

## **AN AGENT-BASED MODEL OF EDIT WARS IN WIKIPEDIA: HOW AND WHEN IS CONSENSUS REACHED**

Arun Kalyanasundaram

Wei Wei

Kathleen M. Carley

James D. Herbsleb

Institute for Software Research

Carnegie Mellon University

5000 Forbes Ave

Pittsburgh, PA 15213, USA

### **ABSTRACT**

Edit wars are conflicts among editors of Wikipedia, when editors repeatedly overwrite each other's content. Edit wars can last from a few days to several years before reaching consensus often leading to a loss of content quality. Therefore, the goal of this paper is to create an agent-based model of edit wars in order to study the influence of various factors involved in consensus formation. We model the behavior of agents using theories of group stability and reinforcement learning. We show that increasing the number of credible or trustworthy agents and agents with a neutral point of view decreases the time taken to reach consensus, whereas the duration is longest when agents with opposing views are in equal proportion. Our model can be used to study the behavior of members in online communities, and to inform policies and guidelines for participation.

### **1 INTRODUCTION**

Discussions are an integral part of the consensus building process. A discussion could be something as trivial as a group of friends deciding where to dine or as intricate as a group of engineers reaching an agreement on a product design. With the rise of the internet, such discussions are now taking place on social media platforms and online forums. A fitting example is Wikipedia<sup>1</sup>, an online encyclopedia community where thousands of editors discuss and create content collaboratively. However, not often do editors discussing a particular topic have the same point of view, especially if it is a controversial topic (Kittur et al. 2007). This leads to a scenario where editors repeatedly overwrite each other's content such that no useful content gets added. These are called *edit wars*<sup>2</sup> and they pose a serious challenge to the Wikipedia community. Edit wars can last from a few days to several years before a consensus is reached often leading to a loss of content quality, increase in trolls, newcomers dropping out and requires administrator's time and effort in monitoring them.

The goal of this paper is to model the process of edit wars in order to study various factors that influence consensus formation and to predict the duration to reach consensus. We use an agent-based approach to model the behavior of editors involved in edit wars. Our core idea is based on the principle that humans learn through repeated interactions with others. An interaction requires a person to perform some action and expect a response for that action. The person then "learns" to interact better based on the response received. If these interactions were discussions in a group, then such repeated interactions would eventually

---

<sup>1</sup><https://www.wikipedia.org/>

<sup>2</sup>[http://en.wikipedia.org/wiki/Wikipedia:Edit\\_warring](http://en.wikipedia.org/wiki/Wikipedia:Edit_warring)

result in a consensus. This approach is used in our model in two stages. First, editors are agents who interact with other agents by performing one of a predefined set of actions. The response that agents receive is modeled as a payoff that depends not only on the agent's actions but also on the actions of other agents. We develop a payoff matrix to compute the payoff of pairs of interacting agents. The use of a payoff matrix is a well known concept in game theory, and we design it such that it captures the underlying incentives of interactions among editors in Wikipedia. In the second stage, agents independently update their beliefs for the set of predefined actions using the Bush-Mostellar reinforcement learning algorithm (Bush and Mosteller 1955). The process is then repeated until agents reach a consensus, which is measured using an existing technique in literature (Erdamar et al. 2014).

In our experiments, we study the influence of the following factors on edit wars: the initial beliefs of agents, the credibility of agents as measured by their ability to make a trustworthy contribution and the distribution of agents with different points of view. We show that increasing the number of credible agents and increasing the number of agents with a neutral point of view decreases the duration to reach consensus. We also study different group compositions of editors and show that the duration is longest when two opposing views are in equal proportion. Our results can be used to inform policies and guidelines for participation in online communities and can be used to study the behavior of their members.

The rest of the paper is organized as follows. We briefly discuss the background and related work in Section 2. We provide the description of our Model in 3. We discuss our experiments and their results in Section 4. We describe the different ways of verifying and validating our model in Section 5. We end the paper with a conclusion of our findings, limitations of our model and scope for future work in Section 6.

