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
University of Lugano
Via G. Buffi 13, 6900 Lugano, CH
giovanni.luca.ciampaglia@usi.ch

***Abstract*—In this study we analyse the structure of a particular form of collective decision-making in Wikipedia, i.e. decisions regarding content inclusion and deletion. Wikipedia's official guidelines require that only topics that meet "notability" standards be included with a dedicated article. Decisions as to whether a topic is "notable" are made by groups of self-appointed reviewers, who assess the alleged encyclopaedic nature of a topic via so called *Article for Deletion* discussions. We analyse the structure and dynamics of these discussions in order to identify possible biases affecting their outcome. We show in particular the effects of voter heterogeneity and herding behaviour on the functioning of these collective deliberation processes.**

I. Introduction

The governance of peer production systems depends on participatory processes that usually involve large numbers of users [1]. Wikipedia is an example of such systems in which deliberative mechanisms are designed to outsource decisions to a population of contributors. Over time, the Wikipedia community has become more and more involved in governance and oversight tasks ([2], [3]). Several maintenance routines, such as suggesting that an article be considered for deletion, can now be directly initiated and run by regular editors. Maintenance tasks in peer production systems are typically self-assigned and contributors can decide to participate in a variety of routines as they see fit. The very strength of a peer production system (its decentralised governance) can also be seen as a possible source of biases and suboptimal solutions, e.g. the allocation of inadequate resources to address specific kinds of task.

The decentralised governance of peer production systems has attracted a growing attention in recent literature (see [4], [5]). Participation in discussions on information quality standards and their enforcement in Wikipedia was first addressed in [6], who found that user participation in information quality decision 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, the authors speculate, may have never come across Wikipedia's quality-related policies and guidelines.

The issue of topic inclusion/exclusion in Wikipedia entries was further addressed by [7], who conducted an extensive analysis of the deletion log of articles in a 3-year period ending in December 2007. Their study focusses on deletion rates of articles in the English Wikipedia and suggests that deletions tend to happen early during the life-cycle of entries. The authors 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 quality standards adopted by the community. Finally, article popularity, measured on the basis of the number of views that an article receives within a given time frame, was compared with the probability of its being deleted and evidence was found that the probability of survival of an article broadly follows its popularity in terms of readership and ranking in search engines.

II. Mechanics of inclusion and deletion

The deliberative process behind user-driven deletion proposals is currently known as an "Article for Deletion" discussion (hereafter: AfD).[1,2] 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.[3]

Editors participating in the discussion can cast any of the following options:

- **Keep** (hereafter: K) to recommend that the article be kept;
- **Delete** (D), to 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

[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.

[3] It should be noted that the guidelines discourage the term "vote" to refer to the AfD procedure, which should be understood as "a means to gauge the degree of consensus reached so far". We use the term "vote" hereafter for the sake of simplicity to refer to options expressed in the context of an AfD, notwithstanding this caveat.

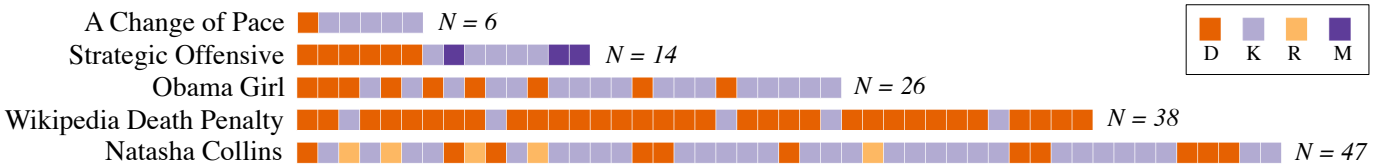

Fig. 1: A sample of 5 AfD sequences, including their length and breakdown by vote type.

sufficient degree of consensus has been reached, and enforce the corresponding decision.

