Modeling the Wikipedia to Understand the Dynamics of Long Disputes and Biased Articles

Rudas, Csilla; Török, János

Historical Social Research
Historische Sozialforschung

Csilla Rudas & János Török:

Modeling the Wikipedia to Understand the Dynamics of Long Disputes and Biased Articles.

doi: 10.12759/hsr.43.2018.1.72-88

Published in:

Historical Social Research 43 (2018) 1

Cite as:

Rudas, Csilla, and János Török. 2018. Modeling the Wikipedia to Understand the Dynamics of Long Disputes and Biased Articles. *Historical Social Research* 43 (1): 72-88. doi: 10.12759/hsr.43.2018.1.72-88.

Historical Social Research
Historische Sozialforschung

All articles published in HSR Special Issue 43 (2018) 1:
Agent-Based Modeling in Social Science, History, and Philosophy.

Modeling the Wikipedia to Understand the Dynamics of Long Disputes and Biased Articles

Csilla Rudas & János Török [*]

Abstract: *»Wikipedia. Eine agentenbasierte Modellbetrachtung zur Dynamik konkurrierender Editionsversuche«*. The Internet has provided us with a number of online collaborative environments, including platforms for open software developments and online encyclopedias such as Wikipedia. Conflicts may arise in the course of such collaboration, but despite differences of opinion consensus can be reached. By investigating the consensus-building processes, we can shed light on the dynamics of social behavior. In Wikipedia, it is not always easy for editors to agree about article content, especially considering people's different tolerance levels towards others and for whatever may be written. In this paper, we focus on how the editors' attitudes, namely being broad-minded or stubborn, affect the consensus-building process in a model of Wikipedia. We further investigate how banning editors affects the speed with which conflicts or debates can be resolved. For the analysis, we use an agent-based opinion model developed to simulate different aspects of Wikipedia. We show that, in most cases, banning agents from editing an article slows down the consensus-building process, and increases the system's relaxation time. We show further, and counterintuitively, that with large groups of "extremists" who hold other than the central opinion, consensus can be reached faster and the article will be less biased.

Keywords: Wikipedia, agent-based modeling, social conflict, collaborative environment, relaxation time, banning, tolerance, consensus.

1. Introduction

Collaboration is indispensable for society to solve problems. As the Internet developed and more and more data became available about people's social and personal activities, it transformed the study of collective behavior and created a new field and approach called computational social science (Lazer et al. 2009;

[*] Csilla Rudas, Centre for Energy Research, Hungarian Academy of Sciences, Konkoly-Thege Miklós út 29-33, H-1121 Budapest, Hungary; csilla.rudas@energia.mta.hu.
János Török, Department of Theoretical Physics, Budapest University of Technology and Economics, Budafoki út 8, H-1111 Budapest, Hungary; torok@phy.bme.hu.

Historical Social Research 43 (2018) 1, 72-88 | published by GESIS
DOI: 10.12759/hsr.43.2018.1.72-88

Watts 2013). To investigate the opinion dynamics and the underlying mechanisms of Wikipedia, we use such an approach.

Other researchers have studied different aspects of Wikipedia, for example, vandalism (Smets et al. 2008; Potthast et al. 2008; Wu et al. 2010), controversy over articles (Vuong et al. 2008; Yasseri et al. 2014; Kittur et al. 2009), and how to detect conflicts (Borra et al. 2015). Most of the literature deals with conflicts but fails to investigate what constitutes the conflict-resolution process. In this paper, we extend our earlier study (Rudas et al. 2016) to investigate how different banning strategies affect the consensus-building process as well as the contents of a jointly edited article at a certain point in our simulation.

Banning users in an online social environment is a sensitive issue. Where there is a central moderator, a decision may be taken to ban certain users based on their behavior, but even in such cases principles are required upon which to make such a decision. Having clear rules for banning is even more desirable if there is no single administrator as in the case of Wikipedia (Wikipedia 2017).

Our earlier study showed that temporal banning does not always help in the consensus-building process (Rudas et al. 2016). Agents do not change their opinion during the banning period and the conflict starts anew when they return to the editorial pool. Here we investigate how having new agents join or banned agents leave the editorial pool effects relaxation time. We have found the interaction of agents with the article and the fluctuation of the bias of the latter are the key features in the consensus-building process. Our results show that banning, in most cases, has a negative effect on relaxation time and helps only if the relaxation time is already small without the ban.

Section 2 describes our model for investigating the Wikipedia consensus-building process and specifies the parameters and their values used in our simulations. Section 3 first summarizes the main effects of using different parameter values in the simulations, then discusses the effects of different banning strategies, and finally shows the results for article content at the end of the simulations. Section 4 presents our conclusions.

2. Methods

The literature on agent-based opinion dynamics is extensive; for a review, see Castellano et al. (2009). To investigate the dynamics of articles on Wikipedia, we use an agent-based model we developed earlier in Török et al. (2013). This model simulates the editor-editor interaction based on the principle of "bounded confidence," which means that agents can accept opinions different from their own up to a certain tolerance threshold (Deffuant et al. 2000). Such models and their generalizations have been applied to explain the dynamics of Wikipedia (Ciampaglia 2011). In this paper, we use a generalization in which,

in addition to communicating with each other, agents also edit and interact via a common product (Iñiguez et al. 2014).

2.1 Computational Model

Opinion dynamics on Wikipedia are modeled by N editors (agents) who have an opinion on a subject, with a corresponding Wikipedia article.

