

RESEARCH ARTICLE

Understanding and coping with extremism in an online collaborative environment: A data-driven modeling

Csilla Rudas¹, Olivér Surányi², Taha Yasseri^{3,4}, János Török^{1,2*}

1 Institute of Physics, Budapest University of Technology and Economics, Budapest, Hungary, **2** Institute of Physics, Eötvös Loránd University, Budapest, Hungary, **3** Oxford Internet Institute, University of Oxford, Oxford, United Kingdom, **4** Alan Turing Institute, London, United Kingdom

* torok@phy.bme.hu



Abstract

The Internet has provided us with great opportunities for large scale collaborative public good projects. Wikipedia is a predominant example of such projects where conflicts emerge and get resolved through bottom-up mechanisms leading to the emergence of the largest encyclopedia in human history. Disaccord arises whenever editors with different opinions try to produce an article reflecting a consensual view. The debates are mainly heated by editors with extreme views. Using a model of common value production, we show that the consensus can only be reached if groups with extreme views can actively take part in the discussion and if their views are also represented in the common outcome, at least temporarily. We show that banning problematic editors mostly hinders the consensus as it delays discussion and thus the whole consensus building process. To validate the model, relevant quantities are measured both in simulations and Wikipedia, which show satisfactory agreement. We also consider the role of direct communication between editors both in the model and in Wikipedia data (by analyzing the Wikipedia *talk* pages). While the model suggests that in certain conditions there is an optimal rate of “talking” vs “editing”, it correctly predicts that in the current settings of Wikipedia, more activity in talk pages is associated with more controversy.

OPEN ACCESS

Citation: Rudas C, Surányi O, Yasseri T, Török J (2017) Understanding and coping with extremism in an online collaborative environment: A data-driven modeling. PLoS ONE 12(3): e0173561. <https://doi.org/10.1371/journal.pone.0173561>

Editor: Alain Barrat, Centre de physique theorique, FRANCE

Received: July 27, 2016

Accepted: November 24, 2016

Published: March 21, 2017

Copyright: © 2017 Rudas et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Wikipedia data: <http://dumps.wikimedia.org/> <http://wwm.phy.bme.hu/>.

Funding: TY has received funding from the European Union’s Horizon2020 research and innovation program under grant agreement No 645043; "HUMANE: a typology, method and roadmap for HUMAN-MACHINE NETWORKS".

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Large scale collaboration has been a central concept in development of both the Internet and the WWW [1–3]. With the ever increasing penetration of the information communication technologies across the globe, and the emergence of the user generated web (sometimes called “Web 2.0”), collaboration of individuals from all around the world to generate public good products is more ubiquitous than ever. A wide range of platforms and protocols facilitate such collaborations between humans and machines at different scales and with different goals [4], e.g., Wikipedia, sourceforge, github.

However, in the more and more globalized world of such social systems, conflicts may arise due to opinion differences. This is even more important in systems where common value production is the goal of the service. In most of the above mentioned systems, due to the bottom-up management in place [5], the conventional tools of conflict resolution are inapplicable. Hence, it is astonishing that in spite of the magnitude of the opinion differences in the world, even in sensitive issues, quality articles are produced on Wikipedia, comparable to the ones in the expert-written encyclopedias [6]. Therefore, the remaining big puzzle about Wikipedia is that “it only works in practice, in theory, it can never work”.

The approach of complex systems science has become more and more relevant to study collective social behavior. The availability of large scale data on our personal and societal activities has transformed the methods and scopes of social sciences considerably, leading to the emergence of the new field of computational social science [7, 8]. In this paper, we take such an approach and use agent-based modeling to shed light on some aspects of Wikipedia opinion and content dynamics.

Wikipedia has been studied by various researchers and from different angles. When it comes to conflicts of Wikipedia, a good amount of research has been focused on vandalism and how to detect it [9–11]. Even though vandalism is very much related to opinion clashes among users, here we are more interested in conflicts between editors who have faith in the whole project and have no negative incentives. Such cases have been studied empirically by various groups. The bursty nature of editorial wars and the separation between peace and war phases in a dynamical framework, are studied in [12] and [13]. More detailed analysis on dyadic interactions between editors and the role of social status are presented in [14]. Wikipedia articles have been ranked based on their controversy scores and the controversial topics have been analyzed in [15–17]. And finally tools for visualizing and detecting Wikipedia conflicts are developed in [18]. However, most of the empirical work on Wikipedia conflicts fail to explain the mechanistic scenarios driving the emergence and resolution of conflicts among editors.

One of the under-researched aspects of Wikipedia edit wars is the role of the “talk pages”. Talk pages are forums in which editors can discuss their opinions on the content of the article and try to reach a consensus before overriding each other’s edits directly on the article [19]. Even though it has been argued that the presence of such facilities would hinder the emergence of edit wars [20], there is little theoretical work to explain this observation. In other related work, the content of the talk pages is analysed using natural language processing tools to explain their functionality, however again, not much of mechanistic modeling is provided [21–23].

Modeling opinion dynamics in an agent-based framework has an extensive literature (for a review see [24]). A successful class of such models are known as “bounded confidence” models, which allow agents to accept opinion alterations within a tolerance threshold [25]. In more recent work, bounded confidence models are generalized to account for emotion dynamics parallel to opinion dynamics [26]. Different directions of generalization of such models have been taken to explain the user dynamics and opinion dynamics on Wikipedia [27, 28]. We make use of one of this generalizations which accounts for the common product (the article) among the agents as well as the indirect interactions between agents through this common product [29].

Previous work has shown that in many cases, a consensus can be reached even if the original pool of opinions was very mixed. It is clear that the major problem in building a consensus is the presence of opinion outliers who have views very different from the majority. We devote this paper to the study of this question, that in what extent are people with extreme opinion impedimental in consensus building and what measures can be applied to decrease the chance

of a frozen conflict. We also implement the process of banning editors who create conflict, a well known procedure in Wikipedia and show the effect of it on the evolution of the conflict. The results of the model are compared with empirical data generated based on Wikipedia activity logs.

2 Methods

In this paper we combine the editorial activity data collected from Wikipedia with a generalized version of the computational model of opinion dynamics that we have developed earlier [27, 29].

2.1 Data

The data collection is carried out using Wikimedia Tool Labs <https://tools.wmflabs.org>, which provide live access to the Wikipedia database containing logs of all the editorial activities. For more details on data collection see [30]. Data Dumps that we generated and used in this work are available at <http://wmm.phy.bme.hu/>.

In obtaining statistics of the editors' activities, we explicitly excluded Wikibots (semi-automated computer codes that carry out large scale simple tasks, e.g., correcting typos or creating inter-language links). In order to do so, we excluded all the users with "bot flag" (an identifier that has to be used by bot-runners to distinguish between bot and human edits). We collected data from different language editions. These editions represent a large range of language editions in size and number of articles, as well as large variety in their local rules and conventions. However, as reported below, most of the observed statistical features are language independent. In particular we selected 13 language editions to have diversity in size, rules, and cultures. The current statistics of these language editions are reported in Table 1.