## **2 BACKGROUND & RELATED WORK**

Consensus formation has been widely studied in multi-agent systems both from the context of coordination among autonomous mobile robots (Ren et al. 2005) and decision making among humans (Chiclana et al. 2013). While consensus among robots can be achieved using a protocol based approach (Ren et al. 2005) and even with limited communication (Savkin 2006), the dynamics of consensus building among humans is significantly complex (Cialdini and Goldstein 2004). One of the most widely accepted theories behind the process of consensus formation among humans is the theory of social influence (Cialdini and Goldstein 2004). The idea is that a person in a group is influenced to comply to the norms of the majority (Friedkin and Johnsen 1990). However, this theory does not always hold (Xie et al. 2011) and therefore, there are many different cognitive models of consensus formation. Measuring consensus is also an important part of building the model. The use of similarity measures is a popular technique in measuring consensus (Chiclana et al. 2013). In our model, we measure consensus using the approach proposed by Erdamar et al. (2014) and apply KL divergence (Kullback and Leibler 1951) as a similarity measure.

Kriplean et al. (2007) showed that the process of consensus formation in online communities such as Wikipedia is based on a set of policies, guidelines and shared mental models. Troitzsch (2009) showed that the act of repeated enforcement of policies leads to the emergence of social norms. However, online discussion forums on the other hand are a loosely coupled community with less restrictive norms for participation. In such cases, reaching consensus in discussions can be modeled as a network of interactions with other members (Sobkowicz 2013). Since the medium on which such interactions happen becomes irrelevant, the process can be modeled using cellular automation to study mutual interactions (Ono et al. 2005). Sometimes, members may not have direct interactions with each other such as on a social tagging website, yet reaching consensus is an important outcome and can be measured using similarity of actions (Robu, Halpin, and Shepherd 2009). However, a major limitation of existing approaches is they either take into account agent behavior or the environment, but not both. Therefore, one of our goals is to build a computational model of agent behavior that also considers the impact of the underlying environment where agents interact.

Discussions among members of online forums may not always reach consensus. Nevertheless, consensus is an important outcome for Wikipedia because the goal is not just to discuss but also to produce useful

content. However, there are situations when members disagree with each other to the extent that they repeatedly overwrite content and no useful content gets added. These are called edit wars and there has been a recent interest in studying and modeling them (Sumi et al. 2011). Yasserli et al. (2012) used the activity patterns of editors on Wikipedia to detect edit wars. They also categorized edit wars as those that either reach consensus or have a sequence of temporary consensus or are never ending wars. However, we model individual editor's cognition and behavior to study various factors that influence edit wars, which could also be used to inform policy decisions.

### 3 MODEL

Our model is based on the cyclical process of interaction and adaptation used in the theory of group stability (Carley 1991). The key idea is that both individual and group behavior can be determined by this cycle of interacting or exchanging information, learning or adapting their behavior, and interacting again and so on. In our model, we consider the actions performed by agents as interactions and the change in their beliefs as learning.

Wikipedia editors perform two basic actions: (1) A *commit* operation, which is adding or editing content in an article and (2) A *revert* operation, which is restoring an article to a previous version. Therefore, the evolution of a Wikipedia article can be seen as a series of commits and reverts. However, during edit wars, one group of editors revert the commits made by another group and vice versa, which often leads to a non-productive cycle where no new content gets added. Our model is based on this premise that edit wars in Wikipedia often have two sides to an issue. For example, one of the popular edit wars in Wikipedia took place in 2006 when the planet Pluto was re-categorized as a dwarf planet. Before this could be updated on Pluto's Wikipedia page, there was an edit war that involved two groups of editors, one supporting the change and the other opposing the change. Let's call these two sides: a) positive (+) and b) negative (-). Therefore, this gives us four possible actions an editor can perform, which are  $C^+$  and  $R^+$  : commit and revert that support the positive side, and  $C^-$  and  $R^-$  : commit and revert that support the negative side. This allows us to view any edit war as a stream of these four actions in different combinations.

Our model uses a simple game theoretic approach, where the editors are the agents in the model. Since editors are humans and the theories of bounded rationality applies to them, we hypothesize that an agent  $i$ 's payoff depends on the action of an agent immediately before and after  $i$ . The payoff mechanism is designed such that it promotes productive editing and suppresses the behaviors of edit warring based on the existing guidelines<sup>3</sup>.

