Beyond notability. Collective deliberation on content inclusion in Wikipedia

Dario Taraborelli
Centre for Research in Social Simulation
University of Surrey
Guildford GU2 7XH, UK
d.taraborelli@surrey.ac.uk

Giovanni Luca Ciampaglia
Faculty of Informatics
Università della Svizzera Italiana
Via G. Buffi 13, 6900 Lugano, CH
giovanni.luca.ciampaglia@usi.ch

***Abstract*—According to its guidelines, content inclusion in Wikipedia is determined by "notability" criteria, i.e. standards that establish whether a topic is sufficiently encyclopaedic to deserve a dedicated article. However, due to the decentralised nature of Wikipedia's governance model, decisions as to whether specific articles are "notable", and should therefore be kept or deleted, are made by groups of users who aim to reach consensus through a procedure known as an *Article for Deletion* discussion. In this paper we study the structure and temporal dynamics of these article deletion discussions as a form of collective deliberative process and indicate biases that affect the outcome of these discussions and, as a result, potentially undermine the role of notability as a quality standard for inclusion.**

I. Introduction

A key aspect of governance and quality control processes in commons-based peer production systems [1] is their open and participatory nature. Traditional forms of hierarchical organisation are sometimes adopted to address particular aspects of the governance of such systems. However, in the general case governance-related decisions in these systems are reached through participatory processes that involve large numbers of users. A project such as the Wikipedia is similarly based on mechanisms designed to allow the effective distribution of maintenance tasks and to support collective decision-making, as required by the daily curation needs of the project. Decentralised curation and deliberation are the only answer to the question of how to govern a project of a similar size in a timely and effective way, but are far from representing the optimal solution. Tasks and functional roles in peer production systems are typically self-assigned, so that contributors can decide to participate in a variety of tasks as they see fit. As a result, the very strength of a peer production system (the decentralised character of its governance) can also be seen as the main source of possible biases and suboptimal solutions, e.g. the allocation of inadequate resources to specific kinds of task. Distributing decisions to self-appointed participants is a phenomenon that deserves attention: collective decision-making in Wikipedia is becoming an interesting new field of research that can shed light on the effects and possible shortcomings of distributed governance mechanisms [2], [3].

A. Related research

Participation in discussions on information quality standards and their enforcement in Wikipedia was first addressed, to our knowledge, in [4], who found that user participation in information quality decisions exhibits a long-tailed distribution, suggesting that a small number of editors participate in almost every vote while most users vote in very few or do not vote at all. The majority of Wikipedia editors, they speculate, have probably never come across Wikipedia's quality-related policies and guidelines.

The issue of topic inclusion/exclusion in Wikipedia entries was further addressed by [5], who conducted an extensive analysis of the deletion log of articles in a 3-year period ending in December 2007. Their study focussed on deletion rates of articles in the English Wikipedia and suggests that deletions tend to happen early during the lifecycle of entries. This study also attempted to identify the potential causes of observed peaks in article mortality as a function of external events and actions undertaken by the governing body of Wikipedia, the *Wikimedia Foundation*, likely to have affected the quality standards shared by the Wikipedia community. Finally, they compared article popularity, measured on the basis of the number of views that an article received within a given timeframe, with the probability of its being deleted and found that article survival probability broadly follows the popularity of the article in terms of readership and presence in search engines.

The debate behind Wikipedia's governance model was reviewed in [6], who looked at the history of the conflict about inclusion norms. By pointing at the lack of formal mechanisms to effectively govern conflicts about inclusion, he suggests that the project should "return to its inclusionist roots" through the creation of a functional process for managing disputes on content inclusion as well as the formulation of an explicit community social contract model to regulate these disputes. He notes that the return to elements of traditional organization may contribute to the long-term sustainability of peer production systems such as Wikipedia.

Content inclusion in Wikipedia is governed by community-produced policies, which evolve and are refined over time. While the authors of [3] observed that the set of commonly

invoked policies has become more and more consistently used across project members, qualitative research conducted by [7] points at the inertia of the collaborative policy making process in Wikipedia, suggesting that there has been little change in policy since Wikipedia's surge of popularity. They suggest that policies regulating participation and content creation in Wikipedia (including norms that govern inclusion/deletion) are crystallised since their original creation: the growing number of active contributors to the Wikipedia is likely to make it impossible to achieve consensus on new policies amending or reforming the original ones. This is particularly crucial for a project that is constantly growing in content and population and hence faces a permanent challenge to identify appropriate forms of governance and content curation strategies.

II. Mechanics of inclusion and deletion

Historically, the right to remove an article, as a measure to fight vandalism, has been a privilege of a selected number of Wikipedia editors with special administrator privileges. Over time, the Wikipedia community has become more and more involved in governance and curation tasks (see [8]). Wikipedia's firm belief in open participation has proved to be one of the main drivers of its growth compared to relatively more closed systems [9]. Supporting open participation, however, raises the question of how to preserve quality and how to achieve balance between quality and quantity of content.