We counted the number of edits to article pages and Wikipedia talk pages and compared this ratio to its equivalent in the model. To be able to study the consensus reaching process, we consider reverts: edits that undo a previous edit. For more details see [31, 32].

2.2 Computational model

We use the model that we introduced in [27] and further developed in [29]. We consider the case with fixed agent pool: N agents try to edit and eventually agree on a descriptive article

Table 1. Overall statistics of the language editions under study.

Language	code	Articles	Edits	Users
English	en	5 M	859 M	29.5 M
German	de	2 M	164 M	2.5 M
French	fr	1.8 M	134 k	2.6 M
Spanish	es	1.3 M	94 M	4.4 M
Portuguese	pt	945 k	48 M	1.8 M
Persian	fa	513 k	23 M	607 k
Arabic	ar	449 k	24 M	1.2 M
Hebrew	he	198 k	20 M	318 k
Hungarian	hu	398 k	18 M	331 k
Romanian	ro	372 k	11 M	386 k
Czech	cz	367 k	14 M	351 k
Japanese	ja	1 M	63 M	1.1 M
Chinese	zh	912 k	43 M	2 M

<https://doi.org/10.1371/journal.pone.0173561.t001>

about a subject. Each agent has an opinion about the article which is represented by a time-varying scalar variable in the range $x_i(t) \in [0, 1]$. The article can also be biased towards an opinion value at any time represented on the same scalar interval $A(t) \in [0, 1]$. At each time-step, a randomly chosen agent with probability r tries to communicate with another randomly chosen agent or otherwise (with probability $1-r$) tries to edit the article. The model for agent-agent interaction is known as the Deffuant model for opinion dynamics in mixed populations [25]. We use a specific case of this model described below.

Two agents can only communicate if their opinion differs less than ϵ_T , the tolerance parameter of the agent-agent interaction, in which case they modify their view on the subject and both adapt a joint opinion half way between their original ones:

$$(x_i, x_j) \rightarrow \begin{cases} ([x_i + x_j]/2, [x_i + x_j]/2) & \text{if } |x_i - x_j| < \epsilon_T \\ (x_i, x_j) & \text{otherwise} \end{cases} \quad (1)$$

The original opinion models [25] don't take into consideration the article editing, the only interaction is the agent-agent communication, which drives the system into stable configuration characterized by opinion groups which do not interact with each other, and the average number of which is determined by ϵ_T .

Agents have different tolerance (ϵ_A) towards the opinion reflected in the article. If they try an editing action and find that the position of the article differs less than ϵ_A from their own opinion, they do not change it, instead adapt their opinion towards it by an amount proportional to a convergence parameter μ_A and the opinion difference. In the opposite case, when the article is intolerable for the agent, it will modify it accordingly:

$$(x_i, A) \rightarrow \begin{cases} (x_i + [A - x_i]\mu_A, A) & \text{if } |x_i - A| < \epsilon_A \\ (x_i, A + [x_i - A]\mu_A) & \text{otherwise} \end{cases} \quad (2)$$

The simulation procedure is thus the following: The system is prepared first by running the agent-agent interaction Eq (1) to create the opinion groups, then each Monte-Carlo step is composed of N actions in which a randomly chosen agent i either talks to another agent or interacts with the article, depending on the parameter r as described above. The relaxation time, in general, is defined as the average number of Monte-Carlo steps needed to reach a consensus where all agents are satisfied with the article, i.e. they are in the tolerance range of the article described by ϵ_A . To cope with the enormous statistical fluctuations which prevent the calculation of a sensible mean, we measured τ by the position of the maximum of the relaxation time density function, which is equivalent to the most probable relaxation time.

If the simulations are started with random initial agent opinion distributions then the number of opinion groups will vary due the intrinsic randomness in the model. The relaxation time measured for a specific set of parameters is the average which includes qualitatively different scenarios of different number of opinion groups. This phenomenon illustrated in Fig 1 where the relaxation time distribution is shown for simulations with exactly the same parameters but with different random seeds creating different initial opinions $x_i(t = 0)$. The distribution of the relaxation time is plotted separately for different number of initial opinion groups (n). Interestingly if initially there are four opinion groups ($n = 4$) then the system behaves similar to either a 3 or a 5 group setup. We categorize the agents by *mainstream* and *extremists* depending on their opinion with respect to the others.

To avoid the above artifacts emerging from this sensitivity, we fix the initial conditions throughout the analysis as follows:

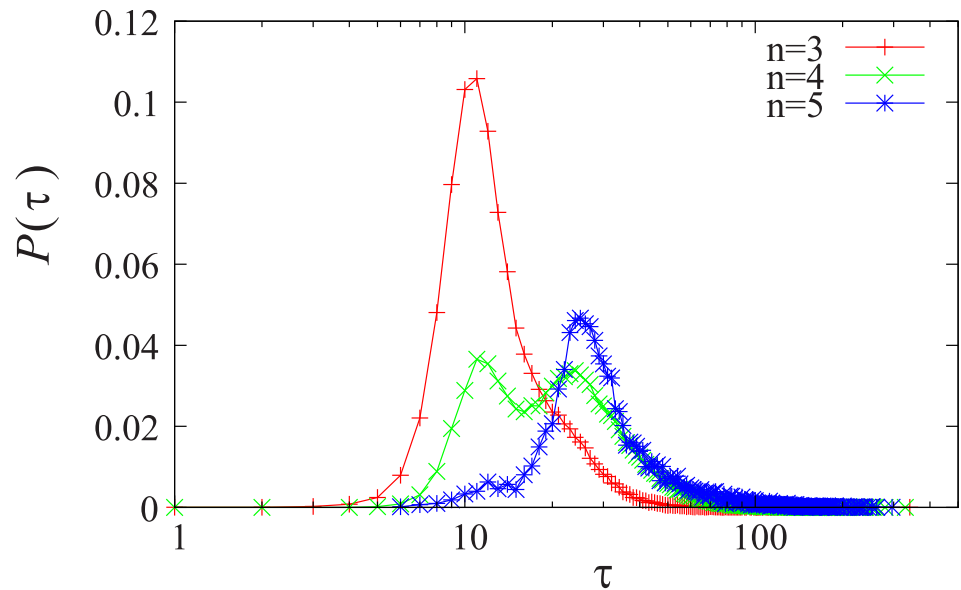


Fig 1. Distribution of relaxation time for different number of initial opinion groups for $\mu_A = 0.65$ and $\epsilon_A = 0.25$.

<https://doi.org/10.1371/journal.pone.0173561.g001>

- **2 groups:** 2 groups at opinions 0.0 and 1.0.
- **3 groups:** 1 mainstream group at 0.5 and 2 extremists at 0.1 and 0.9.
- **4 groups:** 2 mainstream groups at 0.25 and 0.75, and 2 extremist groups at 0.0 and 1.0.