The standard adopted in AfD discussions to decide whether a topic merits a dedicated article is the so-called "notability" of the topic.[4] The common feature of a notable topic, as defined by the guidelines, consists in its been "noticed" to a significant degree by reliable secondary sources. A topic that is deemed *non notable* is by definition unsuitable for inclusion with a dedicated article 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. [7] The interest in studying deletion discussions lies in the fact that what "notable" means if far from granted. The interpretation of "notability" criteria and their application is a matter of ongoing debate in the Wikipedia community. The existence of opposing groups of editors with strongly diverging opinions as to what "notable" means is witnessed by the existence of a long-standing public controversy around "Inclusionism"[5] and "Deletionism"[6], two opposing views on what Wikipedia should include. These views spawned two organised movements of Wikipedia contributors who identify themselves as "Inclusionists" and "Deletionists" and behave accordingly when participating in an AfD discussion.

III. Research questions

In this study we aim to analyse Article for Deletion discussions as a complex form of collective decision-making 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 *per se* may not be the sole reason determining the inclusion/exclusion of content in Wikipedia.

One possibility we consider is that votes expressed by early participants in the context of an AfD discussion may influence the behaviour of subsequent participants. It is also possible, on the other hand, that voting patterns are heavily influenced by the degree of polarisation in the community due to contrasting views on what "notability" means. As a result of these two factors, we expect to observe long series of votes of the same "color" in AfD discussions, as exemplified in Figure 1 for a sample of 5 AfD discussions taken from our dataset.

By focusing on discussions that display a variety of user responses, we aim to study whether there are factors that affect the dynamics of AfD discussion beyond the sheer assessment of a topic's notability. In particular, we intend to tackle the following research questions:

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

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

IV. Dataset

We collected and analysed data on a total of $223,209$ AfD discussions that took place in the period going from January 1999 to July 2010. The dataset includes $1,218,267$ unique votes cast by $68,998$ individual users. The dataset includes, for each vote, the title of the corresponding AfD, the user name of the voter and the option voted. We do not have any record for votes other than the four main options of an AfD and we did not extract the text of the comments added by AfD participants after the vote. In order to compute the timestamp of each vote, we cross-referenced this data with revision data from a recent Wikipedia database snapshot (March 12, 2010) using a simple heuristic: the time of the first revision by a user on the AfD discussion page is taken as the time she cast her preference. Since our dataset is more recent than the database snapshot, this procedure forces us to discard all data from AfD discussions posterior to the date of the snapshot. This leaves us with $948,309$ votes, cast by $57,219$ users in $198,083$ AfD discussions. In this filtered dataset, there are $8,361$ anonymous users (12.1%) that are identified only by the IP address and are responsible for $11,931$ votes.

o	D	K	M	R
f_o	0.6837	0.2543	0.0415	0.0204

TABLE I: Estimated baseline probabilities of AfD options.

Table I gives the baseline probability f_o for the four options $o \in \{K, D, M, R\}$ recorded in our data (i.e. the probability to draw a vote of a given kind among all votes in our dataset). We can see the vast majority of votes is *Deletes* (D), followed by *Keeps* (K), while votes for *Redirection* or *Merge* (M and R) make up a very small fraction of the total number of votes.

V. Results

A. Herding effects

A totally unbiased discussion would require voters' preferences to be homogeneous and that each voter is not influ-

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

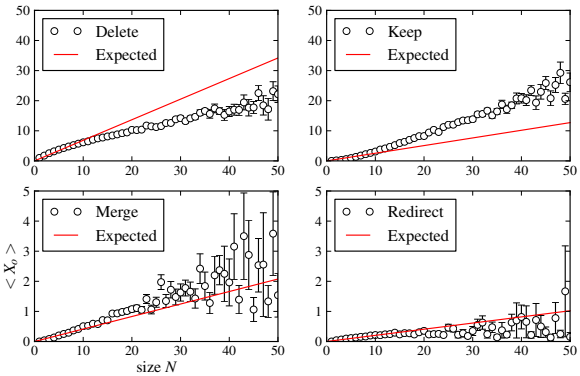

Fig. 2: Expected number of votes as a function of the vote size N. Each plot displays the number of votes X_o of type o for votes of size N). Solid lines are the linear fit with an IID model with baseline probabilities. Standard errors are included.