As a result of this progressive decentralisation of governance, several maintenance routines, such as suggesting that an article be considered for deletion, can now be directly initiated and run by regular editors: an admin's intervention is only required to finalise a decision based on community consensus.

The deliberative procedure behind user-driven deletion proposals has seen important changes over time, but eventually crystallised in what is currently known as an Article for Deletion discussion. An *Article for deletion* [1] (hereafter: AfD) is the procedure whereby Wikipedia editors collectively discuss whether an article should be deleted.[2] Articles that are nominated for deletion are typically discussed for a minimum of 7 days, during which feedback from the community is solicited in order to reach consensus. Editors can participate in an AfD discussion by casting one vote and adding optional comments to motivate their decision. Editors participating in the discussion can cast any of the following options:

- **Keep** (hereafter: K) to recommend that the article be kept as is, possibly after some improvements;
- **Delete** (D, to support the initial nomination and recommend that the article be deleted;
- **Merge** (M), to request that the article be merged with another one;
- **Redirect** (R), to request that the article be removed and its title redirect to another article;

The final step of the process requires an editor with administrator privileges to review the discussion, check if a form of consensus was reached, and enforce the corresponding decision.

[1] http://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion

[2] We do not include in this study other forms of deletion procedures allowed by Wikipedia but not requiring community consensus building.

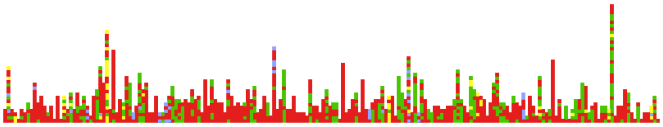


Fig. 1: A sample of 200 AfD sequences (red: *D*, green: *K*, blue: *R*, yellow: *M*). The first vote with which an AfD is created (a *D*) is not represented.

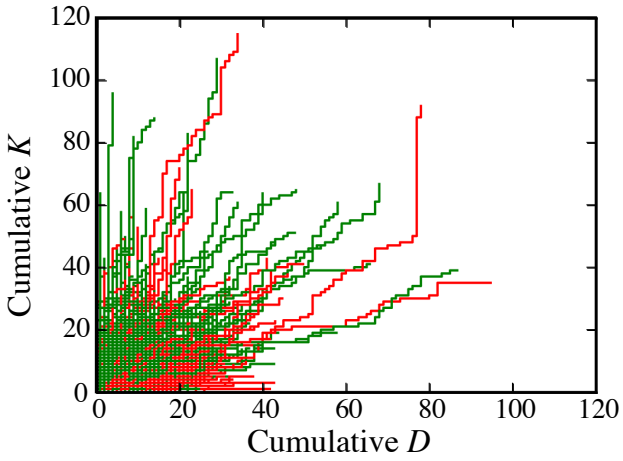


Fig. 2: Vote trajectories and outcomes. Each line represent the cumulative number of *Keeps* and *Deletes* at any given step of the vote sequence. Color codes the outcome of the AfD. Red: votes that resulted in the page being deleted; green: page kept, redirected or merged.

A. Norms regulating deletion

Wikipedia is rich in policies and guidelines that editors are expected to be familiar with before nominating an article for deletion. However, an AfD nomination procedure is fairly simple and anyone can issue such a nomination and trigger a discussion by simply adding a deletion template to an article and thereby initiating an AfD procedure. The deletion guidelines specify the reasons for which an article can be nominated for a deletion as well as the conduct to be held by participating editors throughout and after an AfD discussion. It is interesting to note that the guidelines discourage the term "vote" to refer to the AfD procedure, which should be understood as a "a means to gauge the degree of consensus reached so far". Indeed, the whole process is "conservative" in the sense that the deletion guidelines actually specify that an admin should only delete a page if there is evidence that consensus has been reached to do so. Otherwise, a page should be kept and improved through standard editing, unless another more appropriate recommendation (such as merging) emerges from the discussion.

B. Notability

When content inclusion is debated, the main standard adopted in Wikipedia to decide whether a topic merits a dedicated article is the so-called "notability" of the topic.[3] Notability is the primary quality norm that determines content inclusion and exclusion in Wikipedia and to which the community resorts in case of disputes. As a general rule, notability identifies topics that can be considered of an "encyclopaedic nature". Different kinds of subjects (such as biographies of living people or events) have spawned a complex set of notability criteria. The common feature of a notable topic, however, as defined by the guidelines, consists in its been "noticed" to a significant degree by reliable secondary sources: a topic that is adequately covered by reliable sources and does not conflict with other general guidelines about inclusion in Wikipedia is hence a good candidate for a Wikipedia article. Conversely, an article on a topic that is deemed *non notable* is by definition unsuitable for inclusion in Wikipedia. The alleged non-notability of a topic is by far the main driver for nominating articles for deletion and it has been estimated that up to one third of reasons adopted for deleting an article are indeed related to its notability. [5]

The notability standard has been at the centre of a major, ongoing debate in the Wikipedia community and has been harshly criticised by part of the community of contributors and users. A common flaw in the definition of notability frequently pointed at, is, for example, the fact that it actually shifts the burden of deciding whether a topic is encyclopedic to the "reliability" of the sources that describe it. More generally, the debate on the very nature of notability has spawned the creation of movements of like-minded editors who share similar views on what the mission of Wikipedia should be and what topics it should encompass.

