governance and Rule-Making in Wikipedia

As commons-based communities such as Wikipedia and open-source development grow larger and become more stable, questions of governance become critical [2, 29]. Researchers have examined how policies on Wikipedia are shaped through the creation of proposals that eventually form rules or guidelines. Over time, policies as well as the processes for generating them became more formalized [3, 43] and complex [8], generating hundreds of pages for editors to reference in disputes. Despite this, examination suggested that Wikipedia's governance stayed flexible towards various structures [8] and decentralized when it came to modification and interpretation [15]. More recent analysis of rules on Wikipedia found a shift in favor of deliberation coupled with declining revision activity [20]. Given the impact of deliberation in the continual re-interpretation of policies, it is important that conflicts that affect policy be resolved quickly.

2.2 Processes for Resolving Content Disputes

Broadly, there are two types of disputes in Wikipedia, content-related disputes, which include policy disputes, and user conduct disputes, and numerous formal and informal mechanisms for achieving resolutions for each type. While our focus is on content-related disputes, the line between the two types can blur, as user conduct issues can arise in the course of a deliberation about content. When it comes to resolving a content dispute, editors normally try to resolve it on their own by following Wikipedia policies for achieving consensus² and dispute resolution³ through editing or discussion via the article's talk page.

However, when the dispute cannot be resolved by the involved members, there are a number of ways to receive outside help. First, Third Opinion (3O) is reserved for content-related issues between exactly two editors, and is a relatively informal process for getting an outside opinion. In comparison, the Dispute Resolution Noticeboard (DRN) is used for disputes involving more than two parties or when 3O does not resolve the dispute. Volunteer moderators on the noticeboard provide suggestions and mediation towards the dispute, but this process is primarily limited to simple disputes that can be quickly resolved. If the dispute escalates, there is Formal Mediation, which is provided by a panel of experienced mediators called the Mediation Committee (MedCom) who resolve Requests for Mediation (RfM) once they are filed. At any point in the escalation of dispute resolution processes, editors can turn to Requests for Comments (RfCs) by writing up a proposal or question on the relevant article talk page and then inviting comment by the broader community by posting to various noticeboards.

For this work, we chose to focus on RfCs as it is one of the more common formal processes for resolution due to its flexibility, and because it involves a number of editors across Wikipedia due to the gathering of input from the broader community, as opposed to places like 3O or DRN. Beyond

²<https://en.wikipedia.org/wiki/Wikipedia:Consensus>

³https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution#Resolving_content_disputes

these forums, there are more specialized venues for resolution for particular types of content disputes, such as Articles for Deletion (AfD), which focuses on whether to delete, merge, or move articles. Finally, as mentioned above, there is an entire separate set of noticeboards and processes for the second category of user conduct disputes, culminating in final arbitration by the Arbitration Committee (ArbCom), that we do not study here.

2.3 Research on Conflict and Deliberation in Wikipedia

Some research has analyzed informal processes for managing conflict, such as through edits on article pages or informal discussion on talk pages. Many have studied “edit wars” that break out between editors on certain articles [39, 47], giving rise to policies such as the “Three Revert Rule” [6]. These oftentimes “bursty” edits to content can form an informal process for resolving conflict without explicit communication. Other work has demonstrated the value of both implicit coordination via scaffolding while editing and explicit coordination via communication [21, 45].

Researchers have also analyzed the deliberative discussions that happen on Wikipedia, finding evidence of both constructive behavior and pitfalls [33, 42]. Conversations on talk pages can create long chains of back-and-forth responses in a format much like threaded forums [26]. Analysis of talk page communication found that it scales up to help manage conflict as the number of editors grow [22]. Qualitative analyses of deliberation on Wikipedia found a high level of analytic discussion focused on problem analysis [5], while other work has found examples of debates around information quality [38]. However, researchers have also found lower levels of social aspects of deliberation such as respect and consideration [5], and other researchers found cases of power plays when policies are unclear and advocate for more tools to support the consensus process [24]. These studies demonstrate the importance of deliberative discussions on Wikipedia as well as point to challenges and opportunities for tool-building.

While most existing work focuses on informal coordination and communication, in this work we turn to more formal mechanisms for conflict resolution. There exists some analyses of these formal discussions for the case of Articles for Deletion (AfD) [16], though there the number of participants per discussion is generally small and the emphasis is on voting [41]. Thus, AfD discussions do not represent the best examples of actual deliberation and conflict resolution through consensus. In this work, we present one of the first in-depth analyses of the RfC process, one of the main vehicles for deliberation and formal resolution of content disputes on Wikipedia.

2.4 Promotion Discussions and User Roles in Wikipedia

There are also both formal and informal processes for managing user roles and promotion within Wikipedia. Some of the formal processes involve deliberation, such as the Request for Adminship (RfA) process for selecting administrators on Wikipedia. Research has shown that a model considering factors like strong edit history can predict which users will be voted in as an administrator [7]. There is also research into the emergent, informal roles that form on Wikipedia to handle different tasks, including different editor roles [46], as well as social roles [44], and roles based on discourse acts in talk pages [28]. In this work, we shed light on a particular type of informal editor role that has not been studied in detail, which is that of frequent RfC closer. While there are few restrictions on who can close an RfC, we find that closers tend to be experienced editors. From our interviews, we investigate the motivations of frequent RfC closers to get involved in closing.

2.5 Tools for Deliberation and Consensus

There have been many efforts to improve the interface of talk pages and build tools for consensus. Some have targeted the unstructured nature of talk pages, which can cause difficulty for newcomers, and have developed lightweight tools to add structure [34]. Others have developed models to



Fig. 1. Screenshot of an RfC started by using the RfC template tag `{{rfc}}`.

predict different dialog acts in Wikipedia [14], which could also lend greater structure. Within the MediaWiki platform, interfaces have been developed that make talk pages more like question-answering systems or threaded forums, such as Flow⁴ and LiquidThreads⁵.

Researchers have also sought to support consensus-building on Wikipedia, including tools to summarize behavior and track conflicts as they unfold [24]. Some outside tools could be considered as inspiration for alternative structures for discussion. For instance, Considerit makes use of pro-con lists to get an overview of different perspectives [37]. Similarly, Opinion Space plots users' opinions on a 2-dimensional grid [13]. While traditional polls are sometimes used in RfCs, they are generally used to elicit opinions as opposed to vote on a decision, due to the emphasis on consensus-building over majority rule. Unlike polls, these alternative tools help highlight more nuanced perspectives than is possible through aggregating votes along a single axis. A different kind of tool is Reflect, a system for encouraging active listening by having participants summarize each other's comments [25]. Similarly, Wikum is a tool for summarizing discussions in a bottom-up fashion using a wiki-like mode of collaborative editing [48]. In the discussion of this work, we consider how tools could help improve the problems that we notice with formal deliberations.

2.6 Analyzing the Language of Deliberation

Researchers looking at various communities have studied patterns of discourse in deliberations. Some have built models for politeness, finding that editors on Wikipedia who are polite achieve higher status through elections [12]. Other research analyzing debate communities such as Reddit's ChangeMyView found that persuasiveness aligned with greater interplay between counterarguments and the initiator [40]. Research on language coordination shows that echoes of linguistic style in responses can determine power differentials [11]. We build on this work by analyzing language and releasing a dataset of deliberations on Wikipedia along with their closing statements.

3 INTRODUCTION OF REQUEST FOR COMMENT

In this section we provide a description of Requests for Comment (RfCs). RfCs are a common process use by Wikipedia editors, or volunteers who write Wikipedia articles, for requesting input from uninvolved editors concerning disputes about policy, guideline, or article content. It is a formal way to attract more attention to a problem that is not resolvable with local discussions, and uses a system of centralized noticeboards and bot⁶-delivered invitations to advertise discussions.

3.1 The RfC Process

Initiation: The process for RfCs starts with a content dispute that has already been discussed in a talk page but has not been resolved. At that point, an editor can start a new section within the talk page. Using the RfC template tag `{{rfc}}`, the initiator writes a neutral statement in the form of a proposal or question outlining the issue at hand, optionally selecting one or more topical categories as well, as shown in Figure 1.

Initiator: Any Wikipedia editor can initiate an RfC, as long as they follow the specified procedure.

Dissemination: After the initiator adds the RfC template tag to the page, a Wikipedia bot called Legobot assigns the RfC an ID and posts the RfC on the RfC list page pertaining to that category. Once the RfC tag is removed from an RfC, the category page removes the RfC, keeping only a list of active RfCs on the page. Legobot also notifies a random subset of editors that are watching pages or lists related to the RfC, such as editors who have volunteered via the Feedback Request Service⁷. There are currently 2,360 editors listed as volunteers, though editors also provide a limit on how many notifications to receive a month. Anyone may also post the RfC manually to places such as Village Pump⁸ forums, various noticeboards, talk pages of relevant WikiProjects, and talk pages of related articles or policies, in order to invite more discussion from people not already involved.