A new parameter *RoE* is introduced as the ratio of the agents in the extremist groups. Naturally it is relevant only for 3 and 4 groups.

It was shown in [27] that the above defined model has three different modes of convergence which can be identified by regions in the phase diagram of (μ_A, ϵ_A) . We reiterate here the main findings: Regime I was observed for low values of μ_A, ϵ_A and was characterized by astronomical relaxation time (which prohibits its study for a reasonable system of size $N > 100$) and an ever lasting stable conflict in which a large mainstream group fights an endless war against two small extremist groups.

Therefore, we omit the study of this regime here. Regime II was characterized by an oscillatory behavior of the opinion of the article and the convergence was reasonably fast. Regime III showed the behavior most similar to Wikipedia, with very volatile article behavior, and in parallel, extremists gradually converted to the mainstream opinion. In this study we will focus on Regimes II and III and use the following parameter values for the simulations:

- **regime II:** $\mu_A = 0.45, \epsilon_A = 0.075$
- **regime III:** $\mu_A = 0.7, \epsilon_A = 0.15$

Regime II is characterized by a very controversial topic (ϵ_A small) and a moderately volatile article (μ_A intermediate), while Regime III is less controversial (ϵ_A is larger) and more volatile (μ_A is large). We also note here, that if r is not too small then the opinion groups stay compact during the simulations [27].

3 Results and discussion

3.1 Role of the extremists

In order to understand the dynamics of systems with multiple groups, we start with a two group scenario. The convergence of a two group system is always fast (see Fig 2) and can be understood by the following reasoning: In the case of two groups, the article can only be found between the two groups. If an agent from either sides is chosen to edit the article, two scenarios may occur. Either the article is outside of the tolerance of the agent, when the agent will pull the article towards it's group. In the other case the agent will move towards the article, i.e. towards the center of the opinion pool. The inter-agent talks keep the opinion groups together thus if any member moves, the whole group follows it though the distance the group makes will be n_g times smaller (n_g being the number of agents in the group). Thus the article makes a random walk between the groups while the groups gradually shift towards each other.

However, the article does not make a regular random walk since the step size depends on its position relative to the groups: the farther it is from the group the larger the step size it takes in the direction of it. Therefore the random walk of the article is biased towards the center.

The number of steps the article can make from the center at x_c to the group at opinion x_e can be calculated as:

$$n_s(x) \simeq \int_{\epsilon_A}^x \frac{1}{\mu_A y} dy = \frac{1}{\mu_A} \log(x/\epsilon_A), \tag{3}$$

where $x = |x_e - x_c|$ is the distance of the two groups (the number of steps is the integer part of $n_s(x)$ but we approximate it with a continuous variable).

If the probability of choosing an agent from this group with respect to the other group is p_g ($p_g = 1/2$ for the two group case but the general scenario will be useful later) then the article is within the tolerance level of the article if at least n_s number of steps were made in the

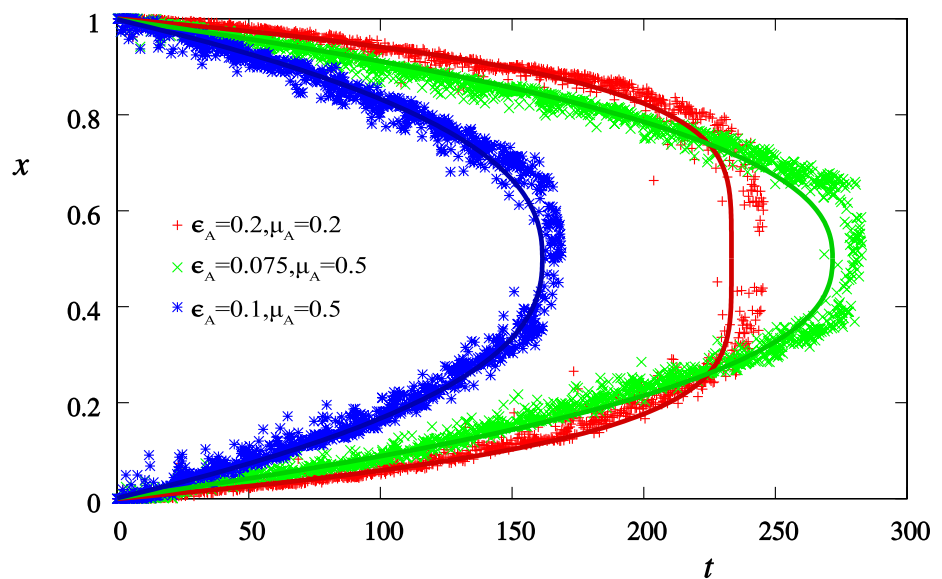


Fig 2. The time evolution of the two opinion groups for three sets of parameters: (a) $\epsilon_A = 0.2, \mu_A = 0.2$, (b) $\epsilon_A = 0.075, \mu_A = 0.5$, (c) $\epsilon_A = 0.1, \mu_A = 0.5$. The points are simulation results, solid lines are result of Eq (6).

<https://doi.org/10.1371/journal.pone.0173561.g002>

direction of the group. The time the article spends in the vicinity of the extreme group is

$$\tilde{t}_g \propto p_g^{n_g}. \tag{4}$$

The velocity of a group is inversely proportional to the n_g number of agents it contains, so:

$$\frac{1}{n_g} v_g(x) \propto \frac{\tilde{t}_g}{n_g} = \frac{1}{n_g} p_g^{\log(x/\epsilon_A)/\mu_A}. \tag{5}$$

Since we only know $v_g(x)$, we will integrate the inverse of the latter to get the time as function of the group position

$$t = \int_0^x p_g^{\log(x'/\epsilon_A)/\mu_A} dx'. \tag{6}$$

The integral can be calculated for fixed values of μ_A and ϵ_A . Fig 2 compares the analytical result with the simulation data. The result of Eq (6) fits the numerical results well and gets bad only at the very end where the number of steps the article can make is small and neglecting the integer part makes an important error.

In the following, we continue with the analysis of the cases with 3 and 4 groups where we can test the effect of the extremists RoE and of the probability of the communication action r .

Fig 3 summarizes the relaxation time as a function of the two parameters: RoE and r . The results are surprising: In case of three groups, the convergence is always faster if the ratio of the extremists is higher. This unintuitive result can be understood by looking at specific examples in Fig 4. The high ratio extremist case behaves essentially as a two group system while on the other hand if the middle group is strong we obtain the converging extremes scenario [27] with much larger relaxation time. The transition is smooth and the minimal relaxation time is observed when there is no mainstream group. This may seem counterintuitive but in our model the big mainstream group acts as conflict stabilizer. This issue will be taken up again in the Discussion.