C. Inclusionism and deletionism

A feature appeared in 2006 in *The Washington Post* [10] noted how "wiki-worthiness has quietly become a new digital divide, separating those who think they are notable from those granted the imprimatur of notability by a horde of anonymous geeks." The existence of opposing groups of editors with strongly diverging opinions as to what "notable" means and what Wikipedia should include is witnessed by the existence of a long-standing public debate around "inclusionism"[4] and "deletionism"[5]. Not only have these two views become the most violently and extensively debated quality-related issues in Wikipedia's policy making. They actually gave rise to organised movements of editors actively involved in promoting a specific interpretation of Wikipedia's inclusion criteria.[6,7] Editors can signal their affiliation to either movements by adding a template and category on their user page, which thus publicly identifies them as affiliates of these associations.

[3]http://en.wikipedia.org/wiki/Wikipedia:N
[4]http://meta.wikimedia.org/wiki/Inclusionism
[5]http://meta.wikimedia.org/wiki/Deletionism
[6]http://meta.wikimedia.org/wiki/Association_of_Inclusionist_Wikipedians
[7]http://meta.wikimedia.org/wiki/Association_of_Deletionist_Wikipedians

A vast majority of users participating in AfD, however, are likely to display inclusionist or deletionist tendencies in an implicit way, without a public affiliation to any of these movements. To understand the general perception of the Inclusionism/Deletionism debate by the community, we conducted an informal survey as preliminary step of the present study. We invited the top 50 participants in AfD discussions (as identified from our AfD datasets, see below) to express their opinion about these movements. All of the 22 editors who responded (including 7 Wikipedia admins) reported being familiar with both the "Inclusionism" and "Deletionism" movements. The majority of the respondents ($N = 10$) declared "considering themselves Inclusionists", while a smaller proportion ($N = 7$) declared being neutral with respect with these movements and a minority ($N = 5$) reported sympathizing with the Deletionist view. Whereas most respondents believe AfD to be the most important place where editors to learn the tacit rules of consensus building in Wikipedia, many expressed concerns about the flaws in the current AfD system and the number of false positives/negative that it can generate. One respondent (who declared his/her neutrality with respect to the two movements), observed: *The inconsistency of results is troubling and in my opinion has worsened the Wikipedia and largely reflects lots of people who know nothing about a subject or minimal experience with Wikipedia and its policies having as much voice as someone who does know and is familiar. AfD discussions are heavily weighted by whoever shows up at any particular discussion, and policies are disregarded to afford those who show up a free voice unimpeded by any checks and balances.* One respondent reported the existence of off-Wikipedia mailing lists, which–in conjunction with an Inclusionist subproject called "Article Rescue Squadron"–coordinate the recruitment of editors in order to stack votes in favour of salvaging AfD nominated articles.

Recruitment of voters through organised movements provides evidence supporting the fact that at least part of the votes cast in AfD discussions may be due to *strategic behaviour*. These considerations raise the question of the accuracy and impartiality of the current AfD system. Our goal in the present study is to identify potential biases in the functioning of this large-scale system of collective deliberation by addressing the following empirical questions:

- are norms regulating AfD discussions generally respected?
- to what extent AfD votes are driven by strategic recruitment as opposed to spontaneous participation?
- do the order, structure and temporal dynamics of AfD discussions affect the behaviour of individual participants?
- to what extent the rate of user participation to AfD determines their overall outcome?
- are user responses to AfD homogeneously distributed over the 4 different options or are there noteworthy asymmetries in voter behaviour?

III. Research questions

In this study we aim to analyse Wikipedia's article for deletion procedure as a complex form of collective decision-making process bearing on the maintenance of quality standards in Wikipedia. Our goal is to identify properties of AfD discussions that may indicate biases in the process of deletion of content. This could suggest that notability may not be the sole reason determining the inclusion/exclusion of content in Wikipedia. An unbiased decision about the deletion of an article would be one in which the notability and enclyclopedic nature of a topic is judged *on its own merit* and in which: (1) participants are familiar with the official norms and policies governing votes for deletions and comply with them when casting their vote; (2) no user is influenced by votes previously cast by other users, i.e. each choice is independent of the other choices; (3) no user is motivated by strategic reasons for voting in favour or against the deletion of an article. If a vote for deletion was entirely unbiased, and assuming that notability could be assessed in a fairly objective way, we expect to observe a fairly straightforward progress towards consensus. Conversely, if votes for deletion are affected by individual and collective biases, we expect to observe more complex patterns of consensus-making, including cases in which consensus cannot be reached at all. This raises the question of how to tell apart controversial from non-controversial votes for deletion.

Tackling this distinction is actually a precondition to be able to tell apart AfD that are genuinely based on the merit of the topic (i.e. driven by notability criteria) from votes where other factors intervene in driving collective decisions. However, since "notability" is not a concept that can be easily operationalised and measured, we need to identify properties of a voting sequence that may be indicative of other factors driving deliberating behaviour. We start by providing a characterisation of what we expect to be the typical properties of a non-controversial vote.