We represent N agents with scalar opinion values $x_i(t)$ in the range of [0,1]. Initially, the agent opinions are randomly sampled from the uniform distribution. The agents can talk to each other only if the difference between their opinion values is less than the tolerance parameter ϵ_T. After talking, the agents adopt a new view halfway between their original ones:

$$(x_i, x_j)_t \rightarrow \begin{cases} \left(\frac{x_i+x_j}{2}, \frac{x_i+x_j}{2}\right)_{t+1} & if\ |x_i - x_j| < \epsilon_T \\ (x_i, x_j)_{t+1} & otherwise \end{cases} \quad (1)$$

This talk-interaction is based on the bounded confidence opinion dynamics introduced in Deffuant et al. (2000). As a result of the talk interactions, the agents form opinion groups, which are determined by the initial conditions and the value of ϵ_T. In addition to the talk interaction, agents can edit an article in common with opinion bias $A(t)$, which has a value in the same [0,1] interval as the agent opinions. The agents edit the article only if they are dissatisfied with it, namely when A is farther from their opinion than an article tolerance ϵ_A. If the agent is not satisfied with the view the article reflects, the agent edits the article, changing it with the product of a convergence parameter μ_A and of the distance between article A and its opinion value x_i, as shown in Equation 2. When the article is acceptable to the agent, the agent changes her or his own opinion regarding the article value similarly. Thus, the editing action is the opposite of the talking action, with the twist of persuasion by reading an acceptable article:

$$(x_i, A)_t \rightarrow \begin{cases} (x_i + [A - x_i] \cdot \mu_A, A)_{t+1} & \text{if } |x_i - A| < \epsilon_A \\ (x_i, A + [x_i - A] \cdot \mu_A)_{t+1} & otherwise \end{cases} \quad (2)$$

In Wikipedia, editors can be banned if they do not follow the rules (Wikipedia 2017). Violations include changing the tone of an article too much or just undoing other edits too often, which leads to "edit wars." The banning or blocking policy on Wikipedia is complex and may even include community decisions that would be too complicated to include in our simple model. Instead, to keep our simulations memoryless and simple, we ban editors with probability p after they have modified the article. In the model, if during the edit action the editor was satisfied with the article, the agent is not banned. But if the agent was unsatisfied with the article and edits it, the agent may be banned with the following probabilities:

a) $p = 0$
b) $p = 0.5$

c) $p = (|x_i - A| - \epsilon_A)^2$

The last choice gives a banning probability that strongly increases with the level of dissatisfaction and thus with the amount of change in the article.

Banning on Wikipedia is generally not intended as a short-term punishment but rather is a decision that a user may not edit a specific article in the future. In the model, agents get banned by virtue of one editing action, but are still able to participate in talk-talk interactions. After the unsuccessful edit, the ban is lifted and they rejoin the editing pool with their same opinion as before. In this sense, the banning process in our model is similar to a short-term article or topic ban in Wikipedia (Wikipedia 2017).

Editors may become dissatisfied after being banned and leave the editorial pool. To simulate this, we introduced the probability of leaving the agent pool P=0.5, which means that each time an agent gets banned a new agent replaces the banned one with probability P, bringing a newly sampled opinion. We set the fourth banning option as:

d) $p = (|x_i - A| - \epsilon_A)^2, P = 0.5$

From the empirical point of view, we find option (d) to be the most realistic, as it best reflects actual Wikipedia rules (Wikipedia 2017).

We defined the relaxation time τ by the number of time-steps needed for all agents (even banned ones) to reach consensus. In each time-step, N interactions are performed. In each interaction, either two agents talk or an agent reads the article and edits it if necessary. Both types of interactions occur with probability 0.5. Consensus is reached if all agents are satisfied with the view expressed by the article, which indicates that the article value is within each agent's tolerance range. The ensemble average of the relaxation times measured in different runs does not represent well the empirical density function of τ, which we found in most cases to be a sharp peek with an exponential tail. Thus, instead of the ensemble average, we use the empirical mode value of the relaxation time distribution.

We measure another parameter of the system, the standard deviation $\sigma(t)$ of the distribution of the article values at time t, as the square root of the average deviation from the expected $A(t) = 0.5$ value in different runs. When measuring the standard deviation, we plot the final article values in a histogram with B = 20 bins and use a E = 500 ensemble. This value of $\sigma(t)$ is lower if the distribution is more concentrated and higher if it is more spread out.

Our earlier study investigated the formation of opinion groups (Rudas et al. 2016). In this paper, we set three initial opinion groups positioned at the following intervals: one mainstream group at 0.45 – 0.55 and two extremists at 0 – 0.1 and 0.9 – 1.

We have also investigated the case with four initial opinion groups but as the results were qualitatively similar to the three-group system, we restricted the present analysis to three initial groups.

Here, we use the term “extremist” in a technical sense to refer to those with opinions different from the mainstream opinion. In real life, such “extremist” opinions may be seen as quite typical.

The choices above facilitate the modeling of how different initial opinion distributions affect the outcome of the simulation. We use the ratio of extremists *RoE* to set how many agents are initially in the extreme opinion groups relative to the entire agent pool. So, if *RoE* = 0.30, then 30% of the agents are distributed equally (15-15%) between the two extremist groups and 70% of the agents are in the mainstream group. We chose symmetric distribution of the agent's initial opinions because we are interested in the symmetry breaking, or the resulting article bias and relaxation time caused by specific model parameters. The study of biased initial condition is left for future studies.