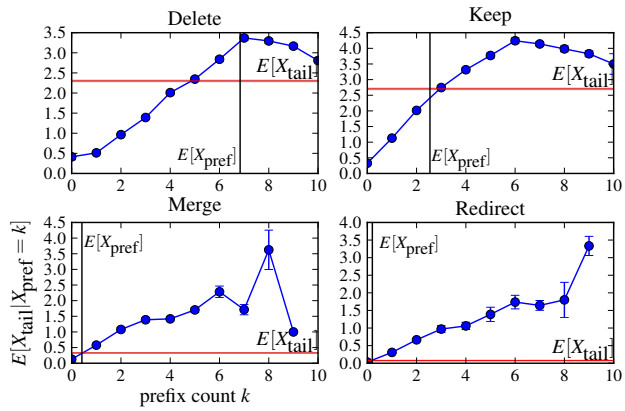

Fig. 3: Influence of initial votes on voting patterns. Expected number of votes in the AfD tail as a function of the number of same options in the prefix (including standard errors). Horizontal red lines indicate: $E[X_{\text{tail}}]$, vertical lines: $E[X_{\text{pref}}]$

enced by the votes already cast by other participants. This is equivalent to say that votes form sequences of IID draws where the probability of choosing (i.e. voting) an option $o \in \{K, D, R, M\}$ is given by the estimated baseline probabilities f_o from Table I. This also means that the expected number of votes X_o that option o receives in a sequence of N votes should be equal to $f_o N$.

In Figure 2 we test this hypothesis by plotting the average number of votes each option receives as a function of the total number of votes N, together with the expected prediction of this simple IID model. The plot shows that there is some level of agreement until $N = 10$. Afterwards, at least for $o = K, D, R$, the fit is clearly not in agreement with the empirical data. One possible explanation for this observed behaviour may be due to the presence of information cascades of votes. An "information cascade", often referred to as "herding behaviour" (see [8]; for a review on herding in humans see [9]), occurs when people form beliefs and opinions on the basis of information obtained through the observation of the behaviour of others. These phenomena are called "cascades" when the opinions expressed by the agents who act first influence the opinions expressed by subsequent agents, which in turn influence later participants and the overall temporal sequence. We make the hypothesis that participants in an AfD discussion are influenced by the level of consensus they perceive at the time and sequential position in which they arrive. We can check this by asking whether an over- or under-expression of votes in the initial prefix of the voting sequence (i.e. the first k votes cast in the AfD) for a given option o will influence users to vote o accordingly 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 3 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, i.e. an over- or under-expression of preferences in the initial part results in an over- or under-expression in the following. A t-test indicates that all differences, with the exception of $k = 3$ for "Keep", $k = 5$ for "Delete" and $k = 9$ for "Merge", are statistically significant (for the latter, the sample consists of 1 sequences and there were no sequences having 10 "Merge" or "Redirect" votes in their prefixes). Even though the number of votes is not a direct proxy for the likelihood of an article's surviving an AfD nomination, this result should be compared with the figures on outcomes of oppose/support votes studied in [4].

Figure 3 is interesting for a number of further reasons. First, the value of k past which a surplus of votes in the tail is observed is roughly equivalent to $E[X_{\text{pref}}]$ for $o \in \{K, M, R\}$ (this is harder to appreciate for M and R in the plots because the expected number in 10 votes is < 1 in for these options). Second, for $k \geq 7$ (*Deletes*) and $k \geq 6$ (*Keeps*) less votes are expected than for lower values of k, i.e. the trend becomes negative. Although we do not have an explanation for this phenomenon we submit that it might be due to a reaction effect (*overturning vote behaviour*) in particularly controversial discussions.