Non-controversial votes. We take a short sequence of votes in which all (or the vast majority) of voters display consensus (i.e. consistently vote in the same way) as indicating a case in which the topic is easy to assess on its own merit. We expect non-controversial votes to be short because if consensus is perceived as being already reached after the initial series of n votes, incentives to participate by casting vote $n+1$ will be low. In other words, we do not expect to observe long sequences of votes of the same kind in which no dissent is expressed and we expect non-controversial votes to terminate more quickly than average votes. Note that even in the case of short, non-controversial votes one can make the assumption that subsequent votes are influenced by the first votes and not based on the merit of the topic. However, we assume that sequences in which votes follow an apparently herd-like pattern but stop after a small number of votes cast are indicative of non-controversial discussions, as the opinion of the first voters are as effective as a direct consideration of the merit of the topic as a ground for a decision. As an example, the article on US pop punk band `A Change of Pace` survived a nomination for deletion, after a short sequence of votes ($K = 5$) rebutting the lack of notability claim raised by the user who started the AfD.

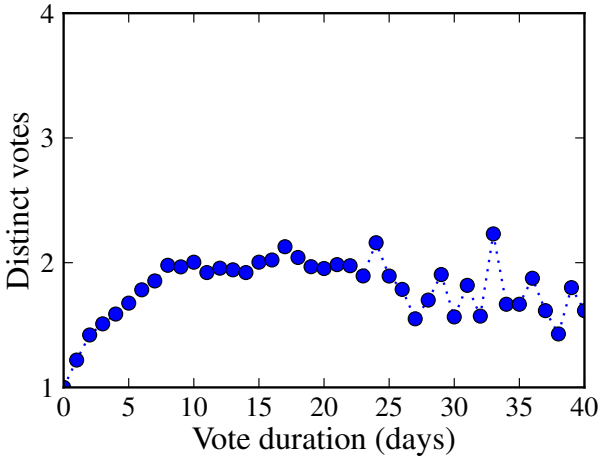


Fig. 3: Number of different types of votes expressed in votes as a function of the vote duration in days.

Controversial votes. By contrast, we characterise controversial votes are longer sequences in which consensus is not as straightforward to reach as in non-controversial votes, i.e. sequences of votes in which people perceive the need to express their opinion despite the fact that a substantial number of votes has already been cast. We assume that reasons for casting the $(n+1)^{th}$ vote after a sequence of n votes may either depend on (a) the desire to express dissent from an opinion previously expressed by a majority of users in order to try and influence the behaviour of subsequent voters (*vote overturning*) or (b) the desire to provide further support to votes already expressed but perceived as not sufficiently strong to resist further opposition (*vote stacking*). This form of selection bias has been found in ratings of books on *Amazon* and movies on *IMBD*, suggesting that this may be a general feature of collective deliberation processes in online systems [11]. It should be noted that an initially controversial vote may later be perceived as non-controversial as soon as a critical mass of voters start aligning their votes in a way that signals consensus to the remaining voters, and thus making it less likely for opponents to cast an overturning vote to express their late dissent. When this happens, we expect the sequence to reach an end in a relatively rapid number of votes after a sequence of homogeneous votes. As an example, the Wikipedia article on British singer `Susan Boyle` was nominated for deletion and triggered 110 votes ($D = 15; K = 94; M = 8$) in slightly more than 3 days of open AfD votation, that ended with the decision to keep the article.

By focussing on votes that display a variety of user responses, we aim to study whether there are factors that affect the dynamics of AfD discussion that go beyond the sheer assessment of a topic's notability. In particular, we intend to

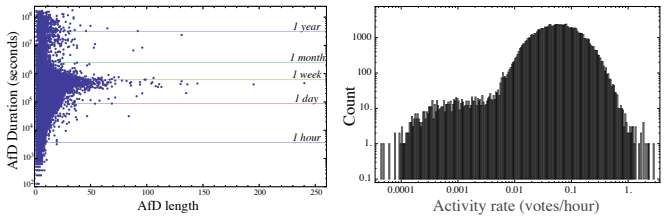


Fig. 4: (*left*) AfD duration as a function of their length; (*right*) Distribution of activity rates for AfD lasting more than 24h;

address the following research questions:

A. *Norm compliance.* Do AfD comply with norms that specify their correct functioning, duration, and mode of participation?

B. *Herding effects.* Is there evidence of informational cascades, suggesting that individual choices may be affected by previously cast votes?

C. *Voter heterogeneity.* Are voters homogeneous in their voting behaviour or are there tendencies that differentiate how Wikipedia users participate in an AfD?

D. *Strategic behaviour.* Is there evidence that AfD may be affected by factions strategically recruiting voters to stack votes in favour or against the deletion of an article?