2.2 Tolerance Inhomogeneity

In our original model (Török et al. 2013), every agent had the same tolerance for others (ϵ_T) and the article (ϵ_A), so the model assumed all participants are equally tolerant or intolerant. To make the current model more realistic, we substitute constant and static values with a linear tolerance distribution that depends on the opinion value of a given agent:

$$\epsilon_T(t) = \epsilon_A(t) = -|x_i(t) - 0.5| \cdot m + c \qquad (3)$$

Where m is the slope parameter and c is the constant offset.

Tolerance parameters are assigned to every agent based on their current opinion value. Each time the opinion of an agent changes, whether because of talking to another agent or reading the article, the tolerance changes as well. Thus, rather than having the same tolerance value for each agent (as in Török et al. 2013; Iñiguez et al. 2014; Rudas et al. 2016), if $m > 0$ agents with central opinion values have larger tolerance values (Weisbuch et al. 2005) and, naturally, agents with extreme views have lower tolerances than the rest. Because of this, ϵ_T in Equation 1 is replaced by the lower tolerance value of the two. In Equation 3, c parameterizes the controversial nature of the subject, with high values meaning less disputed, while the inhomogeneity parameter m describes the polarization of the agent tolerance pool on the subject.

We have chosen the same tolerance parameter for the editor-editor interaction as for the editor-article interaction for two reasons: First, it is always better to have fewer parameters, second, the tolerance of a person applies to a subject, not a person or article. In case of Wikipedia in particular, agent-agent interaction often takes the form of talk page editing, which is the same written interaction as in the case of the article.

2.3 Summary of the Parameter Values

Table 1 shows the values of the parameters used throughout the simulations. In the section that follows, we discuss the phenomena linked with these parameters.

Table 1: Simulation Parameters

Symbol	Name	Values
N	size of the agent pool	100
x_i (t)	opinion value for agent i at time t	∈ [0,1]
A(t)	article value at time t	∈ [0,1]
μ_A	convergence parameter	∈[0.1,0.9]
m	inhomogeneity parameter	0.5, 0.33, 0.2, 0.1
c	sensitivity parameter	0.25, 0.3, 0.35, 0.4
ϵ_T	tolerance towards talking	$c - m \cdot \lvert x_i - 0.5 \rvert$
ϵ_A	tolerance towards the article	$c - m \cdot \lvert x_i - 0.5 \rvert$
p	banning probability	0, 0.5, $(\lvert x_i - A \rvert - \epsilon_A)^2$
P	probability of leaving the agent pool	0, 0.5
RoE	ratio of extremists	∈ [0.1,0.9]
τ_{max}	maximum time-step of the simulation	100, 10 000
B	number of bins in article distribution histogram	10, 20
E	size of ensemble (number of runs with the same parameters)	100, 500

In previous publications (e.g., Török et al. 2013) the thermodynamic limit ($N \to \infty$) was of interest. However it was shown in Yasseri and Kertész (2013) that much fewer people edit the majority of the articles. Hence, $N = 100$ is a good estimate of the number of editors, even for featured articles.

The opinion values of the agents were chosen to represent the most common case, that is, when there is a mainstream group and two extremist groups. We note here that the initial number of opinion groups has a strong influence on the dynamics (Rudas et al. 2016).

Our present implementation also includes a parameter RoE that provides the ratio of the users with extreme opinions $x(i) \cong 0$ or 1 with respect to mainstream agents.

The convergence parameter of the article μ_A may represent the inverse of the length of the given article or section. This μ_A describes the amount of change an agent can make when editing the article. Lower values represent longer articles, as it is more difficult to change the tone of more detailed and extensive articles. We will see that it may play a decisive role in the behavior of the system.

In general, in the bounded confidence models (Deffuant et al. 2000; Török et al. 2013; Rudas et al. 2016), the tolerance parameter for both the agents and the article is a critical parameter that has a drastic influence on the properties of the model. However, in the original definition, the tolerance parameter of all users is the same, while it is known that people with extreme views are less

tolerant. We modified the model in this respect. We set both the agent-agent and agent-article tolerance by Equation 3, which has two parameters: the inhomogeneity parameter, which for large (~0.5) values gives very hard-headed extremists; and c, which controls the sensitivity of the subject with low values of c for debated topics.

Implementing the banning in a probabilistic way was motivated by the fact that dissatisfied users may take actions that may eventually lead to banning. We implemented three cases: no banning $p = 0$; banning with a given probability $p = 0.5$ as an unrealistic reference; and p proportional to the square of the change in the article, which reflects the fact that more dissatisfied users are more prone to take actions that result in banning.

In our implementation, banned users are either returned to the editorial pool after a short time, $P = 0$, or get banned forever, with probability P, and are replaced by new editors with randomly chosen opinions.

An interesting special case is when $0 \leq c/m \leq 0.5$, which means that agents with a zero tolerance level may appear. Since the possibility of communication depends on the lower of the tolerance values of the two agents, communication will not be possible. In addition, and for the same reason, such agents may never find an article to their liking. This has the potential of pulling the article value towards an extreme position. Nevertheless, we included a parameter pair $m = 0.5$, $c = 0.25$ in our parameter set, as there may be topics for which the presence of such intolerant agents is realistic.