Different behavior can be observed in the case of four opinion groups where, as shown in Fig 3, there is an optimum for the relaxation time in function of RoE which is just after a sharp transition at around $RoE = 0.35$. The examples in Fig 5 illustrate the reason behind this transition. If the middle groups are more numerous than the extreme groups, they merge fast as a two group system and the system is converted to a big middle group and two small extremist groups scenario, which is very stable with long relaxation time. On the other hand, if there are more extremists, they first merge with the middle group on their side, thus reducing the four group system to a two group system which converges fast. The difference in the relaxation times between the two cases is of orders of magnitude as indicated by the logarithmic scale in Fig 3. If the mainstream groups merge with the respective extremist, then increasing the size of the extremists will prolong the debate (see Fig 3) because the merged groups will be farther from the middle opinion.

We identify RoE_c the transition point between the above two scenarios when all four groups merge together simultaneously. We will consider the movement of a mainstream group as the result of independent two group scenarios. Unfortunately, the integral in Eq (6) is impossible to evaluate for the general case, therefore we will evaluate only the initial speed of the group. Since according to Eq (5) the smaller groups accelerate faster, the transition point should be between two limiting cases: (i) the middle groups start with 0 velocity, and (ii) the middle groups start with third of the extremist group velocities. These two conditions give us two

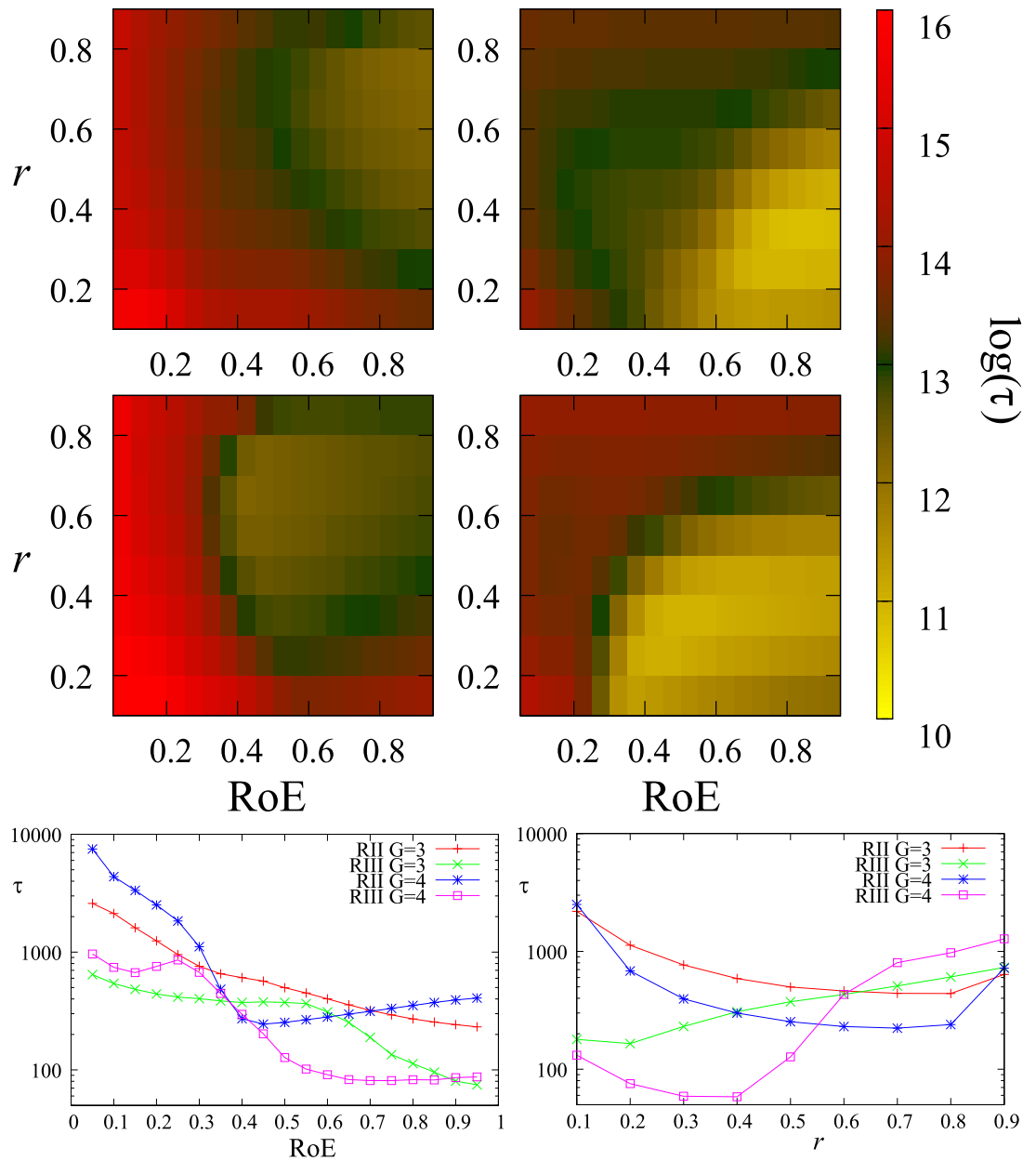


Fig 3. Colormaps of the logarithm of the relaxation time τ in function of RoE and r using simulations with $N = 10000$. Top row 3, bottom row 4 groups, left regime II, and right regime III. The lower two graphs show cuts at $r = 0.5$ and $RoE = 0.5$ respectively. Let us note the log-linear scale and the relaxation time sometimes grows an order of magnitude within small changes of RoE .

<https://doi.org/10.1371/journal.pone.0173561.g003>

implicit equations which can be solved numerically. The two equations are:

$$0 = (1 - RoE)v_{1/2}(0.5) - RoE v_{RoE}(0.25) + RoE v_{RoE}(0.75) \tag{7}$$

$$\frac{1}{2} v_{1/2}(0.5) = 2 \left[(1 - RoE)v_{1/2}(0.5) - RoE v_{RoE}(0.25) \right] \tag{8}$$

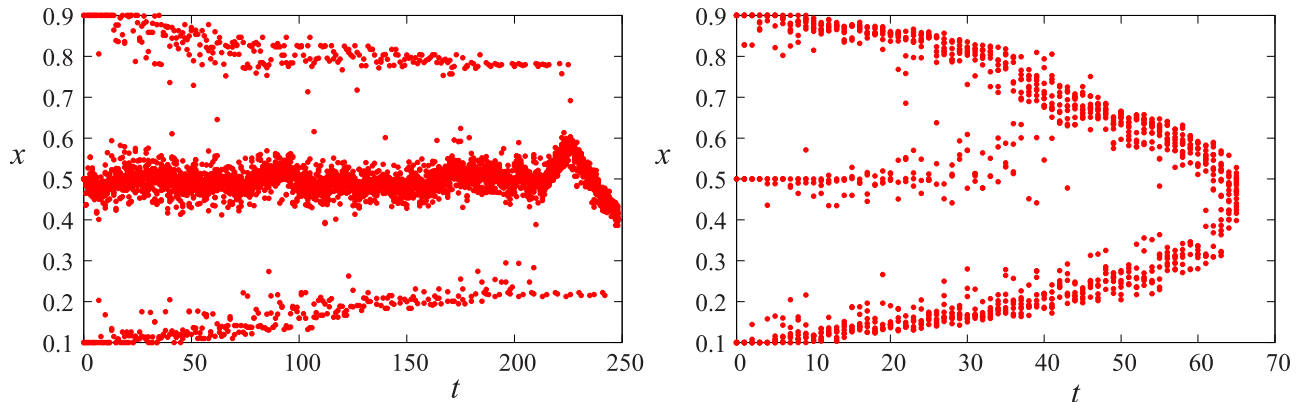


Fig 4. Example time evolution for three initial groups. $N = 1000$, $r = 0.5$, $\epsilon_A = 0.15$, $\mu_A = 0.7$, left: $RoE = 0.5$, right: $RoE = 0.9$.