IV. Dataset

We collected and analysed data on a total of $196,160$ votes for deletion that took place in the period going from January 1999 to May 2010. The dataset includes $1,198,829$ unique votings cast by $68,950$ individual users. Voting users include unregistered users (identified by the IP address of their connection, 14.8% of the total voters) as well as registered users contributing with their username (85.2% of the total number of voters). The dataset records, for each vote, the title of the AfD it was cast in, the timestamp, the user name of the voter and the actual vote. We do not have any record for votes other than the four main choices of an AfD, $\{K, D, M, R\}$ (i.e. abstentions, comments, etc.) nor we have the text of the optional comment inserted by the user.

V. Results

A. Norm compliance

Some macroscopic properties of the dataset provide answers to the question whether AfD comply or not with norms that officially regulate their functioning.

Deletion guidelines recommend that **majority voting** should not be used as a determining factor of whether a nomination succeeds or not. As a consequence one should expect that no direct correspondence be found between the number of votes cast by participating editors and the final decision made by an administrator. Figure 2 displays, for all sequences in our dataset, the tally of votes for the choice of keeping the page versus that of deleting, at any given point during the vote sequence. The plot indicates that there is no clear boundary between votes resulting in the page being deleted (red) vs. other outcomes (green). This can contrasted to the identical plot presented by [2] as part of a study of deliberative process in Wikipedia for admin nomination, which clearly show that these votes function on a principle of majority voting that does not apply to AfD.

An analysis of the **timing of participation** in AfD (III) indicates that a large majority of discussions last well below and above the indicated limit of 7 days from the nomination, with AfD terminating a couple of hours or a single day after nomination as well as a considerable number of discussions lasting several weeks, months or even more than a year. The number of votes cast in an AfD discussion is clearly not correlated to the temporal duration of the discussion.

The variety of temporal spans over which AfD last indicates that the 7-day rule is not only hard to be enforced, but possibly inadequate as a temporal window within which **consensus** has to be reached. Very short sequences indicate indeed that votes are frequently aligned, but longer sequences display a large variety of voting behaviour. A t-test rejects the hypothesis ($p = 0$) that AfD shorter than 1 day have the same number of vote types (i.e. the cardinality of the set of expressed votes) as longer AfD, consistently with our prediction about non-controversial votes. Figure 3 shows that the number of vote types grows with vote duration roughly until 10 days.

B. Herding effects

An "information cascade", often referred to as "herding behaviour" (see [12]; for a review on herding in humans see [13] and references therein), occurs when people form beliefs on the basis of information obtained through the observation of the behaviour of others. These phenomena are called "cascades" when the options expressed by the agents who act first influence the choices made by subsequent agents, which in turn influence later choices and the overall temporal sequence.

Here we want to test for the presence of information cascades in the voting sequences (series of contiguous votes of the same type) as an indicator of potential cognitive biases in AfD participants' behaviour. As previously stated, an unbiased decision would require that voters are homogeneous in their preferences and that they are not influenced by the previous votes. This is equivalent to say that the votes are IID sequences where the probability of voting for option $o \in \{K, D, R, M\}$ is given by the estimated baseline probabilities f_o shown in Table V-B. This also means that the expected number of votes X_o that option o would receive in a sequence of N votes is equal to $f_o N$. In figure 5 we plot this quantity together with the expected prediction of the IID model. The plot shows that there is some level of agreement until $N = 10$, but then,

o	D	K	M	R
f_o	0.6256	0.2992	0.0504	0.0248

TABLE I: Estimated baseline probabilities for each of the four voting options.

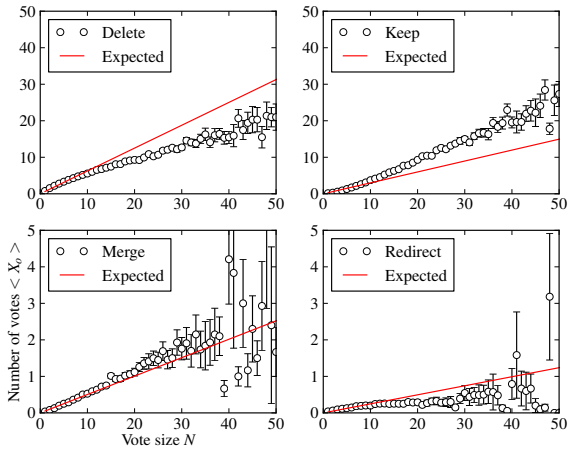


Fig. 5: Expected number of votes as a function of the vote size N. Each sub-plot displays the number of votes X_o of type o for votes of size N (circles). Red solid lines are the linear fit with an IID model with baseline probabilities. Error bars represent the standard error of the mean.