Our model is turn based, where one turn is defined as a random sequence of actions performed by a set of agents. Each agent is allowed one action per turn and the payoffs are computed at the end of each turn. These payoffs are then used in a reinforcement learning algorithm (Bush and Mosteller 1955) for each agent to independently decide a new action for the next turn. The process is repeated until a consensus is reached, which is measured using KL divergence (Kullback and Leibler 1951).

#### 3.1 Payoff Mechanism

An agent's payoff depends on whether her neighboring agents support or refute her view. For example, if an agent  $i$  performs an action  $C^+$  and the next agent  $j$  performs an action  $C^-$ , then both agents have opposing views and since this fosters edit warring, both agents are dis-incentivized.

The amount of disincentive an agent receives depends on a) the type of action, b) agent's credibility and c) the neighboring agent's credibility. Each agent  $i$  has a level of credibility identified by  $\alpha_i$  in  $[0, 1]$ . This simply indicates how accurate or trustworthy an agent's action is. For example, on Wikipedia some editors may have a track record of making good quality contributions and are therefore, perceived as being more credible. This is important because Wikipedia editors often judge a contribution based on its editor's credibility. Since credibility can be viewed as the 'probability of making an accurate action', we use

---

<sup>3</sup>[http://en.wikipedia.org/wiki/Wikipedia:Edit\\_warring](http://en.wikipedia.org/wiki/Wikipedia:Edit_warring)

exponential families to express the dependence of payoff on two independent probabilities. The payoff  $P_i$  received by agent  $i$  due to a neighboring agent  $j$  is given by (1).

$$P_i = 1 - e^{f(\alpha_j) \frac{-\alpha_i}{1-\alpha_i}} \quad (1)$$

Where  $f(\alpha_j)$  depends on the action of  $i$  and  $j$ . Table 1 gives the payoff matrix, except for brevity we only show  $f(\alpha_j)$  since the rest of the equation remains unchanged. Therefore, Table 1 is used to compute the payoff received by  $i$  for a any combination of actions of  $i$  and  $j$ , by substituting  $f(\alpha_j)$  from Table 1 in (1)

Table 1: Payoff matrix - showing  $f(\alpha_j)$  of (1)

		Action of agent $j$			
		$f(\alpha_j)$	$C^+$	$C^-$	$R^+$
Action of agent $i$	$C^+$	$e^{-(1-\alpha)}$	$e^{-\alpha}$	$\alpha$	$1 - \alpha$
	$C^-$	$e^{-\alpha}$	$e^{-(1-\alpha)}$	$1 - \alpha$	$\alpha$
	$R^+$	$\alpha$	$1 - \alpha$	$e^{-\frac{1}{\alpha}}$	$e^{-\frac{1}{1-\alpha}}$
	$R^-$	$1 - \alpha$	$\alpha$	$e^{-\frac{1}{1-\alpha}}$	$e^{-\frac{1}{\alpha}}$

Since a neighboring agent can be either before or after in the sequence, the payoff of  $i$  is the sum of the payoff due to both its neighbors  $\mathbb{N}$ . Hence (1) is rewritten as shown in (2).

$$P_i = \sum_{j \in \mathbb{N}} 1 - e^{f(\alpha_j) \frac{-\alpha_i}{1-\alpha_i}} \quad (2)$$

### 3.2 Reinforcement Learning

The beliefs of agents are represented as a probability distribution of the above four actions. In each turn, agents select one of the four actions based on this probability distribution, which is simply the probabilities for choosing each of the four actions, such that the probabilities sum to one. Suppose, for an agent  $i$  the probability of choosing an action  $k$  is denoted by  $x_{i,k}$  and since our model has has four actions, then  $\sum_{k=1}^4 x_{i,k} = 1$ . At the start of the simulation, each agent is initialized with a particular distribution based on the experimental setup. However, these probabilities are updated at the end of each turn using the Bush-Mostellar reinforcement learning algorithm (Bush and Mosteller 1955). The principle behind the algorithm is that given an agent  $i$ 's payoff  $P_i(t)$  for an action  $k$  at the end of turn  $t$ , then the probability with which  $i$  will choose the action  $k$  in turn  $t + 1$  is given by (3), (4) and (5)