<https://doi.org/10.1371/journal.pone.0173561.g004>

The numerical solutions of the above equation give us the range $0.37 < RoE_c < 0.43$, which corresponds well to the transition point from the relaxation times (see Fig 3).

Thus, if the mainstream views are divided between two options, as in many cases normally, then there is an optimal ratio of the extremists. Of course no extremists would be the best case but as soon as there are few of them they stabilize the conflict. There comes another sharp transition when they get numerous enough that the extremist groups merge with the respective mainstream leading, thus, to a two group system. The key difference here is that the extremists take an active part in the debate.

The talk to edit ratio r also has an effect on the relaxation time. In regime II for both cases of 3 and 4 groups, there is a medium (~ 0.6) optimum talk/edit ratio leading to the shortest relaxation time. On the contrary, in regime III, low values of r are more favorable for short relaxation time which means a lot of editing and little discussion. In summary, in case of oscillations we need agent-agent discussion for fast convergence but in the converging extremes, the major element of the convergence process is the relatively volatile article and extremists adapting a more mainstream opinion due to a semi extremist article. For this we only need editing.

To be able to compare these results with the empirical data, we calculated two ratios: (i) the ratio of edits to talk and to Wikipedia pages and (ii) the ratio of reverts on the articles to

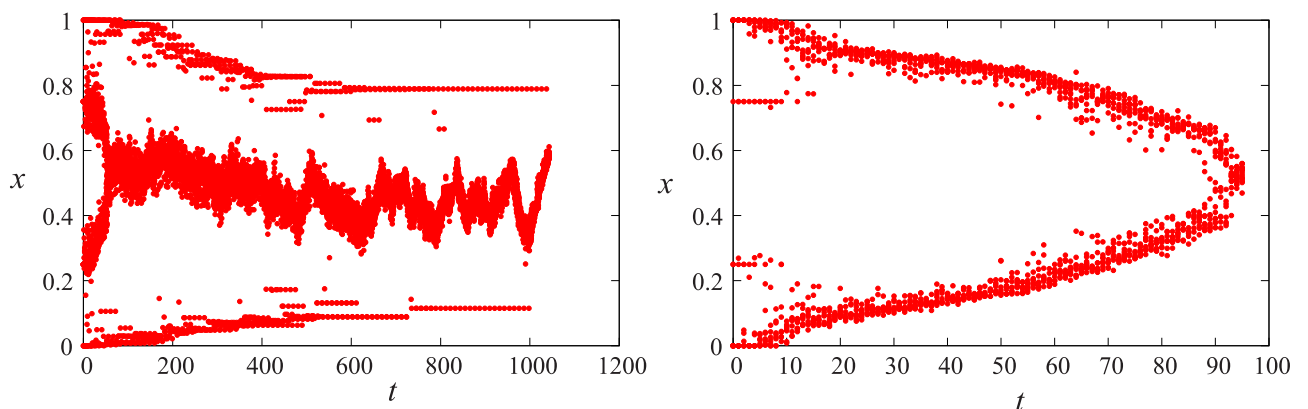


Fig 5. Example time evolution for four initial groups. $N = 1000$, $r = 0.5$, $\epsilon_A = 0.15$, $\mu_A = 0.7$, left: $RoE = 0.2$, right: $RoE = 0.5$.

<https://doi.org/10.1371/journal.pone.0173561.g005>

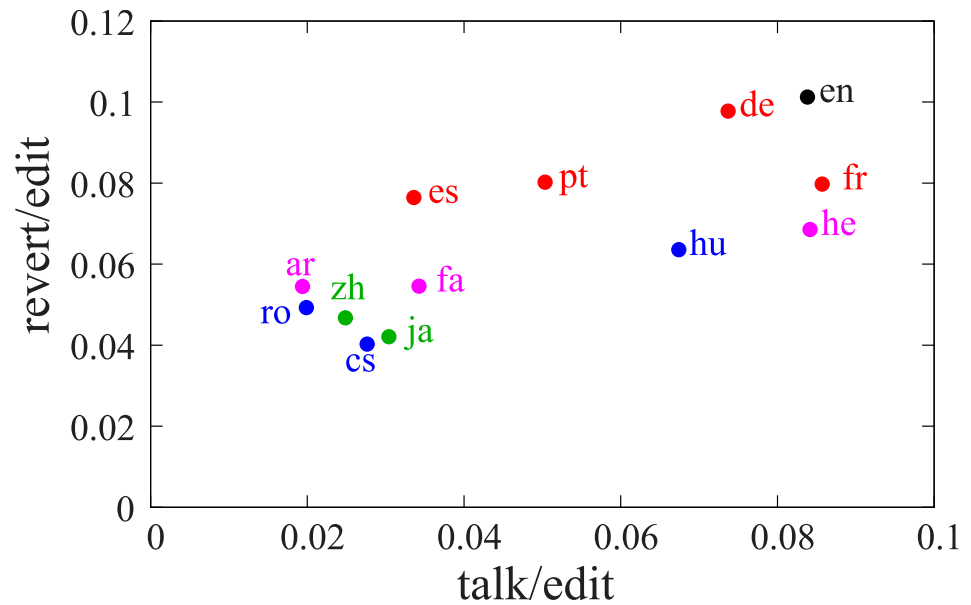


Fig 6. The revert/edit ratio (corresponding to r) vs. the ratio of edits to talk/article pages (corresponding to r) for 13 different Wikipedia language editions (en: English, West European (red): de: German, fr: French, es: Spanish, pt: Portugal, Eastern European (blue): cs: Czech, hu: Hungarian, ro: Romanian, Middle-East (pink): ar: Arabic, fa: Persian, he: Hebrew, Far-East (green): zh: Chinese, ja: Japanese.