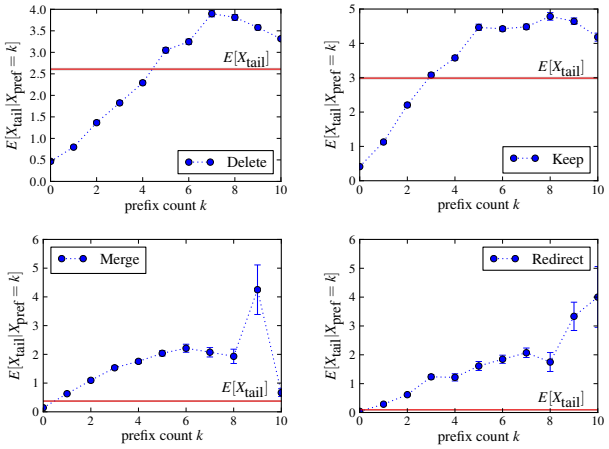


Fig. 6: Influence of initial votes over voting patterns. Blue circles: expected number of counts in sequence tail as a function of the number of preferences in the prefix. Horizontal red line and gray area: baseline count $E[X_{\text{tail}}]$ (red) and standard error bands (gray area).

for $o = K, D, R$ the fit is clearly not in agreement with the empirical data, hence suggesting that a bias is present in the votes.

Indeed, participants in an AfD discussion might be influenced by the level of consensus they perceive at the time and sequential position in which they arrive. We can then ask whether an over- or under-expression of votes for a given option o in the initial prefix of the voting sequence (i.e. the first k votes cast in the AfD) will influence subsequent users to vote for o in the tail of the sequence (i.e. the sequence obtained by removing the prefix from the entire AfD sequence). More precisely, we can ask whether $E[X_{\text{tail}}|X_{\text{pref}} = k]$ depends on k or not, where X_{tail} is the number of votes of type o in the tail of the sequence, and X_{pref} the number of votes in the prefix. Figure 6 shows that that the number of votes in the prefix sequence has a strong influence on the outcome in the remaining part of the vote. A t-test indicates that all differences, with the exception of $k = 3$ for keeps and $k = 10$ for merges are statistically significant (for the latter, the sample consists of 3 sequences). This result conflicts with the findings discussed in [2] suggesting that Wikipedia admin nomination discussions show no evidence of herding as a function of the initial prefix.

Figure 6 is also interesting for a number of reasons. First, in the case of K and D, an over- or under-expression of K and D in the prefix results in an over- or under-expression in the tail. This does not appear to happen for M and R. This might be due to the fact that consensus might be easier to reach in cases where a page just needs to be merged or redirected. Given the very small frequency of these two kinds of votes, however, the differences we see might be then due to the natural tendency for these votes to cluster in AfDs for pages that really need to be merged or redirected.

Second, we observe that $k = 5$ deletes are needed in the prefix to obtain a significant over-representation of D in the tail, whereas only $k = 4$ are needed in the case of K. This asymmetry suggests that herding is more likely to occur in one direction than the other.

Third, for $k \leq 8$ less votes are expected for $k = 7$, i.e. the trend becomes negative. Although we do not have an explanation for this phenomenon we submit that it might be again related to a reaction effect (*overturning vote behaviour*) from the opposing faction in controversial discussions.

We also checked whether cascades are more likely to occur at a specific position in a voting sequence. Figure 7 show the size of the cascades, i.e. the length of the longest subsequence of votes of the same kind starting at a given point. As we expected, there is no privileged position where cascades can happen, since the trend is roughly constant and not very diverging from the one assuming purely random sequences (i.e. permutations of the voting sequences). For K and D, the only privileged point–in the sense of a starting position x in a sequence where longer cascades are observed–seems to be the initial point, i.e. $x = 0$. This can be explained by noting that this deviation happens also for the randomly permuted sequences. This would only happen if the random permutation was not able to effectively break the long cascades, i.e. when the sequence is made up almost entirely of votes of the same kind. But this is exactly what happens for non-controversial votes, so we think the discrepancy at $x = 0$ is not indicative of a real trend but just a byproduct of including short non-controversial AfD in the dataset.

C. Heterogeneity of user voting behaviour

Given the reported existence of different factions of users, namely the inclusionists and the deletionists, it is natural to

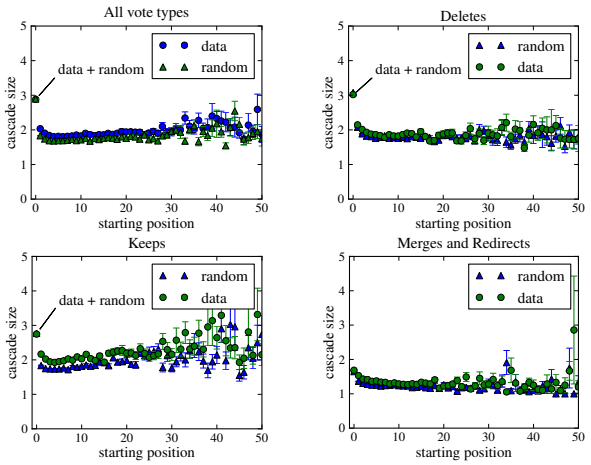


Fig. 7: Size of voting cascades as a function of initial position of the cascade.

ask whether this "global" distribution is good at describing individual users. If "organised factions" exist and affiliation to any of these groups influences an agent's voting tendency, then the baseline probability distribution $f_o, o \in \{K, D, R, M\}$ should perform poorly because of the heterogeneity of the voting tendencies of users. On the other hand, if we observed an overall homogeneity we should conclude that factions are just popular manifestos but do not really influence voting behaviour at a large scale. We can use statistical hypothesis testing to give a precise meaning to such question. For each user, we compute the χ^2 statistics and the associated two-tailed p-value for the hypothesis "the frequencies of vote of the user are taken from $P(o)$". Given the approximate nature of the χ^2 test statistic, we consider only users with more than 5 votes in this computation. This gives a sample of size $N = 13993$. For each significance level Q, we compute the fraction $R(Q)$ of users that attain a p-value greater than Q in the test. As noted by [14], this quantity is the inverse CDF of the χ^2 statistic, hence its expected value is $R(Q) = 1 - Q$.