Discussion: Once initiated and publicized, the discussion unfolds in a threaded fashion using indenting. Some RfCs also include a section for users to indicate their position in a polling process. The default length of an RfC is 30 days, after which Legobot automatically removes the RfC template tag, and it gets removed from RfC lists. Participants can delay this removal if discussion is still ongoing or they can revive the RfC by re-adding the tag later. The RfC may be closed early if consensus is clear before 30 days, though a general practice is to wait at least a week for input.

Participants: Although anyone can participate in an RfC, the system is targeted towards getting input from uninvolved editors who can provide unbiased opinions to help resolve the dispute.

Closure and Conclusion: After a certain period RfCs can conclude with three type of endings, which are a (i) *formal closure*, an (ii) *informal end*, or (iii) simply be left *stale*. These three endings are organized in Table 1. (i) *Formal closure* is a general process for relatively more contentious debates, requesting an uninvolved third party to close and mark the end of the discussion. Anyone may post the RfC to the Wikipedia Administrators' Noticeboard/Requests for closure⁹, a clearinghouse where frequent closers go to find unclosed RfCs. A closer closes the RfC by adding the templates `{{archivetop}}` and `{{archivebottom}}` along with a closing statement surrounding the RfC as shown in Figure 2.

For the remaining RfCs without these templates, there are two possibilities as to what was the outcome of the RfC. First, the RfC could have been (ii) *informally ended* on purpose by participants, the initiator, or another editor by removing the RfC tag manually. This might happen because the initiator reconsiders and chooses to withdraw the RfC, or an obvious consensus may lead participants to agree to withdraw the RfC. Second, the RfC could have (iii) gone *stale*—that is, while waiting for further participation or a formal close, there is a period of no activity for 30 days, and the RfC never gets closed by an individual. In this case, Legobot would remove the RfC tag after 30 days of inactivity, effectively withdrawing the RfC if no one bothers to open it up again.

⁴<https://en.wikipedia.org/wiki/Wikipedia:Flow>

⁵<https://en.wikipedia.org/wiki/Wikipedia:LiquidThreads>

⁶Bots are computer-controlled user accounts that help maintain pages: <https://en.wikipedia.org/wiki/Wikipedia:Bots>

⁷https://en.wikipedia.org/wiki/Wikipedia:Feedback_request_service

⁸https://en.wikipedia.org/wiki/Wikipedia:Village_pump

⁹https://en.wikipedia.org/wiki/Wikipedia:Administrators%27_noticeboard/Requests_for_closure

RfC on income inequality effects

The following discussion is closed. **Please do not modify it.** Subsequent comments should be made on the appropriate discussion page. No further edits should be made to this discussion.

Consensus was to "omit" the material due to concerns that it is off topic. Morphh (talk) 14:31, 17 January 2014 (UTC)

Instead of edit-warring over this excerpt, we clearly need an RFC.

Inequality in land and income ownership is negatively correlated with subsequent economic growth. A strong demand for redistribution will occur in societies where a large section of the population does not have access to the productive resources of the economy. Rational voters must internalize such issues. (Alesina, Alberto (1994). "Distributive Politics and Economic Growth" (PDF). *Quarterly Journal of Economics*. 109 (2): 465–90. doi:10.2307/2118470. Retrieved 17 October 2013. Unknown parameter |coauthors= ignored (|author= suggested) (help); Unknown parameter |month= ignored (help)) High unemployment rates have a significant negative effect when interacting with increases in inequality. Increasing inequality harms growth in countries with high levels of urbanization. High and persistent unemployment also has a negative effect on subsequent long-run economic growth. Unemployment may seriously harm growth because it is a waste of resources, because it generates redistributive pressures and distortions, because it depreciates existing human capital and deters its accumulation, because it drives people to poverty, because it results in liquidity constraints that limit labor mobility, and because it erodes individual self-esteem and promotes social dislocation, unrest and conflict. Policies to control unemployment and reduce its inequality-associated effects can strengthen long-run growth. (Castells-Quintana, David (2012). "Unemployment and long-run economic growth: The role of income inequality and urbanisation" (PDF). *Investigaciones Regionales*. 12 (24): 153–173. Retrieved 17 October 2013. Unknown parameter |coauthors= ignored (|author= suggested) (help))

Should that be included in the Economic effects/Income inequality section? EllenCT (talk) 02:25, 2 January 2014 (UTC)

Survey

- Support inclusion of the passage as a separate paragraph, to explain why income equality is a positive economic effect. EllenCT (talk) 02:25, 2 January 2014 (UTC)
- Omit the paragraph in its entirety, as it does not even approach the subject of taxation, progressive or otherwise. Obviously off-topic and superfluous. Roccodrifi (talk) 02:29, 2 January 2014 (UTC)

RfC: Images used for Planet Nine

(Notifying previously involved editors: Jehochman, prokaryotes, Serendipodous, Fut.Perf. c, Ephraim33, Nergaal, Neutron, Leitmotiv, Kheider, Wnt, Nowa, Ito, Smkolins, Tom Ruen and Jonathunder.)

I really think we need more input regarding the images used on this article. There have been a few previous discussions^{[1][2][3][4][5]} but no clear consensus was demonstrated. I propose that we pool our collective opinions here and put things to a vote. Since there are 2 questionable images I'll split this into 2 subsections.

Artist's impression in the infobox

The infobox currently shows an artist's impression of Planet Nine (right), which appears to be closely based on an image released by Caltech credited to R. Hurt (IPAC). While artist's impressions may help to grab the reader's attention I do not think that it's becoming of an encyclopedia to reproduce them here and I propose that it be removed, or at the very least removed from the infobox. The only information it conveys are *basic assumptions*, which could easily mislead the reader. *A view of the Earth can be found here* in which the planet's size and distance from the Sun is similarly *not* conveyed. Please bear in mind that even if you don't find the picture misleading, others undoubtedly will.

File:Planet-Nine-in-Outer-Space-artistic-depiction.jpg
An artist's impression of Planet Nine

Propose the complete removal of artist's impression. Sorry. nagualdesign 03:43, 4 February 2016 (UTC)

Plenty of artist's impressions are in infoboxes in this encyclopedia. There is no rational reason for its removal. The only issue I have with it is that it is unclear where the Sun is, and how far away it is. SerendiPodous 08:07, 4 February 2016 (UTC)

Could you cite some examples? Last time I asked this question the only example was at *Gamma-ray burst*. The reason artist's impressions are used there is because detailed images aren't available, but we want to convey the formation and structure of GRBs to the reader. They are based on scientific facts. The image of Planet Nine, on the other hand, *supposes* that this hypothetical planet *may well be* an ice giant, whereas the actual hypothesis only deals with orbits and masses. Yes, you could argue that Planet Nine might be found tomorrow and turn out to be an ice giant, but it could also be disproved or found to be something else entirely. As an encyclopedia I think WP only ought to represent what we *do* know - in this case, that the orbits of several TNOs could be explained by the presence of a ninth planet. nagualdesign 22:28, 4 February 2016 (UTC)

Fig. 2. Comparison of a formally closed RfC (top) and one that is not (bottom). Formally closed RfCs have a purple box surrounding the thread and a grey closing statement box. On the other hand, RfCs that are not formally closed have no such template.

	(i) Formally closed	(ii) Informally ended	(iii) Stale
Ended by whom	Uninvolved editor	Participant, initiator, or uninvolved editor	None
RfC tag is removed by whom	Closer	Participant, initiator, or uninvolved editor	Legobot
Exists closing template	Yes	No	No
Dispute is resolved	Yes	Yes	No
Number of RfCs	4,086 (57.65%)	672 (9.48%)	2,329 (32.86%)

Table 1. Differentiation of the three possible outcomes of RfCs.

For the rest of this work, we use the term **“unclosed”** to describe (iii) where RfCs remained stale, without any kind of closure and **“closed”** to describe both (i) and (ii). We use **“formally closed”** and **“informally ended/closed”** when we want to indicate (i) or (ii) respectively.

Closers: Any editor on Wikipedia can formally or informally close an RfC; however, formal closers tend to be more experienced editors on Wikipedia due to their grasp of Wikipedia policy and greater perceived authority within the community. Also, some RfCs do require closure by an administrator if the close involves action that can only be done by an administrator, such as deleting an article or unprotecting a page.