$$s_i(t) = \frac{P_i(t) - E_i}{\sup \forall_k \{ |U_i(k) - E_i| \}} \quad (3)$$

if  $s_i(t) \geq 0$ , then

$$x_{i,k}(t+1) = x_{i,k}(t) + \lambda s_{i,t}(1 - x_{i,k}(t)) \quad (4)$$

if  $s_i(t) < 0$ , then

$$x_{i,k}(t+1) = x_{i,k}(t) + \lambda s_{i,t}(x_{i,k}(t)) \quad (5)$$

Following is an explanation of the notations used in (3), (4) and (5).

- $E_i$  is the payoff agent  $i$  aspires to get. This can be fixed or varying depending on the action  $i$  performs. Estimating  $E_i$  that accurately captures the cognitive state of the agent  $i$  is a challenging problem. In our model, we assume that each agent's aspired payoff  $E_i$  is equal to its credibility  $\alpha_i$ . Since Wikipedia editors have a sense of how other editors perceive their credibility to be, the payoff expected by an editor would be a fraction of this perceived credibility.
- $U_i(k)$  is the maximum and minimum possible payoffs received by  $i$  for action  $k$ . From (1), we know that the payoff can only take values in the range  $[0, 1]$ , hence (3) is reduced to (6) as shown below.

$$s_i(t) = \frac{P_i(t) - E_i}{\sup\{|1 - E_i|, |0 - E_i|\}} \quad (6)$$

- $\lambda$  is the learning rate and is a value in  $[0, 1]$ . It indicates the degree to which agents update their probabilities for turn  $t + 1$  based on the payoff received in turn  $t$ . This is a tunable parameter, however typically a value of 0.5 is used.
- $x_{i,k}(t)$  and  $x_{i,k}(t + 1)$  are the probabilities of agent  $i$  to choose action  $k$  in turns  $t$  and  $t + 1$  respectively. The probabilities for the other three actions that the agent did not choose in turn  $t$  is given by (7)

$$\forall_{l \neq k} x_{i,l}(t + 1) = x_{i,l}(t) + \frac{(x_{i,k}(t + 1) - x_{i,k}(t))}{3} \quad (7)$$

#### 4 EXPERIMENTS AND RESULTS

The goal of our experiments is to study the influence of various factors on the process of consensus formation of edit wars. There are two factors that play a critical role during edit wars in Wikipedia: a) The actions that agents decide to choose and b) The credibility or trustworthiness of agents involved. Our first two experiments aim to evaluate the impact of each of these two factors independently on the duration of edit wars. Our third experiment aims at evaluating different group compositions, where a group is a distribution of agents with different points of view. Our experiments have two dependent variables: a) the duration of an edit war, and b) efficiency. The shorter the duration and higher the efficiency, the less detrimental the edit war is.

Duration is measured as the number of turns to reach consensus, a proxy for the time taken based on existing literature on simulations (Ono et al. 2005). The point of consensus is determined when the mean normalized KL Divergence ( $\overline{KLD}$ ) of two consecutive turns is less than a predefined value  $\epsilon$ , typically 0.01.

$$\overline{KLD} = \frac{\sum_{\forall i,j|i < j} KLD(i,j)}{(\#Agents_2) \cdot \sup\{\forall_{i,j}(KLD(i,j))\}} \quad (8)$$

Where,  $KLD(i, j) = \sum_{k=1}^{k=4} x_{i,k} * \log_e(\frac{x_{i,k}}{x_{j,k}})$ . The edit war is supposed to have reached a consensus when (9) is satisfied with a duration of  $t$  turns.  $\overline{KLD}_t$  is simply the mean normalized KL Divergence computed at the end of turn  $t$ .

$$|\overline{KLD}_{t-2} - \overline{KLD}_{t-1}| + |\overline{KLD}_{t-1} - \overline{KLD}_t| < 2\epsilon \quad (9)$$