B. Heterogeneity of voter behaviour

Given the existence of opposing views on "notability" and of organised movements of Wikipedia contributors endorsing these views (i.e. Inclusionists vs. Deletionists), it is natural to ask whether affiliation to any of these groups influences the voting tendencies of individual users. If this is the case, the baseline probability distribution f_o should perform poorly at describing voting patterns of each individual. Conversely, if we observed an overall homogeneity we should conclude that these movements are just popular manifestos but do not substantially 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

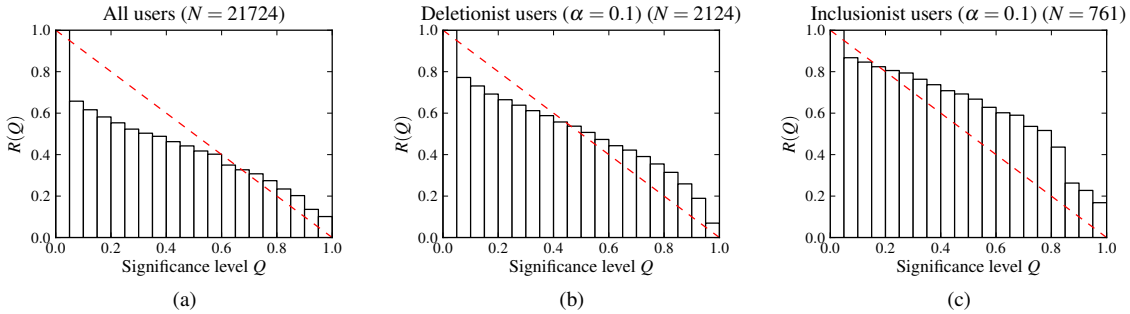

Fig. 4: CDF of the χ^2 test on voters with more than 5 edits. (a): all users; (b) users s.t. $f_K^{(i)} - f_D^{(i)} > \alpha$; (c) $f_D^{(i)} - f_K^{(i)} > \alpha$.

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 us a sample of size $N = 21,724$. 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 [10], this quantity is the inverse CDF of the χ^2 statistic, hence its expected value is $R(Q) = 1 - Q$.

Figure 4a shows the results of this analysis for all users in the sample. We observe that the baseline probability distribution does not perform reliably at the individual level. The reliability improves noticeably, instead, when we test subgroups that are compatible with the "two factions" hypothesis. Figures 4b and 4c show the curve of $R(Q)$ computed with the distribution of voting frequencies of two such groups: a) all users i s. t. $f_D^{(i)} - f_K^{(i)} > \alpha$ (i.e. having a Deletionist tendency) or b) all users s.t. $f_K^{(i)} - f_D^{(i)} > \alpha$ (i.e. Inclusionists). The first group has $2,124$ users while the other 761. The value for the threshold parameter we use is $\alpha = 0.1$. We experimented with other values and found qualitatively similar results.

In conclusion, this analysis shows strong homogeneity at the subgroup level, and suggests that two different factions of users (whether publicly identifiable or not) exist and exhibit heterogeneous voting patterns.

VI. Conclusions

The results presented in this study support the conclusion that, when deliberating about the "notability" of a topic, collective decisions are affected by a number of extrinsic 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 is consistent with the shortcomings that many have identified in the deletion procedure currently in use in Wikipedia. Further research will need to clarify whether the presence of frequent long series of votes expressing an identical option should be regarded as evidence of the effect of psychological mechanisms (which may explain biases at the individual level) rather than strategic behaviour determined by organised movements, such as voter recruitment (which may indicate a bias of a social nature). Simpler behavioural explanations such as a selection bias should also be considered as a possible way to account for these observations [11].

Acknowledgments

We would like to thank Wikipedia user *Betacommand* for providing us with the data on AfD discussions. 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] R. M. Raafat, N. Chater, and C. Frith, "Herding in humans," *Trends in Cognitive Sciences*, vol. 13, no. 10, pp. 420 – 428, 2009.

[10] F. Radicchi, "Human activity in the web," *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, vol. 80, no. 2, p. 026118, Aug 2009.

[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