Post-Close Review: In the case of formal closures, especially for more contentious ones, it is not uncommon for participants to question the close or ask for details. This usually takes place on the closer's user talk page or more rarely the close can be challenged by posting to the Administrator's Noticeboard. There is no specific venue for reviewing RfC closes, unlike AfD decisions¹⁰, so it can be difficult to determine what happened after an RfC ended. Another way to relitigate an RfC is to hold another RfC at a later point in time. While it is frowned upon to hold an RfC soon after a closed RfC on the same topic, they can generally happen since consensus may change over time.

4 DATA COLLECTION

In this work, we set out to better understand the RfC process as a whole as well as uncover issues. We collect from two major sources of data to form our analysis, focusing on English Wikipedia. For this project, we consulted with two members of the Wikimedia Foundation, documented the study on Wikimedia's research wiki, and also discussed the study on Wikimedia's research mailing list.

4.1 Frequent RfC Closer Interview Data

Many RfC discussions are formally closed by a neutral third party, which involves writing a summary statement and final decision. We conducted semi-structured interviews with 10 of the most frequent closers on English Wikipedia. In order to find interviewees, we compiled a list of frequent closers. As we did not have a dataset of RfCs yet, we instead scraped the archives of Wikipedia's Administrator's Noticeboard/Requests for closure, a board dedicated to finding closers for an RfC. This yielded links to 2,034 RfCs. We contacted 17 editors who were the most frequent closers and still active on Wikipedia, with 10 accepting.

The interviews were conducted over phone or video call, with the exception of two that were conducted over back-and-forth emails. For the calls, the interviews lasted anywhere from 45 minutes to 1 hour and 30 minutes. Interviewees were compensated \$15 for their time. Due to their desire for anonymity, we only have demographic information for 4 of the 10 interviewees. The average age for the four is 40.75, and all four are male. On average, interviewees have been editors on Wikipedia for 9.9 years, with only 2 of 10 with an edit history under 5 years. 3 out of 10 are administrators.

After asking general questions about interviewees' experience with RfCs, we asked interviewees to walk through the process they go through to decide what RfCs to close and how they go about closing an RfC. We asked them to consider if there were any problems with the RfC process and whether any tools or collaborations could help make the process easier or faster.

Interviews were conducted by the first and second authors. After each interview, it was transcribed and coded by them using a grounded theory approach [9] due to the exploratory nature of the study. As interviews were ongoing, the codes were discussed by all authors and grouped into major themes, including around common concerns about the RfC process as a whole and reasons for why RfCs go stale.

4.2 RfC Discussion and Closing Data

To supplement the interview data, we collected a comprehensive dataset of RfCs. On Wikipedia, there is no archive of pages containing links to all past RfCs. The closest is the Administrator's

¹⁰https://en.wikipedia.org/wiki/Wikipedia:Deletion_review/Active

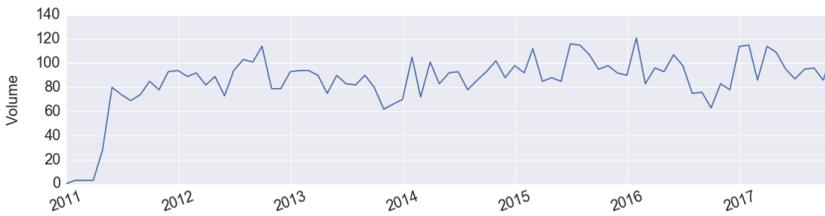


Fig. 3. The number of RfCs initiated each month in our dataset from 2011 to end of 2017.

Noticeboard/Requests for closure but it only contains links to the 2,034 RfCs where someone has explicitly sought a formal close by a neutral third party. Many RfCs get publicized elsewhere, such as on topical lists dedicated to RfCs or on a WikiProject page, or get formally closed through other means. For this reason, we focused on edits left by Legobot, a bot that is automatically triggered when the RfC template tag `{{rfc}}`¹¹ is added to a discussion to create an RfC. Using this strategy, we collected a dataset of 7,316 RfCs beginning from 2011, when Legobot began running, to the end of 2017. From the link of the RfC, we were able to extract all initiator, participant, and closer information, as well as all comments and initiator and closing statements, keeping reply structure intact through the use of libraries such as MediaWiki WikiChatter¹².

We used this dataset to analyze characteristics of contributors as well as the lifecycle of RfCs, from initiation to a final outcome. From this dataset, we can determine RfCs that have been (i) *formally closed* using a template as shown in the left of Figure 2. Analyzing the dataset and the interviews revealed, however, that among the RfCs that did not have the template, not all were simply left stale. Thus, we differentiated between (ii) *informally ended* RfCs and (iii) *stale* ones by tracking the revision history to find when the RfC tag was removed and then retrieving the user account that removed the tag. If it was removed by Legobot, we considered it stale; if the RfC tag was removed by an editor, it was treated as informally ended. While not perfect—for instance, participants might choose to withdraw their RfC but neglect to remove the RfC tag—this method represents our best approximation from the data available to reconstruct what happened.

We were able to categorize 7,087 RfCs out of 7,316 RfCs using this method. Some RfCs were unable to be categorized due to parsing issues. 57.65% of the RfCs ended up formally closed while 42.35% have no formal resolution. Among the unclosed ones, 78% (2,329, 32.86% of all RfCs) remained stale without any closure, while 22% (672, 9.48% of all RfCs) were informally ended. Among the 672 informally ended RfCs, 522 were ended by participants or initiators who took the tag off while 150 were ended by uninvolved editors. Although the former is considered the norm, inspecting the 150 RfCs showed that in some cases uninvolved editors take the RfC tag off if they believe it is no longer necessary or should not have been created. Since in these 150 cases an editor ended a discussion by taking the action of removing the tag, we counted it as informally ended.

To understand qualitatively why RfCs do not get formally closed and the distinction between going stale and informally ending, we randomly selected 40 RfCs that did not get closed and manually inspected and coded the discussion to understand why they were never closed. This analysis was coded by the first and second authors and then discussed by all the authors. Since the reasons may not always be immediately apparent from the discussion, the reasons we were able to identify were informed by our prior discussions with interviewees as well as informal conversations with top RfC participants on Wikipedia.

¹¹<https://en.wikipedia.org/wiki/Template:Rfc>

¹²<https://github.com/mediawiki-utilities/python-mwchatter>

RfC category	Num RfCs initiated	RfC category	Num RfCs initiated
Politics, government, & law	2650	Religion & philosophy	949
History & geography	2573	Wikipedia style & naming	749
Biographies	2123	Wikipedia proposals	634
Wikipedia policies & guidelines	1767	Economy, trade, & companies	585
Uncategorized	1732	Wikipedia technical issues & templates	381
Society, sports, & culture	1634	Language & linguistics	372
Art, architecture, literature, & media	1601	WikiProjects & collaborations	259
Maths, science, & technology	1165		

Table 2. Number of RfCs issued from 2004 to 2017 by categories. One RfC may have multiple categories, for example, `{{rfc|econ|bio}}`.

	Initiators	Participants	Closers
Total number of people	3,346	14,815	759
Percentage of administrators	7.41%	5.11%	23%
Avg (σ) number of edit counts	23,432.16 (74,417.6)	14,055.43 (56,749.5)	39,759.46 (89,639.2)
Median number of edit counts	4,590.5	1,257	17,556
Avg (σ) account age (days)	3,076.63 (1,338.2)	2,260.05 (1,226.1)	3,289.3 (1340.2)
Median account age (days)	3,230.81	2,331.71	3,635.67

Table 3. Overall information about RfC initiators, participants, and closers. The values for initiators and participants was calculated using the whole dataset including unclosed ones as well.

5 PARTICIPATION, PARTICIPANTS, TOPICS, AND DYNAMICS OF RFCS OVER TIME

In this section, we characterize our RfC dataset to demonstrate how the RfC process works currently and how it has evolved over time¹³.

Initiation: From looking at Figure 3, we can see that the number of RfCs initiated over time has remained fairly steady since mid-2011, with 86.5 initiated per month on average across our dataset. Table 3 provides information about the initiator population, which overall is smaller and more experienced than the participant population.

Dissemination: Table 2 shows the number of RfCs initiated within each category from 2004 to 2017. These category counts can give us a rough understanding of areas of relatively higher and lower levels of contention within Wikipedia. When it comes to using RfCs as a means to attract outside input, we find that they appear to work reasonably well. On average, 56.5% of the participants of an RfC are newcomers to the topic of the RfC, determined by considering whether the participant had previously made any edits on the talk page where the RfC took place. However, participants are relatively less experienced than initiators or closers, as shown in Table 3.