<https://doi.org/10.1371/journal.pone.0173561.g006>

Wikipedia page edits for each language edition. One should keep in mind that this ratio is not exactly the same as in the model. In reality, edits to the articles are finer than edits to talk pages which generally add a larger part of the text to the discussion. Hence the talk/edit ratio in the model overestimates the same quantity measures in Wikipedia data. However, we expect to see the same trends, namely that with increasing talk/edit ratio the consensus reaching time should increase. As a proxy to the consensus reaching time, we measured the ratio of reverts to all the edits in each language editions. Fig 6 shows the relationship between these two ratios. We observe that in the language editions for which the talk/edit ratio is higher, there are generally more reverts. Similar relationship has also been reported for the article label [13]. This behavior is more similar to the case of 3 opinion groups in regime III where the optimal value of r is very small.

3.2 Banning of extremists

In Wikipedia, many different tools are used to control and eventually settle conflicts, for example freezing controversial articles, or banning users temporarily or permanently, who are not obeying community conventions. The aim of these measures is mainly to calm down editors, an aspect which is not included in our model. It has been shown that the editors who are banned more often, have a higher focus on smaller number of articles and they contribute directly to the editorial wars [30].

In our model, we have shown that oscillations have a key role in reaching consensus. Here we investigate how the introduction of banning can effect the consensus reaching process. Our expectation is that by banning agents, the fluctuations (e.g. number of active agents in extremist groups) speed up the oscillations and the relaxation time will be reduced.

We implement banning in our simulations as follows: If an agent in an edit process changes the article, it may get banned with a probability proportional to the square of the distance between the opinion of the agent and of the article. This reflects the fact that agents who find an article completely opposing their views are more likely to take action resulting in a banning. The probability of banning can be formulated as

$$p = \begin{cases} (A - x_i)^2 & \text{if } |A - x_i| > \epsilon_A \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

Banned agents, when selected for editing action will do nothing but get back their normal status and later will be able to edit the article. Agents are thus banned from one editing action while they still participate in the talking. In the simulation, the following agent opinion distribution was used ($RoE = 0.5$):

- $N/4$ extremist agents with opinion $[0.05, 0.15]$
- $N/4$ extremist agents with opinion $[0.85, 0.95]$
- $N/2$ mainstream agents with opinion $[0.45, 0.55]$

We can assume that as soon as the middle group and the article get in the vicinity of an extremist group, the members of the opposite extremist group get banned more easily, enhancing thus the oscillatory process of the article. To test this, we measure the relaxation time (τ_b) and compare it to the case without banning (τ).

Fig 7 shows the logarithm of the relaxation time of the model without banning, and the ratio of the relaxation times with/ without banning (normal scale). In the presence of banning, in regime III, patches of relaxation time decrease can be observed, while a more pronounced $\sim 10\%$ increase is observed in regimes I and II. Moreover, the increase of the relaxation time is larger where the relaxation time of the original model was already large. Therefore, it seems that this type of banning may only help to make consensus reaching faster when it was already fast without it. A possible explanation why the relaxation time increases could occur, is that when agents are banned, the debate is just delayed, because with banning it is less likely that a chosen agent is able to edit. So by banning agents, all we do is slow down the editing process. It

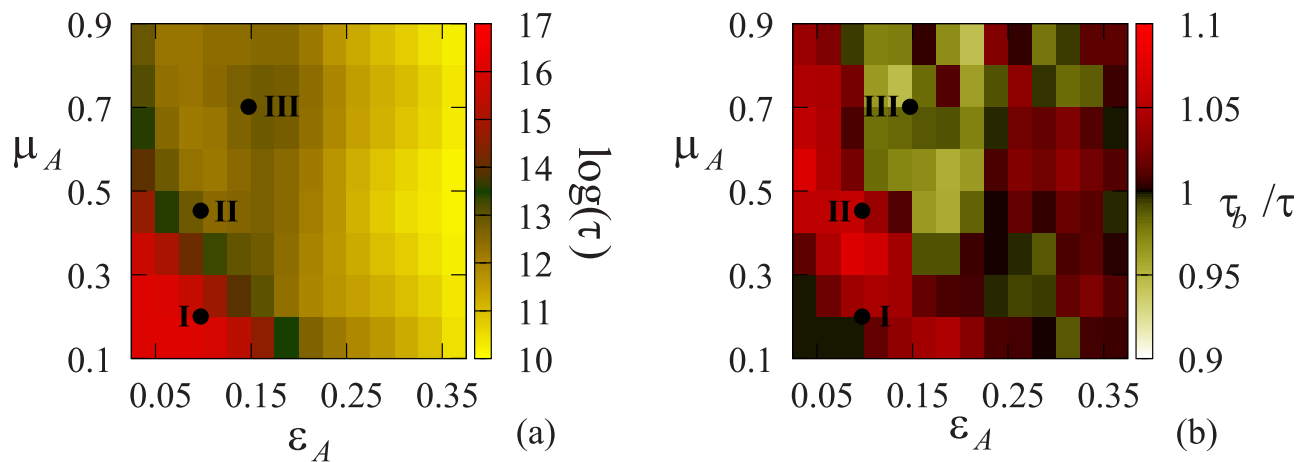


Fig 7. Simulation results: (a) The logarithm of the relaxation time of the original model with three opinion groups and $RoE = 0.5$, (b) The ratio of the relaxation times with and without banning.

<https://doi.org/10.1371/journal.pone.0173561.g007>

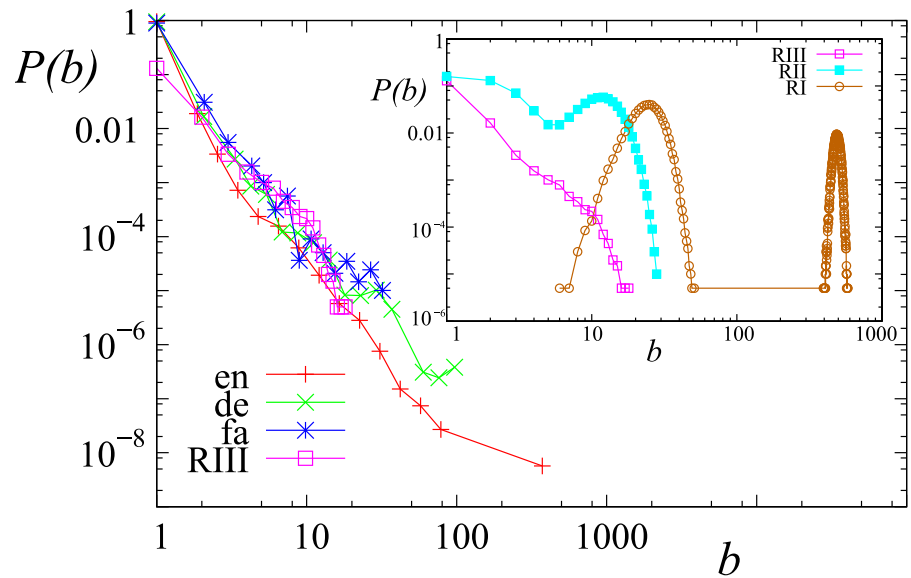


Fig 8. The distribution of the number of times users were banned in different regimes and in Wikipedia. Notation: en: English, de: German, fa: Persian, RI, RII, RIII stands for Simulation in regime I, II, III respectively. On the main plot only regime III is shown the other regimes for the model are shown in the inset.