3. Results

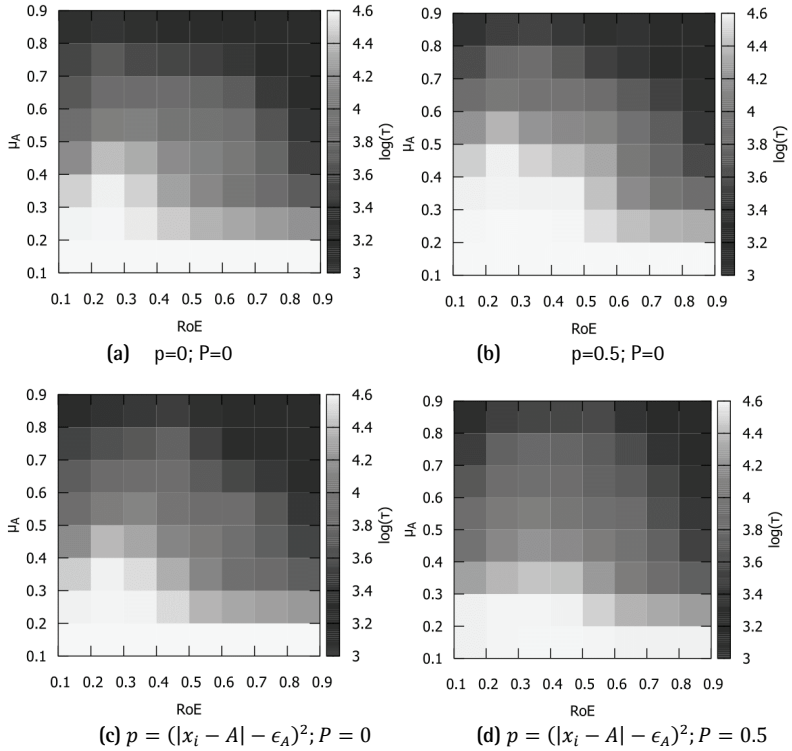
First, we calculated the relaxation times for all the combinations of m and c and the 9 different values of μ_A and RoE each from 0.1 to 0.9, for an ensemble of 100 independent runs. There are parameter combinations when it is practically impossible to reach a consensus. To tackle this issue, we set a maximum value for the number of time-steps performed in the calculations τ_{max}.

3.1 The Behavior for Different Parameters

Banning Strategies

We implemented four different banning strategies (compared these in Figure 1): (a) no banning at all; (b) banning with fixed probability; (c) banning proportional to the dissatisfaction level; and (d) banning proportional to the dissatisfaction level, with banned users getting replaced by new ones with probability P. Despite these varying scenarios, the differences of the logarithm of the relaxation times is only marginal. This is illustrated in Figure 1 for the choices $c = 0.4$ and $m = 0.5$.

Figure 1: Heat Maps of the Logarithm of the Relaxation Time Log (τ) as the Function of *RoE* and μ_A for Different Banning Strategies. c=0.4, m=0.5, τ_{max}=100



(a) p=0; P=0 **(b)** p=0.5; P=0

(c) $p = (|x_i - A| - \epsilon_A)^2; P = 0$ **(d)** $p = (|x_i - A| - \epsilon_A)^2; P = 0.5$

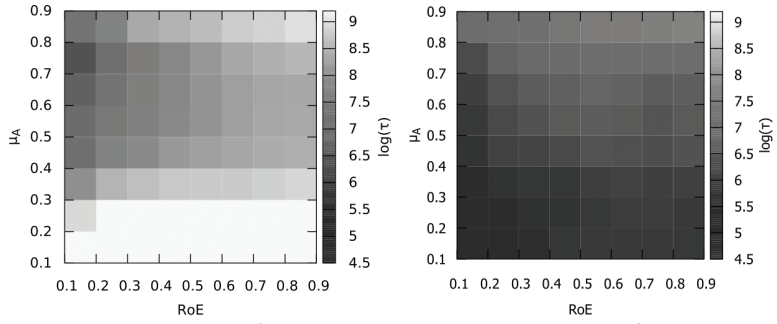
We can observe that there are similar patterns in the relaxation time plots for all banning strategies. For example, all results are very sensitive to the value of μ_A, so that low μ_A values result in extremely high relaxation times. This means that allowing editors to change only a limited amount of the text (and the bias of the article) is unfavorable for the consensus. In our model, similar to Rudas et al. (2016), editors must take an active part in the discussion and the editing process to achieve consensus.

We observe that a high ratio of extremists speeds up the consensus-building process, but (contrary to Rudas et al. 2016) a low level of extremists can also produce fast convergence. It seems there is a range of ratio of extremists that results in long relaxation. We found this to be around $RoE \cong 0.2–0.5$ in our model.

There is a case ($c = 0.25$, $m = 0.5$) in which consensus was rarely reached in the given simulation time: $\tau_{max} = 100$ at first. After raising the maximum

time step value to τ_{max} = 10,000, we found that in this case, contrary to previous findings, scenarios (a)-(c) produced similar results, but that the banning scenario (d) had much smaller relaxation times. This is shown in Figure 2. The relaxation time heat maps in (c) and (d) of Figure 2 show completely different trends. The first three banning strategies mostly resulted in disagreement (only shown for scenario (c)); surprisingly, consensus was always reached when unsatisfied agents were replaced with probability P (scenario (d)). This is because in this special case, there can be editors at extreme values of x with almost zero tolerance and who cannot be persuaded. Banning alone cannot help in this case; rather, the replacement of these agents is needed to manage and resolve the conflict, which happens only in the banning strategy (d).