Discussion: A discussion's size and shape can affect both the reading and commenting experience. RfCs in our dataset had on average 34.37 comments between 11.79 participants. As a sign of how unwieldy these discussions can get, the highest number of comments on an RfC is 2,375, while the highest number of participants is 831. Both values come from the same RfC¹⁴. Not only can there be many comments but they can create long threads of replies. On average across RfCs, the depth of the longest thread in the discussion was 5.15 comments, while the average depth of any

¹³More details including steps to get data not included in the paper are at https://figshare.com/articles/rfc_sql/7038575

¹⁴https://en.wikipedia.org/wiki/Wikipedia:VisualEditor/Default_State_RFC

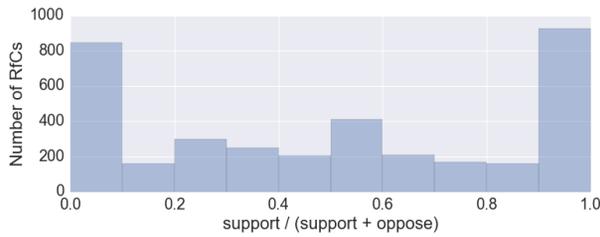


Fig. 4. Ratio of support votes among all votes in RfCs that contain a binary poll.

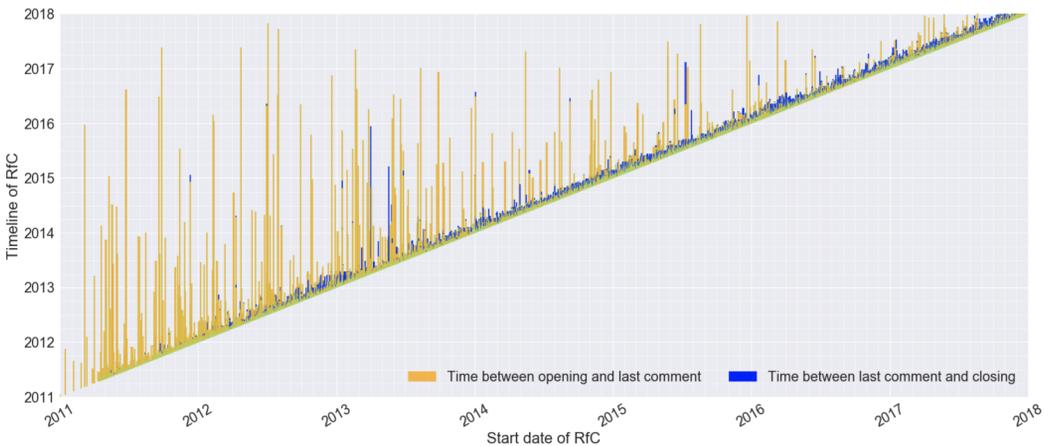


Fig. 5. Timeline of all RfCs showing the length of time for discussion of an RfC after opening it, as well as the length of time between the last comment and the close, if it exists. For each RfC, we draw a vertical line whose x coordinate is the start date and whose y coordinate ranges between start and end date.

comment was 0.39, where a comment that is not a reply to any other comment has a depth of 0. This suggests that RfCs have a mix of deeper back-and-forth discussion as well as many comments simply responding to the initial prompt. Some of these non-threaded comments may come from a dedicated polling section within the RfC. We found that 49.6% of the RfCs in our dataset had an area for a poll. Among RfCs where there was a binary decision, on average there were 5.09 supports and 4.57 opposes, and most polls have a ratio strongly in one direction or the other (Figure 4).

When we calculated the length of the discussion period, we found that the average time between the first comment and the last recorded comment was 44.44 days, with a standard deviation of 160.16 days due to a heavy tail of RfCs that drag on for many months. As noted in our data collection, this duration distribution does include RfCs that were open at the time of this writing. It is also possible that at a future point in time, an editor may reopen any unclosed RfC. When considering only RfCs that were closed, the average length of the discussion was 28.17 days ($\sigma = 75.37$). In Figure 5, we plot the timeline of all RfCs in our dataset, with the yellow lines representing the discussion period and the blue lines representing the time from the last comment to the closing of the RfC if formally closed. As can be seen, there are many discussions that drag on for long periods of time, even years. On average, after the initial proposal, it takes 16.47 days ($\sigma = 76.89$) for the first comment to be made. This is due again to a long tail, and thus the median is 3.91 days.

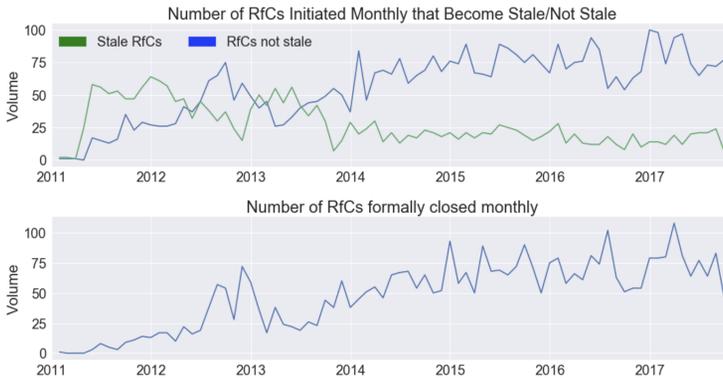


Fig. 6. On the top, the number of RfCs initialized per month is broken down into RfCs that became stale versus RfCs that were either informally ended or formally closed. The number of RfCs formally closed each month is on the bottom.

Closure and Conclusion: As seen in Figure 6, there was a steady increase in RfCs formally closed from 2011 to 2015, lining up with a higher ratio of RfCs that were closed versus RfCs that went stale over that time period, even as RfC initiation volume stayed fairly steady. After 2015, around 20 RfCs still get initiated every month that do not get formally closed. As visualized in Figure 5, the time taken to close a discussion can also be long. For RfCs that eventually were formally closed, on average it took 16.74 days ($\sigma = 25.90$) after the last comment in the RfC. In total, the average RfC time period from initiation to closure for RfCs that were formally closed was 45.56 days ($\sigma = 81.14$). This is about 1.5 times longer than the default 30 days that Legobot allots, with 37% of the time spent on waiting for the closing statement.

As seen in Table 3, closers make up the most experienced but also smallest population, with 23% administrators. From analyzing the closer population over time, we found that the number of active closers has generally been rising since 2011. Some frequent closers we interviewed echoed this finding, saying: “I have a feeling that the backlog is shorter now”. However, this population is also skewed, with 57% of the 759 closers having only closed one RfC, while the account with the most number of closes has closed 352 RfCs.

Post-Close Review: While there are no ways to automatically track what happens to an RfC after conclusion, there is a manually curated page of RfC closure reviews¹⁵ primarily maintained by two editors. It contains 80 RfCs from 2011 to mid-2017, representing 1.1% of the RfCs in our dataset. Of these, 40% of the closes were upheld, and 25% were changed by either being withdrawn, overturned, reverted, or reopened.

6 WHY DO RfCS NOT GET CLOSED?

Through quantitative analysis of our RfC dataset, we found a significant number of RfCs—almost half—that do not get formally closed, with about 78% of those going stale and about 22% ended informally. RfCs going stale can be a problem for maintaining productivity on Wikipedia as editors involved in the RfC may be waiting on the outcome before they feel they can continue editing. It can also be discouraging if an RfC never gets closed when editors put effort into participating in the RfC. While less of an issue, informally ended RfCs may also indicate wasted time due to RfCs that were improperly created or that should never have happened because of pre-existing consensus.

¹⁵https://en.wikipedia.org/wiki/Wikipedia:Administrators%27_noticeboard/Closure_review_archive

We also saw that RfCs can linger for weeks and sometimes months before getting closed. This is problematic if the discussion has gone out of date in that time. One frequent closer stated it as “*There’s also the danger of resurrecting the six month old RfC to close it, that unless you’re really on top of everything that goes on in Wikipedia, which is almost impossible, you just don’t know what’s changed since that RfC.*”

To understand why RfCs do not get formally closed, we conducted a qualitative analysis of 40 randomly selected RfCs that did not get formally closed and also interviewed frequent closers to understand why they would avoid closing some RfCs. Out of the 40 RfCs we analyzed, 22 RfCs contained meta-comments about the issues behind the RfC itself, including warnings against the initiator or participants’ actions, that revealed 6 explicit reasons for staleness or informal ending.

6.1 Problems with Initiators and Initial Proposals

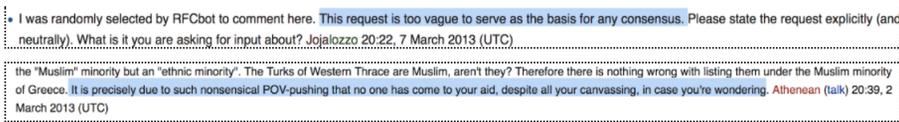


Fig. 7. The first meta-comment points out the initial proposal is too vague while the second notes the initiator’s biased actions.