Figure 8a shows the results of this analysis for all users. We observed that the baseline probability distribution does not perform very reliably at the individual level. The effects improve noticeably when we test specific subgroups. Figures 8b and 8c show the curve of $R(Q)$ computed for the distribution of voting frequencies on the restricted groups of all users i such that $f_D^{(i)} - f_K^{(i)} > \alpha$ or s.t. $f_K^{(i)} - f_D^{(i)} > \alpha$, respectively. The first group has 626 users while the other 157. The value for the threshold parameter we use is $\alpha = 0.1$. We experimented with other values and found they give qualitatively similar results.

In conclusion, these graphs show strong homogeneity at the subgroup level, and thus suggest that two different factions of users (whether publicly identifiable or not) exist and exhibit different voting patterns.

D. Strategic behaviour

In the previous sections we observed that distinct subgroups of users can be identified by similar voting patterns and that individual voters are strongly affected by the perceived state of the deliberation process at the time in which they cast their vote. In this final section we check whether the existence of these factions influences the overall voting dynamics in a systematic way–for example whether members of this factions act under some form of orchestration. One hypothesis could be that Deletionist users tend to open AfDs and, whenever a topic is controversial, Inclusionist users react by voting *en masse*. If this is the case, we would observe a clear non-stationarity in the temporal voting patterns.

In figure V-D we plot the probability $p_o(n)$ of voting for a given option o after $(n-1)$ votes have been already cast. We observe indeed that there is a strong non-stationarity in the case of K and D. D tend to be more likely to be cast in the initial part of the sequence but the probability of voting this option decreases rapidly until it balances with that for voting K. This behaviour could also be explained as a selection bias of voters [11], but given that users are also clustered in their voting behaviour, this picture is very significant in telling us that factions do have an influence in the dynamics of voting.

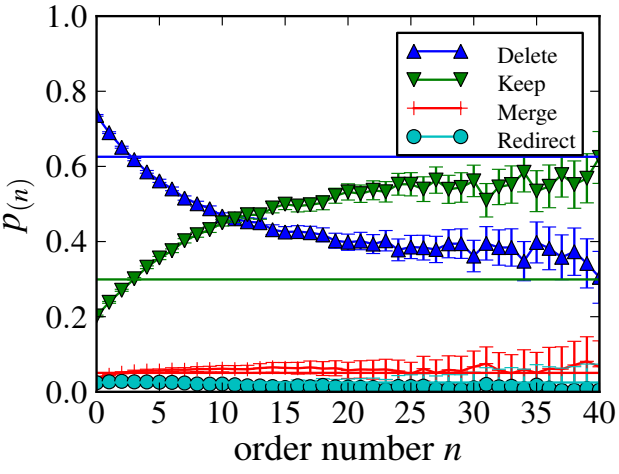
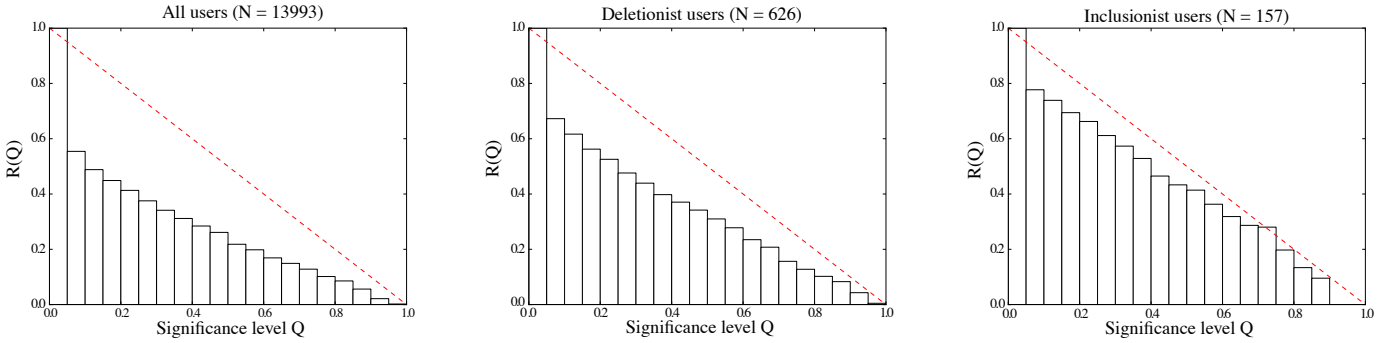