Efficiency is measured as the ratio of sum of commit actions of all agents to the total number of actions. If  $\#C_i^+$  is the number of  $C^+$  actions made by agent  $i$  then, efficiency is given by (10). The reason we measure efficiency is that edit wars can be either constructive or destructive, and efficiency gives us a way to quantify this outcome. For example, an efficiency of less than 0.5 implies that there were more reverts made than commits, which effectively means a destructive edit war with no new content added. Therefore, higher the efficiency, the more constructive the edit war is.

$$Efficiency = \frac{\sum_i (\#C_i^+ + \#C_i^-)}{\sum_i (\#C_i^+ + \#C_i^- + \#R_i^+ + \#R_i^-)} \quad (10)$$

The experiments are performed using an agent based simulation software written in Java. The complete source code is made available online<sup>4</sup>. Each replication in our simulation starts by initializing the attributes of each agent depending on the experimental condition. Agents are then chosen at random to perform an action based on their probability distribution of actions. Each agent updates its probability distribution using the reinforcement learning algorithm. This process is repeated until the stopping condition as given in (9) is reached. The number of turns to terminate is recorded as the duration of the edit war for one replication. We run a total of one thousand replications for each experimental condition. We use a fixed agent size of 100 because when analyzed our outcome variables with agent size =  $\{100, 300, 500\}$ , we found the results to be independent of the number of agents. We also use a fixed learning rate of  $\lambda = 0.5$ , since we ran our experiments with  $\lambda = \{0.2, 0.5, 0.8\}$  and found that although the value of our outcome variables change with  $\lambda$ , the variance between different conditions of our experiments remain insensitive to  $\lambda$ .

#### 4.1 Experiment 1: Likelihood to Commit ( $L_C$ )

The goal of this experiment is to evaluate the effect of the agent’s initial beliefs on the duration of edit wars. The belief is represented as a probability distribution of the four possible actions. Hence, to study this effect we will have to vary four different variables. However, using a simple technique we can reduce this to just one variable. We note that in real world, agents in an edit war start by supporting one of the two sides of an issue. In other words, people often have strong opinions about the issue they are discussing, at least at the start of an edit war. Therefore, we can divide our agents into two groups depending on the side of the issue they support. These agents can either perform a *commit* or a *revert* action to support their side. Since these probabilities should sum to one, we can model our agents using only the likelihood of making a commit action as explained below.

Let  $L_C$  and  $L_R$  denote the likelihood to commit and revert respectively. Therefore, the following relations hold,

$$L_C = 1 - L_R \qquad L_C = L_{C^+} + L_{C^-} \qquad L_R = L_{R^+} + L_{R^-}$$

Suppose we randomly assign agents to the positive side of the issue, then based on our real world assumption,  $L_{C^-}$  and  $L_{R^-}$  should be infinitesimally small. Therefore, by using a single independent variable  $L_C$ , we can study the effect of agent’s initial beliefs. For each replication of an experimental condition, we generate a normal distribution with mean  $L_C$  and standard deviation 0.05. Each agent  $i$  is then randomly assigned  $L_C(i)$  from this distribution and the initial probability distribution of actions for  $i$  is computed depending on the side  $i$  was assigned at the start of the simulation. For example, suppose  $L_C = 0.2$  and  $L_C(i) = 0.23$  and  $i$  is assigned the positive side, then the probability distribution of actions for  $i$  is  $\{C^+, R^+, C^-, R^-\} = \{0.23, 0, 0, 0.77\}$ . However, in our implementation we assign a small value of 0.001 instead of zero for the non-supporting side. The value of standard deviation is chosen such that there is almost no overlap in the distributions of two conditions. Since our experiment has four conditions with an interval of 0.2 between each condition, a standard deviation of 0.05 will have about 95% of values in the interval of  $\pm 0.1$ . Table 2 shows the configuration of different variables used in this experiment.