Figure 2: Heat Maps of the Logarithm of the Relaxation Time Log (τ) as the Function of *RoE* and μ_A for Different Banning Strategies. c=0.25, m=0.5, τ_{max}=10,000

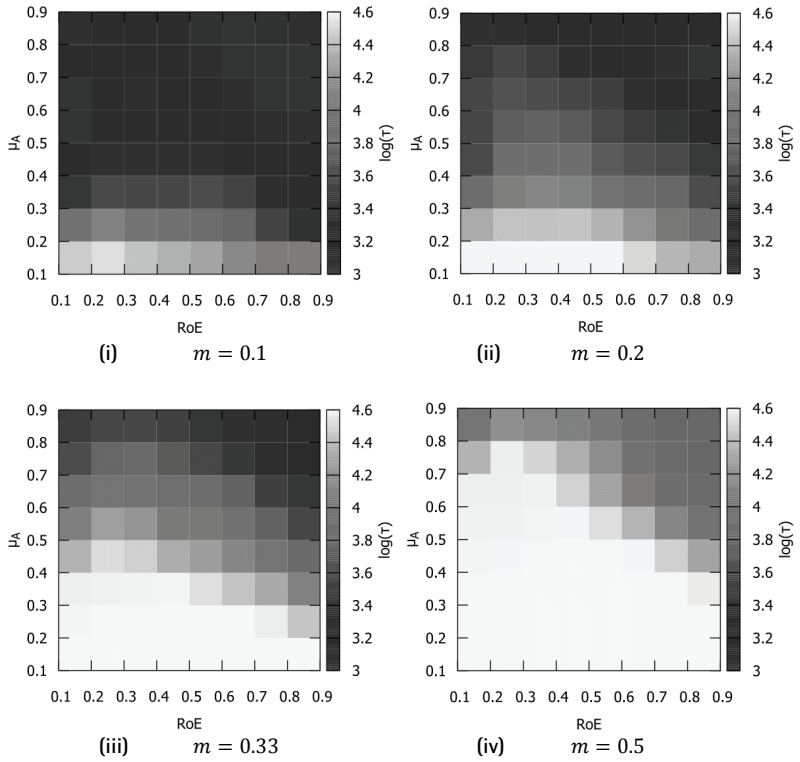


(c) $p = \left(|x_i - A| - \epsilon_A\right)^2; P = 0$ (d) $p = \left(|x_i - A| - \epsilon_A\right)^2; P = 0.5$

Inhomogeneity Parameter

Changing the slope (m) of the tolerance distribution has a considerable effect on the logarithm of the relaxation time, as Figure 3 shows. Here, we fixed the sensitivity parameter (c) of the tolerance distribution and studied the heat maps of the logarithm of the relaxation time for different slopes. In accordance with our expectations, the best case is when the slope is small (i) m = 0.1; the worst case is when the slope is the steepest (iv): m = 0.5.

Figure 3: Heat Maps of the Logarithm of the Relaxation Time Log (τ) as the Function of *RoE* and μ_A for Different Inhomogeneity Parameters. c=0.35, p=0.5, P=0, τ_{max}=100



(i) $m = 0.1$ (ii) $m = 0.2$

(iii) $m = 0.33$ (iv) $m = 0.5$

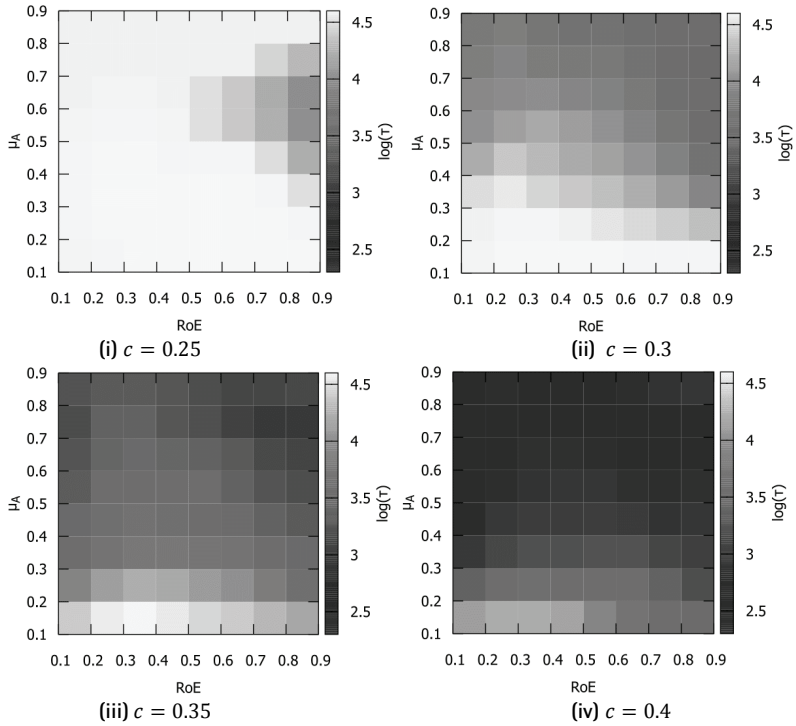
We explored negative values for m, which means that editors with opinions closer to 0.5 are less tolerant and that editors with more extreme opinions (closer to 0 or 1) are more tolerant towards each other and towards the article. In this case, the relaxation time was always small, and the conflict was resolved quickly. The initial groups merged rapidly into one final group in the middle, as the extremists on the sides with higher tolerance were persuaded and pulled easily towards the stubborn middle group.

Sensitivity Parameter

As expected, the consensus building takes a lot of time for low c values, as shown in (i) of Figure 4, where in most cases consensus is not even reached in the given $\tau_{max} = 100$ time. Increasing the value of parameter c accelerates the process, as more agents are able to talk to each other and become satisfied with

the article. Combining this result with those for the different inhomogeneity parameters, one sees that it is more important for fast consensus to have no intolerant agents than to have many very tolerant ones.