According to our random sample, issues with initiators had a lot to do with producing unclosed RfCs. 14 out of the 22 RfCs with a meta-comment had an issue related to initiator actions. For instance, sometimes the initiator was not clear with the wording of the request, potentially related to their level of experience. On the other hand, there were more severe cases when the initiator went against the normal consensus decision-making process by biasing the wording of their initial proposal or attempting to canvass by soliciting participation in a non-neutral way, either in their wording or recruitment of certain editors. A few of our interviewees (2/10) mentioned issues with initiators, with one interviewee saying “*An RfC not well-formed—this can happen when the results are unclear because of the structure of the RfC. For example, the RfC might have no clear question...*” This closer went on to say that despite this issue, it can still be possible for a closer to determine editors’ opinions and make a deliberation on what editors actually ended up talking about.

6.2 Behavior of Participants: Bickering and Sock-Puppeting

Four of the RfCs that we examined explicitly mentioned excessive participant bickering, including by the initiator sometimes, which led to more complicated and longer threads that were difficult for newcomers and potential closers to examine. The back-and-forth argumentation was often caused by participants who had a history with each other and had been involved in previous discussions.

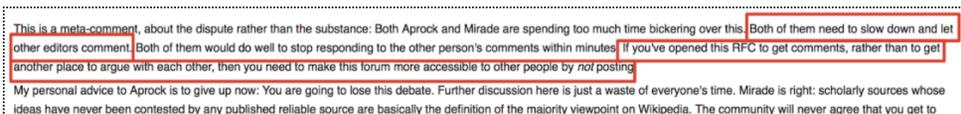


Fig. 8. Meta-comment revealing that the participants’ bickering is making it difficult for other new participants to engage in the RfC.

Three of the frequent closers we interviewed also pointed out that RfCs with lots of bickering would be unlikely to become closed. One interviewee said “*..no one really cares about [the RfC] that*

just gets a lot of bickering back and forth without a lot of substantive discussion. That's the kind of RFC that will often sit for a few months." Another interviewee described how excessive bickering between a few participants might also push away future potential participants: "If one or two participants are trying to reply to everyone who disagrees with them, others may simply not be taking them seriously or have grown tired of repeating themselves."

Three of the interviewees also mentioned actions by participants that try to influence the outcome of the decision by creating multiple fake accounts to create the appearance of consensus (called "sock-puppeting") or by recruiting editors to join a discussion on behalf of that editor (called "meat-puppeting" if recruiting off-wiki and "canvassing" on-wiki). When this happens and another editor notices, an investigation can be called, and the offending editor is routed to formal processes for user conduct. One frequent closer said "If I would have a suspicion that there was socking going on, I probably wouldn't be closing it." This was also a reason why several interviewees spoke strongly about how RFCs should not become a voting process, and mentioned that they give less attention to votes that do not include any rationale or are not based in existing policies due to these concerns.

6.3 Obvious Consensus

There were also cases when the outcome was an absolute consensus, and the participants seemed to think there was no need for a closure. 4 RFCs that we examined were in this category. In these cases, after numerous comments all on one side, eventually a participant just takes the RFC tag off (2/4). The other two RFCs had the tag taken off by a bot, where the participants may have just left the RFC after seeing consensus. Interviewees that mentioned this (2/10) also mentioned that many of these cases are fine to just informally end: "When you have an RFC that has 15 people in support of something and one very loud person opposing it, those are very clear cut outcomes usually and it doesn't necessarily need formal closure". If an initiator is repeatedly starting RFCs to fight a general consensus, they may get referred to a user conduct forum. This category also included cases we saw when many participants responded to the initiator that there is no need for the RFC to begin with, which could be chalked up to lack of initiator expertise.

6.4 Lack of Interest or Expertise from Uninvolved Editors

Other than the three reasons mentioned above that were explicitly mentioned, there were also times where the reason was not clear from the discussion. Among these 18 RFCs without explicit comments about the RFC, we saw both long and short discussions. One possible reason why they did not get closed could be that there was simply lack of interest in the RFC from uninvolved editors. We noticed even in the long discussions, participants were primarily those that were already involved in the discussion before the RFC began.

Should this split proceed?
 would just do it, but lack of objection is not quite the same thing as a show of support, especially on a page with few watchlists. I'm including a |style parameter in the RFC tag since followers of Wikipedia-internal style discussions are usually also interested in the progress of our reader-facing articles on such subjects (or should be!).

Fig. 9. Comment revealing that the lack of overall interest on the page which may influence the outcome of the RFC.

Two of our interviewees also brought this up as a reason why RFCs in topics that attract only a small number of editors might go stale. One interviewee mentioned his own lack of interest in a topic being a factor, saying "When no one cares enough because even if you get it wrong, you've affected one small part of one article that might get 15 views a day, or whatever...I've definitely passed on an RFC because I thought 'this doesn't matter. My time is better used elsewhere.'"

A related issue that several closers (6/10) brought up was lack of expertise in the topic behind the RfC. While closers do not need to be experts on a topic to close it, and in fact should not be too involved in the topic so that they maintain neutrality, they still need to have some knowledge of it or be willing to invest time to learn about it. One interviewee said *“...in some cases a certain amount of background may also be a requirement. This is especially relevant for more technical subjects, such as the sciences... You may be able to remedy this by studying, or it may be better to leave the discussion for someone else to close.”* And although anyone on Wikipedia can close an RfC, if the topic is too esoteric to the majority of frequent closers, then it may never get closed.

6.5 RfC is Too Complicated or Too Contentious

Two other reasons that we were not able to uncover by analyzing RfCs using meta-comments but that were mentioned by several interviewees were RfCs that were too complicated or contentious, with these problems often overlapping. Although there were no meta-comments, we noticed two long discussions containing 136 comments and 84 comments. Three interviewees mentioned that when the RfC is hard to close due to severe contentiousness, they tend to leave it to other closers who can handle it, mostly ones they felt had more authority. One interviewee said *“There were a few that I avoid just because I look at it and think, ‘Whoa, no way.’ Usually it’s the policies and guidelines, anything with like 300 plus comments or where feelings are running very high. Eventually I...think ‘Hmm. That needs one of Wikipedia’s big names to close.’”* Another closer mentioned that they could tell that for some RfCs, no matter how they close it, participants will follow them to their user talk page to question the close, and so they just didn’t want to bother.

Other interviewees (6/10) talked about RfCs that were just too complicated to make sense of. These could be RfCs that were contentious but could also include ones that had a great deal of back-and-forth or many participants, a lot of links to outside sources or relevant policy, or a particularly content-heavy topic. One interviewee described it as, *“And I tried to read it, I looked it over and I realized I couldn’t make heads or tails of it.”* In these cases, an RfC could stay open indefinitely if no closer wants to take on the time to make sense of the discussion and all relevant materials. We also noticed from talking to closers that most of them cited spending on the order of several hours, sometimes over the course of multiple days, closing their most complicated RfCs.

6.6 Interpersonal Issues and “Wikipolitics”

As closers are humans, interpersonal reasons also had to be considered for closures. Two RfC closers mentioned that they do not close RfCs that are related to participants with whom they have a negative relationship. Although this is not a direct reason for staleness overall, it implies that an RfC with an involved editor that has many negative relationships with other editors is more likely to stay open. One interviewee said *“...my relationship with some of the contributors...is not very good. Now suppose people with whom I do not share a particularly good relationship...has initiated the RfC, I don’t generally close it.”* Related to this as well as to the previous reason of an RfC being too complicated, two interviewees discussed how “wikipolitics” play into their decision to close an RfC. One interviewee said *“I closed a discussion where these two people were fighting and they represented two huge factions on Wikipedia...because I did that, if you read my request for [role], that was one of the key points that people opposed it...if you have people who don’t like something you did, even if you did something according to policy, if it’s not popular amongst enough people, they can join their voice with something else and sway a discussion.”* For this reason, a potential closer interested in growing their social capital might steer away from the more contentious discussions.

7 PREDICTING THE LIKELIHOOD OF AN RFC GOING STALE

Building on our analyses of the factors related to closure, we used the RfC dataset we collected to develop classifiers to predict the likelihood of an RfC going stale. Our prediction task is framed as a binary classification problem, taking into account features related to the initiation and unfolding discussion in the RfC as well as characteristics about the article or policy page in question. We first classify RfCs into formally or informally closed versus stale using all the historical data we have on each RfC, minus the closing statement if it exists, to learn what features distinguish stale RfCs. We then consider how a model for predicting the likelihood of an RfC going stale performs as the RfC's life-cycle moves forward in time from initiation.