Fig. 9: Probability of voting for an option o as a function of the position n of the vote in the AfD sequence. Horizontal lines represent the baseline probabilities calculated on the complete sample

VI. Conclusions

The results presented in this study support the conclusion that, far from being exclusively driven by "notability" criteria and by an objective assessment on the merit of a topic, decisions about content inclusion in Wikipedia are strongly affected by a number of heterogeneous factors. The presence of biases due to the way in which AfD participants are allocated to discussions should not be necessarily regarded as a shortcoming of the system itself if there are reasons to believe that the current mechanism has some other benefits (e.g. scalability at the cost of accuracy). However, the empirical evidence emerging from our analysis supports the criticism that has been previously



(a) (b) (c)

Fig. 8: CDF of the χ^2 test statistic on the population of voters with more than 5 edits. (a): all users; (b) users s.t. $f_K^{(i)} - f_D^{(i)} > \alpha$; (c) $f_D^{(i)} - f_K^{(i)} > \alpha$ ($\alpha = 0.1$).

levelled against the deletion procedure currently in use in Wikipedia, which seems to lend itself to arbitrary deletion of content contributed by editors. We showed that a large share of AfD discussions show discrepancies with the official guidelines that govern such discussions. More interestingly, we shed light on the alleged role played by herding and strategic behaviour in determining the evolution of an AfD. Further research will need to clarify whether the evidence of frequent long series of votes expressing an identical option should be regarded as the result of psychological mechanisms (which may explain biases at the individual level) rather than strategic behaviour determined by organised movements (which may indicate a bias at the social level). Herding and strategic behaviour occupy extreme positions in a range of possible explanations, as the herding hypothesis assumes that what drives individual behaviour is merely the order and type of previously expressed votes, whereas the strategic behaviour hypothesis assumes that the only driver of individual decisions is a predisposition to vote consistently for a given option (e.g. keeping or deleting an article) regardless of the behaviour of other voters. Both hypotheses offer interesting empirical challenges to the idea that public collective deliberation in peer production systems may be secured against against biases of an individual and collective nature.

Acknowledgments

We would like to thank Wikipedia user *Betacommand* for providing us with the data on AfD votings; the participants in an informal survey on AfD mechanisms for their valuable feedback; the Wikimedia Foundation for granting us access to the archive of deleted revisions. GLC acknowledges the financial support of the Swiss National Science Foundation (grant no. 200020-125128). DT's work was partly supported by the FET programme of the European Commission through project QLectives (grant no.: 231200).

REFERENCES

[1] Y. Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, May 2006. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0300110561

[2] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Governance in social media: A case study of the wikipedia promotion process," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'10)*, 2010.

[3] I. Beschastnikh, T. Kriplean, and D. W. McDonald, "Wikipedian self-governance in action: Motivating the policy lens," in *Proceedings of the second ICWSM conference*, 2008.

[4] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Information quality work organization in Wikipedia," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 983–1001, 2008. [Online]. Available: http://dx.doi.org/10.1002/asi.20813

[5] S. K. Lam and J. Riedl, "Is Wikipedia growing a longer tail?" in *GROUP '09: Proceedings of the ACM 2009 international conference on Supporting group work*. New York, NY, USA: ACM, 2009, pp. 105–114. [Online]. Available: http://dx.doi.org/10.1145/1531674.1531690

[6] V. Kostakis, "Identifying and understanding the problems of Wikipedia's peer governance: The case of inclusionists versus deletionists," *First Monday*, vol. 15, no. 3, March 2010. [Online]. Available: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2613/2479

[7] A. Forte, V. Larco, and A. Bruckman, "Decentralization in wikipedia governance," *Journal of Management Information Systems*, vol. 26, no. 1, pp. 49–72, July 2009. [Online]. Available: http://dx.doi.org/10.2753/MIS0742-1222260103

[8] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie," in *ALT.CHI*, 2007.

[9] C. Roth, D. Taraborelli, and N. Gilbert, "Measuring wiki viability. An empirical assessment of the social dynamics of a large sample of wikis," in *WikiSym '08: Proceedings of the 4th International Symposium on Wikis*. New York, NY, USA: ACM, September 2008. [Online]. Available: http://nitens.org/docs/wikidyn.pdf

[10] D. Segal, "Look me up under N for Nobody; On Wikipedia, deletion looms for the patently non-notable," *The Washington Post*, p. D01, December 2006.

[11] F. Wu and B. A. Huberman, "Public discourse in the web does not exhibit group polarization," May 2008. [Online]. Available: http://arxiv.org/abs/0805.3537

[12] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *The Journal of Political Economy*, vol. 100, no. 5, pp. 992–1026, 1992. [Online]. Available: http://dx.doi.org/10.2307/2138632

[13] R. M. Raafat, N. Chater, and C. Frith, "Herding in humans," *Trends in Cognitive Sciences*, vol. 13, no. 10, pp. 420 – 428, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6VH9-4X6PPCY-1/2/38f26f1994570f7a58d587bb5a7a0569

[14] F. Radicchi, "Human activity in the web," Mar 2009. [Online]. Available: http://arxiv.org/abs/0903.2999