Figure 4: Heat Maps of the Logarithm of the Relaxation Time Log (τ) as the Function of *RoE* and μA for Different Sensitivity Parameters. m=0.25, $p=(|x_i - AI - \epsilon_A|)^2$, P=0.5, τ_{max}=100



(i) $c = 0.25$ (ii) $c = 0.3$

(iii) $c = 0.35$ (iv) $c = 0.4$

3.2 Banning

In this section, we discuss in greater detail the effect of banning on the relaxation time. We use the results of very long ($\tau_{max} = 10{,}000$) simulations during which, in most cases, consensus was reached.

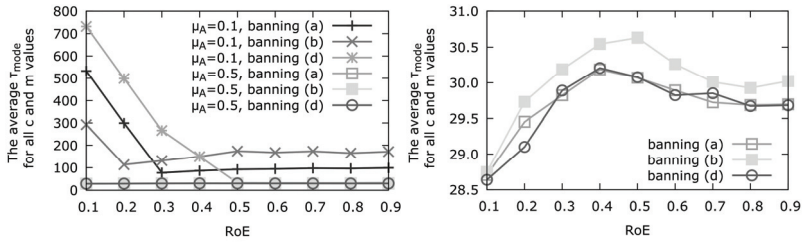
We measured the relaxation time for $\mu_A = 0.1$ and $\mu_A = 0.5$. Figure 5 shows the calculated average τ for all the combinations of the four values of c and m. The left side shows all the results together. For $\mu_A = 0.1$, we can see extremely high relaxation time values for low ratios of extremists. In these cases, the consensus is sometimes not reached, even for $\tau_{max} = 10{,}000$. According to our model, this means that a small change in the position of the article (determined by low μ_A) is undesirable, as it may leave editors with extreme views frustrat-

ed. We have already shown in Rudas et al. (2016) that where there is a populous mainstream group and two small extremist ones, the system is stable and barely reaches consensus. The combination of these two elements leads to a gigantic relaxation time for $\mu_A = 0.1$ and $RoE < 0.2$.

Figure 5 (i) compares the effectiveness of the different banning strategies in our model. Strategy (b) with fixed banning probability is effective when $RoE < 0.3$, but is the least effective for $RoE > 0.4$. Strategy (d), which uses editor replacement, has the smallest relaxation time for $RoE > 0.45$. For RoE values of 0.3 and 0.4, no banning is the best. We find it interesting that different banning strategies are optimal depending on the ratio of extremists. This indicates that more realistic modeling would require good estimates of the RoE values that occur in Wikipedia.

The results for $\mu_A = 0.5$ are indistinguishable in this case, as they are an order of magnitude smaller than the results for $\mu_A = 0.1$. These are the cases of fast relaxation. The details can be seen in Figure 5 (ii), where the relaxation time averaged for different values of c and m is shown only for $\mu_A = 0.5$. The displayed range of the vertical axis is only 7% of the maximum, so the values are, for practical purposes, equal. Nevertheless, it is interesting that here the lower number of extremists is optimal and that banning either makes things worse (strategy (b)) or changes nothing (strategy (d)) compared with no banning.

Figure 5: The τ Values for Different Banning Strategies and RoEs Averaged for Different Values of c and m. τ_{max}=10,000



(i) $\mu_A = 0.1, 0.5$ (ii) $\mu_A = 0.5$

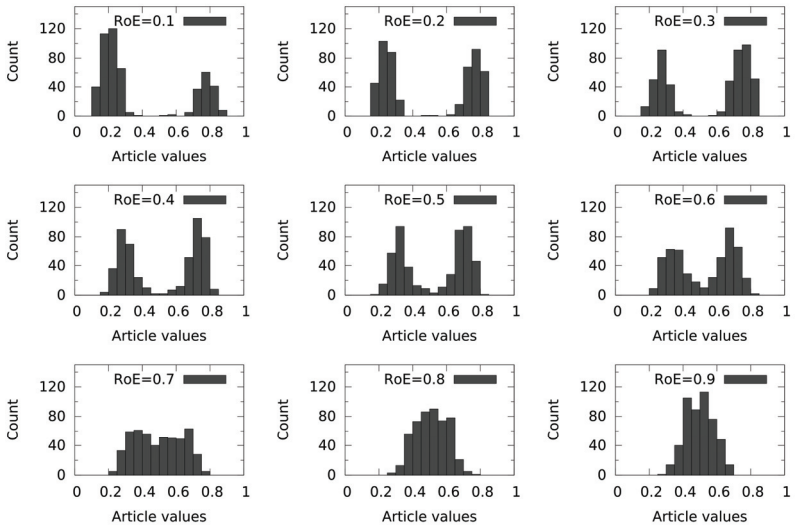
3.3 Bias of the Article

In this section, we study the position of the article value after $\tau_{max} = 100$ and $\tau_{max} = 10{,}000$ iterations. In some cases, consensus has not yet been reached. Nevertheless, the results are relevant because in real life every Wikipedia article has been through a number of editing cycles, possibly without having reached a consensus. Also in most cases in which consensus was reached in the given time, the final article's position can be regarded as the final content of a disputed article on Wikipedia.

In the description, we refer to an article as biased if its position is not equal to 0.5. Translated to the distribution, a biased case is when the article distribution has more than one peak (disregarding the small statistical fluctuations). We measured the standard deviation of the distribution σ as the square root of the average distance of the bars in the histogram from the expected $A = 0.5$ value. Hence, σ describes the bias of the article distribution well.