Figure 1(a) shows the time taken (measured as the number of turns) to reach consensus for different values of  $L_C$ . The solid gray triangle on each box plot marks the mean value, and we see that the mean number of turns to reach consensus is non-decreasing with increase in  $L_C$ . This might seem counter intuitive that increasing the likelihood to commit also increases the time taken to reach consensus. However, in reality the reverts are a necessary evil, which serve to suppress biased opinions. However, this effect is only noticed for low values of  $L_C$ , that is the number of turns increases initially and then almost saturates for  $L_C > 0.4$ . We then performed a Tukey’s HSD (Honest Significant Difference) test to find if the difference between any two levels is statistically significant. The figure also shows the level of statistical significance between any two conditions. For example, in Figure 1(a), the difference in the number of turns between

<sup>4</sup>[https://github.com/arunk054/Modeling\\_Edit\\_Wars](https://github.com/arunk054/Modeling_Edit_Wars)

Table 2: Independent and Control variables in Experiment 1

Variable	Description	Nature / Number of values	Values used
$L_C$	Mean Likelihood to commit, normally distributed with a std. dev. of 0.05	4	0.2,0.4,0.6,0.8
$\alpha_i$	Credibility of agent $i$ , uniformly randomly distributed in $[0, 1]$	Random	-
$\lambda$	Learning rate	1	0.5
Agent population	Number of agents	1	100
# Replications	Number of replications per condition	1000	-
# Runs	Total number of independent runs	4000	-

$L_C = 0.4$  and  $L_C = 0.2$  is statistically significant with a mean difference of 3.7 turns and a 95% confidence interval of  $[1.7, 5.6]$ . Whereas, the difference between  $L_C = 0.6$  and  $L_C = 0.4$  is not statistically significant. On the other hand, the change in efficiency is statistically significant for every 0.2 increase in  $L_C$  as shown in Figure 1(b), with the highest difference being between  $L_C = 0.8$  and  $L_C = 0.4$  of 0.48 with a 95% confidence interval of  $[.45, 0.51]$ . Therefore, even if the time taken to reach consensus were to remain same, the efficiency increases with increase in  $L_C$ .

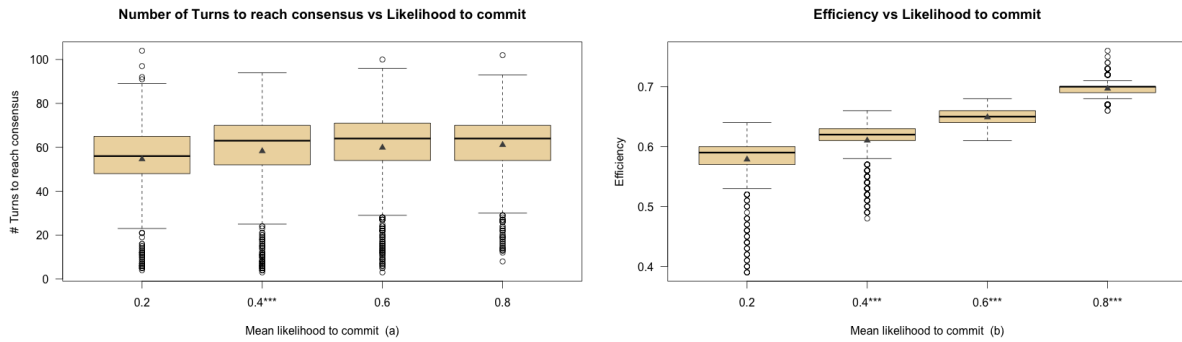


Figure 1: (a) Number of Turns to reach consensus. (b) Efficiency - for different values of  $L_C$ . (Solid triangle marks the mean.) \*\*\* $p < 0.001$

#### 4.2 Experiment 2: Credibility ( $\alpha$ )

The goal of this experiment is to evaluate the effect of agent’s credibility on the duration of edit wars. We can intuitively argue that an edit war among highly credible agents will last shorter and would be more efficient than if it had agents with lower credibility. In this experiment we aim to validate this relationship. We study four different  $\alpha$  values -  $\{0.2, 0.4, 0.6, 0.8\}$ , where each value is the mean of a normal distribution

with standard distribution 0.05. Therefore, at the start of each replication, each agent  $i$  is assigned an  $\alpha_i$  from the corresponding distribution of  $\alpha$  depending on the experimental condition. The other independent variables are: a)  $L_C$  is uniformly randomly distributed in  $[0, 1]$ , b)  $\lambda$  is 0.5, c) agent population size is 100. We run a thousand replications per condition.