We used four classification algorithms and compare the performance. The four algorithms are Logistic Regression (LR), Adaptive Boosted Decision Trees (ADT), Random Forests (RF), and Support Vector Machines (SVM) with a radial-basis function kernel. We conduct training and testing on 7,087 RfCs using 61 features. For features with missing data, such as deleted user accounts, we used imputation¹⁶ to insert the mean value instead. 50 trials were conducted with random 40% testing splits, and the resulting performance values were averaged. We also used a tree-based feature selection algorithm to find the most important features, shown in Table 6 based on the feature importance calculated by the ADT model. To determine feature importance we calculated Gini Importance (I) which is the normalized total reduction of the criteria due to the feature.

7.1 Features

Initiator Experience: From the interviews, we learned that initiators may have a large impact on producing RfCs that do not get closed due to lack of experience. For this reason, we calculate measures related to initiator expertise before the RfC took place, such as the *initiator edit count* and *age of the initiator account* in days. The initiator might also be well versed in Wikipedia but a newcomer to the discussion around the topic in question. Thus, we also calculate the *number of revisions to the talk page of the RfC by the initiator*. We finally considered *whether the initiator is an administrator*.

Participant Interest: Another aspect related to likelihood of closure was the ability to attract outside participation towards the RfC, which is the main goal of RfCs to begin with. Thus, we calculate the overall *number of participants* in the discussion so far, as well the *ratio of new participants* so far, where a new participant is one that has not participated on the talk page prior to the RfC.

Participant Experience: In addition to attracting participants, we saw that it was also important that participants have experience. First, an RfC that failed to attract experienced editors may be a factor in lack of interest from frequent closers, who are often also experienced editors. Experienced editors also bring a knowledge of policy and norms, potentially contributing to the quality of the discourse. Finally, sock-puppeting was noted as an issue affecting closure. This could potentially be determined by an unusually low level of experience from participants. We calculate a number of measures related to participant expertise, including the *age of the account of participants*, incorporating the average, standard deviation, sum, and maximum over those values, as well as the *participant edit count*, incorporating the average and sum.

Size and Shape of Discussion: We also found that the size and complexity of the discussion was related to the likelihood of closure. RfCs that generate a lot of discussion may have higher than usual interest and perhaps importance to the community, leading to a vested interest in closure. At the same time, these discussions might scare away potential closers who do not want to invest the time or do not feel like they have the authority. On the other hand, RfCs with very few comments may suggest lack of interest in the topic at hand. To capture these characteristics of both volume

¹⁶<http://scikit-learn.org/dev/modules/impute.html>

Algorithm	Precision	Recall	F1	AUC	Accuracy
LG	0.762	0.868	0.812	0.657	0.73
ADT	0.788	0.864	0.825	0.695	0.753
RF	0.75	0.909	0.822	0.645	0.736
SVM	0.71	0.955	0.815	0.58	0.709
Baseline (most frequent)	0.672	1	0.803	0.5	0.672

Table 4. Average performance of classifiers over 50 trials to predict the closure of RfCs from full data.

and complexity, we measure the *number of comments*, *average depth of replies* per comment, and the *average number of replies* to each comment.

Contentiousness: We learned from the interviews that a discussion’s contentiousness is an important factor considered when deciding to avoid closing a discussion. To measure this, we calculated, for RfCs that had binary polls, *number of supports/opposes*, *ratio of supports over total votes*, and average and sum of *number of replies that support/oppose comments receive*. We also calculated *weighted reciprocity*, which is a measure of the degree of back-and-forth between participants [35].

Tone of Participant Discourse: Bickering was a separate concern that was mentioned in interviews. To get a sense of the tenor of conversations, we calculated features using the frequency of terms taken from commonly used lexicons (indicative word sets) from the Linguistic Inquiry and Word Count (LIWC) software [31]. We examined the average frequency of indicative words over all comments in the discussion so far. First, we considered negative emotionality and affect, using dictionaries for *hostility*, *swear words*, and *anger*, as well as *positive affect*, *negative affect*, and *affect terms in general*. Conversely, we calculated measures for *cognition (cogmech)*, *percept*, and *insight*. Related to prior work on the importance of social aspects of deliberation [5], we also calculate measures for the use of *first-person singular words*, *inclusive language*, and *exclusive language*. Finally, we calculate measures for *certainty* and *tentativeness*.

Initial Proposal Tone and Length: Besides expertise of the initiator, we learned that the quality of the initial proposal can be important, such as if it is too short or has biased language. Thus, we measure the *number of words and characters* in the initial proposal. We also measure all the LIWC terms described in the prior feature category related to tone of participant discourse.

Popularity of RfC and Topic: Finally, we learned from interviews that the interest in the RfC and the underlying topic in question can be a factor. To measure popularity of the RfC, we calculated the *number of words and characters in the RfC* so far, reasoning that longer and more comments indicate greater interest. To calculate interest in the general topic, we also included the *total number of revisions made on the talk page* where the RfC is located. We also look at more recent interest leading up to the RfC, including *number of revisions made 1 week, 2 weeks, 3 weeks, 1 month, and 2 months* prior to the initiation.

7.2 Results

First, we consider the performance of classifiers that make use of features calculated from all data from an RfC up to its closure, if there is one. We report accuracy, precision, recall, F1, and area under the curve (AUC) in Table 4. Adaptive Boosted Decision Trees perform the best overall except for the recall score. They achieve 75.3% accuracy while Support Vector Machines with a radial-basis function kernel perform the worst with 70.9% accuracy. The best accuracy shows a 8.1% increase over the baseline performance of 67.2% of simply picking closed for an RfC’s outcome.

In Table 5 we report precision, recall, F1, AUC, and accuracy for an ADT classifier when using features from only one category at a time. Additionally, in Table 6, we show the top 14 features

Category	Precision	Recall	F1	AUC	Accuracy
Size and Shape of Discussion	0.75	0.903	0.819	0.644	0.733
Participant Experience	0.757	0.86	0.805	0.647	0.72
Participant Interest	0.722	0.897	0.8	0.595	0.699
Contentiousness	0.674	0.98	0.799	0.506	0.669
Popularity of RfC and Topic	0.687	0.947	0.797	0.533	0.675
Tone of Discourse	0.691	0.925	0.791	0.54	0.673
Initiator Experience	0.675	0.984	0.801	0.508	0.672
Initial Proposal Tone and Length	0.673	0.978	0.798	0.504	0.667

Table 5. Performance of ADT classifier to predict the closure of RfCs using features from each category.

Features	Importance	ρ	p
Number of comments	0.08	-0.053	< 0.0001
Maximum Wikipedia age of participants	0.06	0.12	< 0.0001
Cognitive tone of RfC	0.06	-0.049	< 0.0001
Average Wikipedia age of participants	0.06	0.003	< 1
σ of Wikipedia age of participants	0.04	0.215	< 0.0001
Sum of edit counts of participants	0.04	0.147	< 0.0001
Average edit counts of participants	0.04	0.146	< 0.0001
Number of participants	0.04	0.13	< 0.0001
Average reply depth of comments	0.04	-0.13	< 0.0001
Average number of replies	0.04	0.061	< 0.0001
Affective tone of RfC	0.04	-0.054	< 0.0001
Wikipedia age of RfC initiator	0.04	0.028	< 0.05
Hostile tone of initial proposal	0.04	0.013	< 0.5
First person singular word usage of RfC	0.04	0.015	< 0.5

Table 6. Top 14 features in the ADT model incorporating all data, including correlation to closure.

among all 61 features using ADT. Overall, we see that features related to *size and shape of the discussion* best model the data to predict closure, with all three features appearing in the top 10 features. Interestingly, *average number of replies* positively correlated with closure while *number of comments* and *average reply depth of comments* negatively correlated. This may be because longer depth and more comments signify greater complexity and back-and-forth arguing, which may turn some closers off. However, a greater number of replies as opposed to just one-off comments may signal greater interest in the discussion.

Another feature category that models the data well is *participant experience*, with features related to the Wikipedia age of and number of edits by participants listed as important. All of these features were positively correlated with closure, indicating the importance of experienced participants.

While not performing as well altogether, a few features related to *tone of participant discourse* and *tone of initial proposal* were included in the top 14 features. For instance, the affective tone of the discussion was weakly negatively correlated with closure, possibly because words related to emotion may hinder progress of a deliberative discussion.

Lastly, *Wikipedia age of RfC initiator* was also included in the top 14 features with a weak positive correlation with closure. This implies a higher level of an initiator's expertise may help prevent an RfC from going stale.