We ran the simulations for all combinations of *RoE, m, c*, and the four banning strategies, with an $E = 500$ ensemble and $\mu_A = 0.5$. We plotted the final article values in 20 bin histograms (shown in Figure 6) for a specific set of parameters, although for almost every combination of *m, c,* and *p* (except $m = 0.5$, $c = 0.25$) a similar picture was obtained. For small *RoE*, the article is very biased, with two symmetric peaks rather far from the middle. When the *RoE* is raised, the peaks move towards each other, until around $RoE \sim 0.6 - 0.7$, where the distribution turns unimodal. Raising the *RoE* further, the distribution becomes more concentrated around $A(\tau_{max}) = 0.5$.

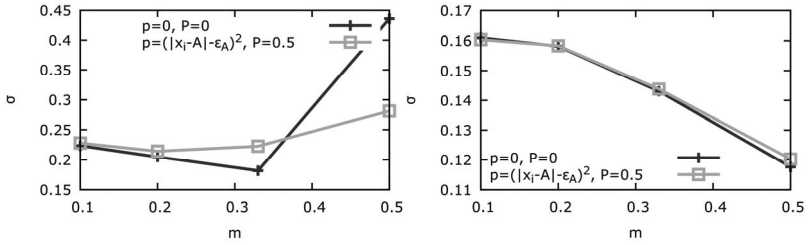
Figure 6: Article Value Distribution with τ_{max} for Different *RoEs*, *m*=0.33, *c*=0.25, *p*=0.5, *P*=0, *B*=20, *E*=500



The comparison of the charts in Figure 6 yields very surprising results. Not only is it faster to reach a consensus with an extremist agent majority, but the resulting article also becomes less biased. We always began the simulations with an equal number of upper and lower extremists, but it seems again that if there are few hard-headed extremists, one group will eventually win and the bias of the resulting article will be high. Conversely, if there are more extrem-

ists they take an active part in the discussion, which moderates them to the middle (De Vries and Edwards 2009; Adams et al. 2006).

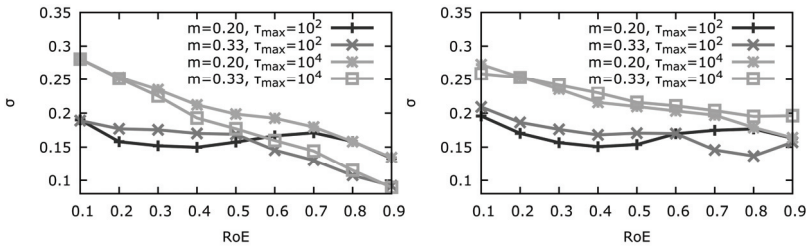
Figure 7: The Standard Deviation of the Article Distribution (σ) for c=0.25, 0.40. *RoE*=0.2, τ_{max}=10000



(i) $c = 0.25$ **(ii)** $c = 0.40$

We measured the standard deviation of the histograms and plotted them against the inhomogeneity parameter m. The data are shown in Figure 7 for τ_{max} = 10,000 at RoE = 0.2 for different values of c and banning strategies. Banning strategies (b) and (c) produced results very similar to (a), so graphs were omitted for better readability. Clearly, for large values of c, the difference between the banning strategies is small and the standard deviation is reduced with increasing m. For smaller c and larger inhomogeneity, banning strategy (d) leads to a lower standard deviation. This means that for the case of c = 0.25 and m = 0.5, banning scenario (d) helps make the articles less biased, but for a high c value it has no effect on the bias.

Figure 8: The Standard Deviation of the Article Distribution (σ) for All *RoE*, c=0.25, (a) and (d) Banning Scenarios, m=0.2, 0.33, τ_{max}=100, 10,000



(a) banning scenario **(d)** banning scenario

Figure 8 compares the standard deviations for τ_{max} = 100 and 10,000. We see that with rising RoE the standard deviation decreases, except for m = 0.2 and τ_{max} = 100, probably because consensus was not reached in that case. For small values of RoE in particular, the value of σ is constant and small for τ_{max} = 100; after more time, τ_{max} = 10,000 increases considerably. This

suggests that articles with low *RoE* may be largely unbiased even during discussion, but become more and more biased as the discussion continues over time.

The above results indicate that it is important to study the effect of banning and tolerance on the bias of the resulting article. Undisputed articles (with quick consensus) tend to be unbiased if there are many extremists (see Figure 6), but in other cases involving extremists very biased outcomes may occur. In this respect, banning helps only if the editors are fully uncooperative (see Figure 7).

Our model results suggest (see Figure 6) that the larger the mainstream group, the more biased the article may become.

4. Conclusion

In this paper, we studied the effect of banning on the time needed to reach consensus and on the bias of the article. We used the model introduced to describe the process of Wikipedia editing (Török et al. 2013), but modified it so agents have a unique tolerance towards a given subject.

In terms of banning, we have shown that, in most cases, the relaxation time to consensus is slightly increased or unchanged by banning. There were only two cases in which banning helped a lot. First, when there are intolerant editors, the only chance for consensus is if these agents leave the editorial pool, which can only be achieved with permanent banning. Second, if the article can be changed only by a small amount (described by μ_A in our model), consensus may be reached very slowly. In this case, depending on the ratio of extremists, different (or no) banning strategies can be optimal.

By introducing an opinion dependent tolerance, we have shown that for the consensus it is more important not to have intolerant editors than to have very tolerant ones.