<https://doi.org/10.1371/journal.pone.0173561.g008>

seems, that the effect of positive feedback is too small, and the fluctuations are not large enough to compensate the loss of actions.

We found that banning users hinders consensus building in most cases. As already mentioned, in reality in some cases banning may still help the process of consensus building but for other reasons not included in the model, for example: banned editors leave the pool, or with cooler heads after the banning period, editors can be more constructive. Thus further improvements of the model could be the implementation of these effects, for example: changing ϵ_A after banning. But the main message is similar to the previous section: one needs active interaction of all participants to achieve a consensus.

It is possible to verify our model by real Wikipedia data as there are two quantities which can be measured both in the model and on Wikipedia. First we show the distribution of the number of times users were banned b . The results are shown in Fig 8. The results for regime III look the same as the empirical data except for users banned once. This small deviation could be explained by the vandal editors on the Wikipedia, who deliberately delete articles or replace entire pages with nonsense. These editors are only banned once permanently, a feature (deliberately) missing from our model.

Now we turn to the relationship between the number of edits versus the number of bans. The simulation results are compared with actual Wikipedia data in Fig 9. The positive correlation between the number of edits and the number of times each user is banned reflects the old proverb that “it’s only those who do nothing that make no mistakes”. The correlation is very similar to the empirical observations. The only deviation is for large numbers of edits, where the model overestimates the number of bans. This may be the result of the ability of human editors to learn how to avoid mistakes resulting in banning, something that agents do not do in our model.

Obviously, regime III is the most similar to the Wikipedia data, which is important since relaxation time is the smallest in this regime which allows editors to reach a consensus within a reasonable time limit. This could be one reason behind the sustainability of Wikipedia.

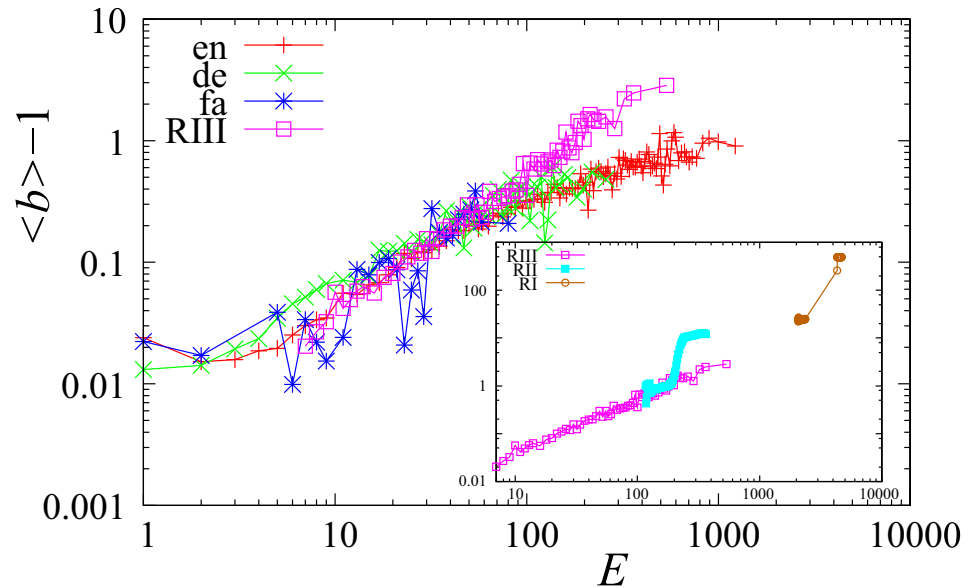


Fig 9. Number of edits versus number of bannings (minus 1) in Wikipedia and in the model. Notation: en: English, de: German, fa: Persian. On the main plot only simulation results of regime III is shown the numerical results of the other regimes of the model are only shown in the inset.

<https://doi.org/10.1371/journal.pone.0173561.g009>

We note here that the apparent similarity between regime III of the model and the empirical data is not a result of fitting. We have varied model parameters e.g. the exponent 2 in Eq (9) and found similar results. Of course one can distort the definition of the model such that even regime III results does not match with Wikipedia data, however the robustness of the results suggests that in Wikipedia the primary way to consensus is similar to the converging extremes behavior.

4 Conclusions

In this paper we investigated the role of people with extremist views in a value production environment like Wikipedia using both modeling and data analysis. Two questions were studied: the influence of the ratio of the extremists on the characteristic time to reach consensus and the effect of temporal banning.

We found that in order to achieve fast consensus in our model all participants need a constructive access to the common product in order to converge to a consensus. The worst scenario is when there is a strong mainstream group which punishes all moves of the extremist groups (a very general scenario in everyday life). We have even found that there is a phase transition like abrupt change in the relaxation time in the four opinion group system as function of the ratio of the extremists and, unexpectedly, the high relaxation time regime is for a low ratio of extremists.

We have also found out that, in general, there is an optimal ratio of talk/edit and it is never zero but especially in Regime III is very small. This is again a counterintuitive result, meaning that more discussion does not help consensus, very often it just freezes the front lines. We have shown that in real Wikipedia, the articles with higher talk/edit ratios have relatively more conflict.

Translating it to real life would mean that if debates are mediated by people with average (politically correct) views, then the debate will not lead to consensus as the extremist groups

will remain frustrated forever. If the extremists have a chance to see their opinion reflected in the medium, for at least a small amount of time, they are more inclined to change their views which is necessary for the consensus. Thus, the active participation of the extremists is needed for a consensus. Furthermore, too much discussion just strengthens the position of the people in their opinion group and does not allow them to leave it. We believe these observations are well reflected in other fields of opinion difference e.g. politics.

We also included banning, a general procedure of Wikipedia, and we found it counterproductive in most cases as it only delays the consensus building. The probability distribution of the number of times a user was banned in Regime III, matches very well with the Wikipedia data suggesting that the converging extremes is the most general convergence method on Wikipedia. It means that the point of view of the article is volatile and extremists become satisfied with a temporarily biased article while they also alter their views on the subject while at the end they accept a more mainstream version of the article.

Acknowledgments

TY has received funding from the European Union's Horizon2020 research and innovation program under grant agreement No 645043; "HUMANE: a typology, method and roadmap for HUMAN-MACHINE NETWORKS".

Author Contributions

Conceptualization: JT TY.

Data curation: TY.

Formal analysis: CsR OS TY JT.

Funding acquisition: TY.

Investigation: CsR OS TY JT.

Methodology: CsR OS TY JT.

Software: CsR OS TY JT.

Supervision: JT.