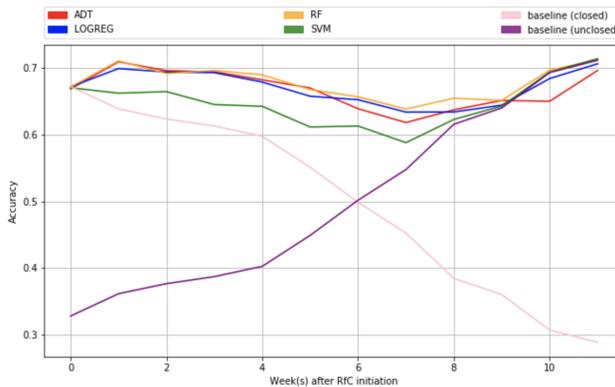


Fig. 10. Change in accuracy over time after initiation up to 11 weeks.

7.2.1 Predicting closure as RfCs progress. While we demonstrated that we can classify closed versus unclosed RfCs from our dataset when provided with all the RfC participation, a more interesting question is how soon after an RfC is initiated can we begin to predict the likelihood of closure with reasonable accuracy. To understand this, we built models that predict closure at different points in time after the start of an RfC. Immediately after initiation, features from the categories of *initiator experience*, *initial proposal's tone and length*, and *popularity of RfC and topic* can be used. As time goes by and participants join the conversation, we can make new predictions about the likelihood of closure using all 61 features and updating their values with historical data.

As time moves forward from initiation, we perform a prediction each week. However, some RfCs get closed during that time—since we already know the outcome of those RfCs looking back in time, we can discard already-closed RfCs in each week's prediction. This means that at each week, we only make predictions on the RfCs that are as of yet unclosed. Since as time goes by, some unclosed RfCs may start to go stale as there are no new comments, we also add a time-based feature to these models which is *the number of days since the last comment up to the current point in time*. We choose to do this instead of discarding inactive RfCs from our prediction since any unclosed RfC might be re-opened at any time by an editor, and this is unknown ahead of time.

As the accuracy over time in Figure 10 shows, all four classifiers start out quite close to a baseline which simply predicts closure for all RfCs, achieving around 66% accuracy. However, as time moves forward across RfCs and only unclosed RfCs remain, the baseline for simply predicting closure for all remaining RfCs drops while the baseline for simply predicting going stale improves. Similarly, as time progresses, the accuracy of the classifiers begin to approach the value presented above with all the RfC participation data baked in, demonstrating how our models can provide timely feedback to participants even just a week after the RfC is initiated. As time goes to 11 weeks after initiation, the baseline prediction of marking all RfCs unclosed begins to approach our models' performances, as most RfCs that are still unclosed at this point are likely to go stale.

8 DISCUSSION

Through a comprehensive analysis of RfCs on English Wikipedia, we examined how RfCs get initiated, discussed, and closed. We found that while the closer population and the proportion of RfCs getting closed is increasing over the last seven years, a large portion of RfCs still do not get closed in a timely manner. From interviews and qualitative analysis of unclosed RfCs, we notice various factors including the nature of discourse and the characteristics and number of discussion

participants can indicate the likelihood of resolution. Using measures informed by interviews and inspection of RfCs, we were able to develop a model that can predict the likelihood of closure at above 70% even a single week after initiation of the RfC. These suggest design considerations for tools that could potentially help make formal deliberations on Wikipedia more effective.

8.1 Tools to help initiators and participants

First, our development of a model for predicting closure could be helpful as a tool for initiators or participants in an RfC to consider ways to avoid going stale. From the model utilizing all participation data, we find that the *participants' interest and experience* were some of the most important factors. In terms of participants' *interest*, it seems crucial to find a way to properly promote an RfC to experienced Wikipedians. Although we did not include it in this work, it would be interesting to find what are the most effective ways to gather interest in an RfC. For example, it might be effective for certain topics to publicize an RfC in particular forums within Wikipedia. Or perhaps certain ways of phrasing the solicitation for participation or closure makes a difference. This kind of feedback, in addition to the feedback that our existing model provides, could help suggest actions for users to take when waiting for more participants or a closer.

As the results imply that participants' expertise is crucial for an RfC to become resolved, this demonstrates the need for designs that can provide editors with relatively lower level of expertise to communicate or receive feedback from more experienced participants. As an interviewee mentioned, participants learn how to provide more reliable sources and policies as evidence by observing or even being won over by more experienced editors' comments during deliberation. A system that can match and invite a group of experienced editors to an RfC that has relatively inexperienced participants could be helpful. Future work could analyze the Feedback Request Service, one of the primary drivers for soliciting participants, to consider whether alternative designs such as pings to volunteers that are not simply random or that happen at different points in the RfC's life-cycle could be beneficial. This is also the case for helping out initiators when writing the proposal, as the initiator's experience was the most crucial factor at the time when one is initiating an RfC.

8.2 Tools to help closers

In addition, we learned about how the *size and shape of discussions* is predictive of going stale. This finding echoes interviewee responses that mentioned spending hours combing through long and deep discussions before writing a resolution, as well as sometimes purposefully shying away from RfCs that were too complicated or contentious. This suggests that tools to better parse and organize these long threaded discussions could potentially help manage the workload. For instance, systems like Wikum [48] that break down a large threaded discussion into manageable chunks to tag, group, and summarize might be of help. A complementary direction could be to consider how similar tools could facilitate closing larger RfCs collaboratively as opposed to by a single individual. While frequent closers tell us that these do happen on rare occasion in Wikipedia on an ad hoc basis, they generally involve collaborations over the draft closing statement through back-and-forth email as opposed to collaboratively understanding and organizing a massive discussion. Additionally, by sharing responsibility it might lesson concerns about "wikipolitics" or lack of authority.

It would also be interesting to consider ways that participants in a deliberation could enrich the representation of the discussion to provide more information that can help closers. For instance, sites like Reddit's ChangeMyView allow discussants to mark when a particular argument has changed their mind on a topic. Since RfCs are meant to be consensus-driven as opposed to voting-based, the deliberation should ideally be causing people to come together over time. Illuminating points of consensus and persuasive arguments would be helpful to closers and may speed up consensus since new participants will more quickly get up to speed. Similarly, an idea that a frequent closer

mentioned was a tool to allow one to see the RFC discussion unfolding over time, so that he could notice changes in people's interest and opinions as time went on. Currently, he achieved this by going through the revisions on the RfC page by page, which he found to be tedious.

8.3 Implications beyond Wikipedia

Our study also presents insights that can be valuable to systems and processes within peer production and deliberative communities beyond Wikipedia.

8.3.1 Task- and Proposal-based processes. The findings from our model may be of help to peer production communities that must assign and track tasks or issues that community members propose. Examples include open source communities that have task or issue processes with a definite start and end, similar to the RfC process. Other examples include platforms centered around creating proposals, voting and discussing them, followed by implementation, such as Climate CoLab [18] or Decidem Barcelona [1]. In the case of open source contributions on sites like GitHub, researchers have uncovered problems with contributors reporting issues that are incomplete or invalid, which causes difficulties for developers [4]. Newcomers also face difficulties in contributing to open source projects because of lack of answers to their inquiry or their own communication behavior [36]. The findings related to *participants' expertise and interest* as well as *initiator's expertise* in our model could be helpful for mitigating these problems by providing insights on how resources including experienced initiators and participants should be allocated.

8.3.2 Deliberative processes. Communities seeking to have productive discourse could also benefit from the implications from our model. Many platforms for discussion do not have definitive formal resolution processes like RFCs, focusing only on the deliberation aspect. For instance, in platforms like Kialo¹⁷ or ConsiderIt [37], the discussion artifact, or resulting issue map, is the desired outcome. These platforms do not aim for a definitive end of the discussion but rather aim to have a fair and productive deliberation while mapping the space of opinions.

Whether or not the platform requires a definitive "task" to complete, systems seeking productive discourse will likely benefit from the findings of our model's features related to *contentiousness, excessive bickering, and various tones of discourse*. Another finding from the RfC process that might translate to deliberative systems is the more formal starting and ending nature of RFCs. Systems where discussions go on indefinitely or where threads with the same issue repeatedly arise might benefit from having a procedure that lets participants stop and move on to something else or work towards some conclusion. An interviewee mentioned: "*RfCs can bring even the most intractable disputes to a conclusion and allow editors to move forward despite holding extremely diverse opinions. A few times, I've even seen an entire topic area return entirely to quiet, 'normal' editing at the conclusion of a particularly important RfC*", emphasizing that RFCs provide a way for Wikipedians to move on and not get stuck on a particular issue. This is healthy for the community because editors can allocate their resources to different issues instead of wasting effort on a single one. This nature of RFCs may provide insights to platforms like Reddit's ChangeMyView, where there may exist participant fatigue around certain topics. Systems like Wikum [48] for collaborative summarization of discourse might be a vehicle for providing a sense of productivity or resolution.