Our results indicate that consensus is reached extremely slowly if the bias of the article can be changed only by a small amount. To resolve the conflict faster, one must either increase the change of bias in one edit or the ratio of extremists. In general, the latter cannot be controlled deliberately, but the former can be influenced.

In Wikipedia, there is already a method aimed at resolving disputes of that sort. The solution is to move the disputed questions into a new section (or page) where they can be discussed freely. The new trend to move disputed parts of the article into the *Criticism* or *Controversy* sections is a good way to handle this problem. Assigning sensible arguments and opinions to a small section of the article that is much easier to modify makes the full article less disputed. Thus, tolerance towards the main article increases, and even though tolerance towards the small *Criticism/Controversy* section may decrease significantly, that section's limited extent will afford it a much larger μ_A. Together

with the likely higher *RoE*, these two effects may result in a faster consensus. With this method, real Wikipedia disputes may avoid the low μ_A and *RoE* range and, therefore, be resolved more quickly. For definite confirmation of these effects, further investigation is needed.

Studies on consensus building to date have focused on relaxation time, but we show that the bias of the resulting article can be even more important. Oddly enough, our model indicates that the more extremists there are, the less biased the article becomes, provided that the size of the two opposing extremist groups are equal. Moreover, the bias of the article often increases as it approaches consensus.

We confirmed that the agent-based model presented in this paper has strong relevance for Wikipedia despite its simplicity (Iñiguez et al. 2014; Rudas et al. 2016). It makes a number of surprising predictions that may be particularly relevant for Wikipedia, even though some are quite counterintuitive. A few results are in line with present trends in Wikipedia (e.g., large μ_A vs. moving controversial issues to separate parts), but other predictions such as increased article bias if extremists are banned could be tested with newer text analysis tools (Callahan and Herring 2011).

References

Adams, James, Michael Clark, Lawrence Ezrow, and Garret Glasgow. 2006. Are niche parties fundamentally different from mainstream parties? the causes and the electoral consequences of western European parties' policy shifts, 1976-1998. *American Journal of Political Science* 50 (3): 513-29.

Borra, Eric, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. Societal controversies in wikipedia articles. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 193-6. New York: ACM.

Callahan, Ewa S., and Susan C. Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology* 62 (10): 1899-915.

Castellano, Claudio, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of modern physics* 81 (2): 591.

Ciampaglia, Giovanni Luca. 2011. A bounded confidence approach to understanding user participation in peer production systems. In *International Conference on Social Informatics*, 269-82. Berlin, Heidelberg: Springer-Verlag.

De Vries, Catherine E., and Erica E. Edwards. 2009. Taking Europe to its extremes extremist parties and public euroscepticism. *Party Politics* 15 (1): 5-28.

Deffuant, Guillaume, David Neau, Frédéric Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems* 3: 87-98.

Iñiguez, Gerardo, János Török, Taha Yasseri, Kimmo Kaski, and János Kertész. 2014. Modeling social dynamics in a collaborative environment. *EPJ Data Science* 3 (1): 1-20.

Kittur, Aniket, and Bongwon Suh Ed H. Chi. 2009. 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, 1509-12. New York: ACM.

Lazer, David, Alex S. Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Life in the network: the coming age of computational social science. *Science* 323 (5915): 721.

Potthast, Martin, Benno Stein, and Robert Gerling. 2008. Automatic vandalism detection in wikipedia. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, 663-8. Berlin, Heidelberg: Springer-Verlag.

Rudas, Csilla, Olivér Surányi, Taha Yasseri, and János Török. 2016. Understanding and coping with extremism in an online collaborative environment. *PLOS ONE* 12 (3): e0173561. <https://doi.org/10.1371/journal.pone.0173561> (Accessed January 25, 2018).

Smets, Koen, Bart Goethals, and Brigitte Verdonk. 2008. Automatic vandalism detection in wikipedia: towards a machine learning approach. In *AAAI Workshop Wikipedia and Artificial Intelligence: an Evolving Synergy*, WikiAI08, 43-8. Association for the Advancement of Artificial Intelligence.

Török, János, Gerardo Iñiguez, Taha Yasseri, Maxi San Miguel, Kimmo Kaski, and János Kertész. 2013. Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Physical Review Letters* 110 (8): 088701.

Vuong, Ba-Quy, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, Hady Wirawan Lauw, and Kuiyu Chang. 2008. On ranking controversies in wikipedia: Models and evaluation. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, 171-82. New York: ACM.

Watts, Duncan J. 2013. Computational social science: Exciting progress and future directions. *The Bridge on Frontiers of Engineering* 43 (4): 5-10.

Weisbuch, Gérard, Guillaume Deffuant, and Frédéric Amblard. 2005. Persuasion dynamics. *Physica A: Statistical Mechanics and its Applications* 353: 555-75.

Wikipedia (2017). *Banning policy*. <https://en.wikipedia.org/wiki/Wikipedia:Banning_policy> (Accessed January 14, 2017).

Wu, Qinyi, Danesh Irani, Calton Pu, and Lakshmish Ramaswamy. 2010. 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, 1797-1800. New York: ACM.

Yasseri, Taha, and János Kertész. 2013. Value production in a collaborative environment. *Journal of Statistical Physics* 151 (3-4): 414-39.

Yasseri, Taha, Anselm Spoerri, Mark Graham, and János Kertész. 2014. The Most Controversial Topics in Wikipedia: A Multilingual and Geographical Analysis. In *Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration,* ed. Pnina Fichman and Noriko Hara, 25-48. New York: Rowman & Littlefield.