Validation: CsR OS TY JT.

Writing – original draft: CsR TY JT.

Writing – review & editing: CsR TY JT.

References

1. Segal B. A short history of Internet protocols at CERN. Professional webpage April <http://ben.home.cern.ch/ben/TCPHIST.html>. 1995;.
2. Leiner BM, Cerf VG, Clark DD, Kahn RE, Kleinrock L, Lynch DC, et al. A brief history of the Internet. ACM SIGCOMM Computer Communication Review. 2009; 39(5):22–31. <https://doi.org/10.1145/1629607.1629613>
3. Beaudouin-Lafon M, Beaudouin-Lafon M. Computer supported co-operative work. Wiley New York; 1999.
4. Tsvetkova M, Yasseri T, Meyer ET, Pickering JB, Engen V, Walland P, et al. Understanding Human-Machine Networks: A Cross-Disciplinary Survey. arXiv preprint arXiv:151105324. 2015;.
5. Eide AW, Pickering JB, Yasseri T, Bravos G, Følstad A, Engen V, et al. In: Kurosu M, editor. Human-Machine Networks: Towards a Typology and Profiling Framework. Cham: Springer International Publishing; 2016. p. 11–22.

6. Giles J. Internet encyclopaedias go head to head. *Nature*. 2005; 438(7070):900–901. <https://doi.org/10.1038/438900a> PMID: 16355180
7. Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*. 2009; 323(5915):721.
8. Watts DJ. Computational social science: Exciting progress and future directions. *The Bridge on Frontiers of Engineering*. 2013; 43(4):5–10.
9. Smets K, Goethals B, Verdonk B. Automatic vandalism detection in Wikipedia: towards a machine learning approach. In: *AAAI Workshop Wikipedia and Artificial Intelligence: an Evolving Synergy. WikiAI08. Association for the Advancement of Artificial Intelligence*; 2008. p. 43–48.
10. Potthast M, Stein B, Gerling R. Automatic vandalism detection in Wikipedia. In: *Proceedings of the IR research, 30th European conference on Advances in information retrieval. ECIR'08. Berlin, Heidelberg: Springer-Verlag*; 2008. p. 663–668.
11. Wu Q, Irani D, Pu C, Ramaswamy L. Elusive vandalism detection in wikipedia: a text stability-based approach. In: *Proceedings of the 19th ACM international conference on Information and knowledge management. CIKM'10. New York, NY, USA: ACM*; 2010. p. 1797–1800.
12. DeDeo S. Conflict and Computation on Wikipedia: a Finite-State Machine Analysis of Editor Interactions. *Future Internet*. 2016; 8(3). <https://doi.org/10.3390/fi8030031>
13. Yasseri T, Sumi R, Rung A, Kornai A, Kertész J. Dynamics of conflicts in Wikipedia. *PloS one*. 2012; 7(6):e38869. <https://doi.org/10.1371/journal.pone.0038869> PMID: 22745683
14. Tsvetkova M, García-Gavilanes R, Yasseri T. Dynamics of Disagreement: Large-Scale Temporal Network Analysis Reveals Negative Interactions in Online Collaboration. *arXiv preprint arXiv:160201652*. 2016;.
15. Vuong BQ, Lim EP, Sun A, Le MT, Lauw HW. On ranking controversies in wikipedia: models and evaluation. In: *Proceedings of the international conference on Web search and web data mining. WSDM'08. New York, NY, USA: ACM*; 2008. p. 171–182.
16. Yasseri T, Spoerri A, Graham M, Kertész J. The most controversial topics in Wikipedia: A multilingual and geographical analysis. *Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration*, Fichman P, Hara N, eds, Scarecrow Press. 2014;.
17. Kittur A, Chi EH, Suh B. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In: *Proceedings of the 27th international conference on Human factors in computing systems. CHI'09. New York, NY, USA: ACM*; 2009. p. 1509–1512.
18. Borra E, Weltevrede E, Ciuccarelli P, Kaltenbrunner A, Laniado D, Magni G, et al. Societal controversies in Wikipedia articles. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems. ACM*; 2015. p. 193–196.
19. Wikipedia. Help: Using talk pages;. [Online; accessed 17-May-2016]. http://en.wikipedia.org/w/index.php?title=Help:Using_talk_pages&oldid=492609161.
20. Viegas FB, Wattenberg M, Kriss J, van Ham F. Talk Before You Type: Coordination in Wikipedia. In: *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*; 2007. p. 78.
21. Yasseri T, Kornai A, Kertész J. A practical approach to language complexity: a Wikipedia case study. *PloS one*. 2012; 7(11):e48386. <https://doi.org/10.1371/journal.pone.0048386> PMID: 23189130
22. Laniado D, Tasso R, Volkovich Y, Kaltenbrunner A. When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages. In: *5th International AAAI Conference on Weblogs and Social Media. ICWSM 2011*; 2011. p. 177–184.
23. Iosub D, Laniado D, Castillo C, Morell MF, Kaltenbrunner A. Emotions under discussion: Gender, status and communication in online collaboration. *PloS one*. 2014; 9(8):e104880. <https://doi.org/10.1371/journal.pone.0104880> PMID: 25140870
24. Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. *Reviews of modern physics*. 2009; 81(2):591. <https://doi.org/10.1103/RevModPhys.81.591>
25. Deffuant G, Neau D, Amblard F, Weisbuch G. Mixing beliefs among interacting agents. *Advances in Complex Systems*. 2000; 3(01n04):87–98. <https://doi.org/10.1142/S0219525900000078>
26. Sobkowicz P. Extremism without extremists: Deffuant model with emotions. *Frontiers in Physics*. 2015; 3:17. <https://doi.org/10.3389/fphy.2015.00017>
27. Török J, Iñiguez G, Yasseri T, San Miguel M, Kaski K, Kertész J. Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Physical review letters*. 2013; 110(8):088701. <https://doi.org/10.1103/PhysRevLett.110.088701> PMID: 23473207
28. Ciampaglia GL. A bounded confidence approach to understanding user participation in peer production systems. In: *International Conference on Social Informatics. Springer*; 2011. p. 269–282.

29. Iñiguez G, Török J, Yasseri T, Kaski K, Kertész J. Modeling social dynamics in a collaborative environment. *EPJ Data Science*. 2014; 3(1):1–20.
30. Yasseri T, Kertész J. Value production in a collaborative environment. *Journal of Statistical Physics*. 2013; 151(3-4):414–439. <https://doi.org/10.1007/s10955-013-0728-6>
31. Sumi R, Yasseri T, Rung A, Kornai A, Kertész J. Characterization and prediction of Wikipedia edit wars. In: *Proceedings of the ACM WebSci'11*, Koblenz, Germany; 2011. p. 1–3.
32. Sumi R, Yasseri T, Rung A, Kornai A, Kertész J. Edit wars in Wikipedia. In: *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom)*; 2011. p. 724–727.