9 FUTURE WORK AND LIMITATIONS

There were several features and related datasets in Wikipedia that we were interested in collecting but require additional work and some manual effort to gather. For instance, we could see whether models perform differently for RFCs in different categories. To match categories to RFCs, we need

¹⁷<https://www.kialo.com>

to parse the revision histories for the RfC's talk page to find the revision that adds the RfC tag. This is because the revision histories of each RfC category page does not provide the link or title of the RfC, only RfC IDs that disappear once an RfC tag is taken off.

One limitation of our study is that participants' number of edits made on Wikipedia were not captured over time since they were retrieved by using the MediaWiki API, which only provides users' number of edit counts up to now. If we can easily collect the number of edits made on Wikipedia by an editor at the time of each comment, it may be that we can achieve more accurate results. Finally, some of the unclosed RfCs that were most recent in our dataset, such as ones in 2017, might at this point or in the future become closed because of more recent activity after our data collection.

10 CONCLUSION

In this work, we provide a case study of Requests for Comment on English Wikipedia to examine online dispute resolution. We learned from interviews and qualitative analysis the reasons why many RfCs go unclosed. We identified features that distinguish formally closed RfCs and informally ended ones from ones that remained stale, achieving 75.3% accuracy. The results show that participants' interest and experience are significant factors, along with size and shape of the discussion. We also built models to timely predict RfCs from the start of the initiation to different points in their progression, with the best model reaching above 70% after just one week.

ACKNOWLEDGMENTS

The authors would like to thank the interviewees and the members of the Wikimedia Foundation, especially Jonathan Morgan who gave valuable feedback.

REFERENCES

- [1] Pablo Aragón, Andreas Kaltenbrunner, Antonio Calleja-López, Andrés Pereira, Arnau Monrde, Xabier E Barandiaran, and Vicenç Gómez. 2017. Deliberative Platform Design: The case study of the online discussions in Decidim Barcelona. In *International Conference on Social Informatics*. Springer, 277–287.
- [2] Yochai Benkler. 2002. Coase's Penguin, or, Linux and" The Nature of the Firm". *Yale law journal* (2002), 369–446.
- [3] Ivan Beschastnikh, Travis Kriplean, and David W McDonald. [n. d.]. Wikipedian Self-Governance in Action: Motivating the Policy Lens. In *ICWSM*. 27–35.
- [4] Tegawendé F Bissyandé, David Lo, Lingxiao Jiang, Laurent Réveillere, Jacques Klein, and Yves Le Traon. 2013. Got issues? who cares about it? a large scale investigation of issue trackers from github. In *Software Reliability Engineering (ISSRE), 2013 IEEE 24th International Symposium on*. IEEE, 188–197.
- [5] Laura W Black, Howard T Welser, Dan Cosley, and Jocelyn M DeGroot. 2011. Self-governance through group discussion in Wikipedia: Measuring deliberation in online groups. *Small Group Research* 42, 5 (2011), 595–634.
- [6] Luciana Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. 2006. Temporal evolution of the wikigraph. (2006).
- [7] Moira Burke and Robert Kraut. 2008. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 27–36.
- [8] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1101–1110.
- [9] Kathy Charmaz and Linda Liska Belgrave. 2007. Grounded theory. *The Blackwell encyclopedia of sociology* (2007).
- [10] Justin Cranshaw and Aniket Kittur. 2011. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1865–1874.
- [11] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 699–708.
- [12] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* (2013).

- [13] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1175–1184.
- [14] Oliver Ferschké, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 777–786.
- [15] Andrea Forte and Amy Bruckman. 2008. Scaling consensus: Increasing decentralization in Wikipedia governance. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*. IEEE, 157–157.
- [16] 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*. ACM, 201–202.
- [17] Marit Hinno Saar, Toomas Hinno Saar, Michael Kummer, and Olga Slivko. 2017. Wikipedia Matters: a significant impact of user-generated content on real-life choices. (2017).
- [18] Joshua Introne, Robert Laubacher, Gary Olson, and Thomas Malone. 2011. The Climate CoLab: Large scale model-based collaborative planning. In *Collaboration Technologies and Systems (CTS), 2011 International Conference on*. IEEE, 40–47.
- [19] Sirkka L Jarvenpaa and Dorothy E Leidner. 1998. Communication and trust in global virtual teams. *Journal of Computer-Mediated Communication* 3, 4 (1998), JCMC346.
- [20] Brian Keegan and Casey Fiesler. 2017. The Evolution and Consequences of Peer Producing Wikipedia’s Rules. (2017).
- [21] 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.
- [22] 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.
- [23] 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*. ACM, 453–462.
- [24] Travis Kriplean, Ivan Beschastnikh, David W McDonald, and Scott A Golder. 2007. Community, consensus, coercion, control: cs* w or how policy mediates mass participation. In *Proceedings of the 2007 international ACM conference on Supporting group work*. ACM, 167–176.
- [25] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. 2012. Is this what you meant?: promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1559–1568.
- [26] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. 2011. When the wikipedians talk: Network and tree structure of wikipedia discussion pages.. In *ICWSM*. 177–184.
- [27] Christoph Lattemann and Stefan Stieglitz. 2005. Framework for governance in open source communities. In *System Sciences, 2005. HICSS’05. Proceedings of the 38th Annual Hawaii International Conference on*. IEEE, 192a–192a.
- [28] Keith Maki, Michael Yoder, Yohan Jo, and Carolyn Rosé. 2017. Roles and Success in Wikipedia Talk Pages: Identifying Latent Patterns of Behavior. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 1026–1035.
- [29] Elinor Ostrom. 2015. *Governing the commons*. Cambridge university press.
- [30] Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society* 6, 2 (2004), 259–283.
- [31] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
- [32] Jenny Preece. 2000. *Online communities: Designing usability and supporting sociability*. John Wiley & Sons, Inc.
- [33] Jodi Schneider, John G Breslin, and Alexandre Passant. 2010. A content analysis: How Wikipedia talk pages are used. (2010).
- [34] Jodi Schneider, Alexandre Passant, and John G Breslin. 2011. Understanding and improving Wikipedia article discussion spaces. In *Proceedings of the 2011 ACM Symposium on Applied Computing*. ACM, 808–813.
- [35] Tiziano Squartini, Francesco Picciolo, Franco Ruzzenenti, and Diego Garlaschelli. 2013. Reciprocity of weighted networks. *Scientific reports* 3 (2013), 2729.
- [36] Igor Steinmacher, Tayana Conte, Marco Aurélio Gerosa, and David Redmiles. 2015. Social barriers faced by newcomers placing their first contribution in open source software projects. In *Proceedings of the 18th ACM conference on Computer supported cooperative work & social computing*. ACM, 1379–1392.
- [37] Hans Stiegler and Menno DT de Jong. 2015. Facilitating personal deliberation online: Immediate effects of two ConsiderIt variations. *Computers in human behavior* 51 (2015), 461–469.
- [38] Besiki Stvilia, Michael B Twidale, Linda C Smith, and Les Gasser. 2008. Information quality work organization in Wikipedia. *Journal of the Association for Information Science and Technology* 59, 6 (2008), 983–1001.
- [39] Róbert Sumi, Taha Yasseri, et al. 2011. Edit wars in Wikipedia. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 724–727.

- [40] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 613–624.
- [41] Dario Taraborelli and Giovanni Luca Ciampaglia. 2010. Beyond notability. Collective deliberation on content inclusion in Wikipedia. In *Self-Adaptive and Self-Organizing Systems Workshop (SASOW), 2010 Fourth IEEE International Conference on*. IEEE, 122–125.
- [42] Fernanda B Viegas, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. 2007. Talk before you type: Coordination in Wikipedia. In *System sciences, 2007. HICSS 2007. 40th annual Hawaii international conference on*. IEEE, 78–78.
- [43] 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.
- [44] Howard T Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. 2011. Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference*. ACM, 122–129.
- [45] Dennis M Wilkinson and Bernardo A Huberman. 2007. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis*. ACM, 157–164.
- [46] Diyi Yang, Aaron Halfaker, Robert E Kraut, and Eduard H Hovy. 2016. Who Did What: Editor Role Identification in Wikipedia. In *ICWSM*. 446–455.
- [47] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. 2012. Dynamics of conflicts in Wikipedia. *PLoS ONE* 7, 6 (2012), e38869.
- [48] Amy X Zhang, Lea Verou, and David R Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *CSCW*. 2082–2096.

Received April 2018; revised July 2018; accepted September 2018