Using Tukey’s HSD test we found that the difference in the number of turns between all subsequent levels is statistically significant, with the highest decrease being between  $\alpha = 0.8$  and  $\alpha = 0.6$  of 33.2 turns (about 58%) with a 95% confidence interval of  $[29.2, 37.1]$ . The box plot in Figure 2(a) shows that the number of turns decreases with increase in agent credibility, except for  $\alpha$  in the range  $[0.4, 0.6]$ . This has an intuitive explanation; humans find it easier to judge the actions of another person, when they know that the person is either highly trustworthy or less trustworthy. However, they have difficulty in making this judgment if the trustworthiness or credibility of the other person is uncertain. The same analogy can be used to explain the scenario in Figure 2(a), when agents have a credibility of around 0.5, it is almost a random toss of a coin to predict whether the actions of an agent are credible or not. Therefore this uncertainty slightly increases the number of turns to reach consensus when  $\alpha$  is between 0.4 and 0.6, which eventually decreases considerably when  $\alpha = 0.8$ .

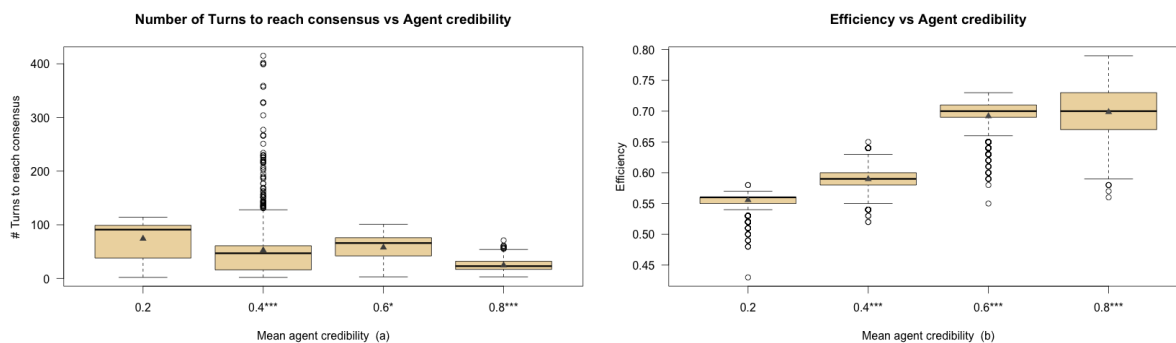


Figure 2: (a) Number of turns to reach consensus. (b) Efficiency - for different values of mean agent credibility ( $\alpha$ ). \*\*\* $p < 0.001$ , \* $p < 0.05$

Figure 2(b) shows the increase in efficiency between all subsequent levels of  $\alpha$  is statistically significant. The increase in efficiency is maximum (about 17%) when  $\alpha$  is in the range  $[0.4, 0.6]$ , and this complements our earlier argument that agents find it difficult to make judgments about other agent’s actions when  $\alpha$  is around 0.5. As a result of this uncertainty, agents have to add more content (i.e. make more commits) before they can reach a consensus. Although the efficiency at  $\alpha = 0.8$  is highest, the increase is only about 0.9% from the efficiency at  $\alpha = 0.6$ . This is because higher values of  $\alpha$  implies that consensus is reached faster, and hence involves lesser deliberations, which in turn causes the efficiency to reach a saturation.

### 4.3 Experiment 3: Group Composition

Our experiments so far have ignored one key aspect of edit wars in real world, that there is not always an equal proportion of agents from both sides of the issue. In our earlier experiments we assumed equal number of agents from both the positive and negative side. However, previous research on Wikipedia has shown that diversity of members leads to higher content quality (Arazy et al. 2011). On the other hand, Ono et al. (2005) showed that having more mediators or in our case editors with a neutral point of view reduces the time to reach consensus in arguments. Therefore, we identify three broad categories of editors: a) Positive campaigner (P), b) Negative campaigner (N) and c) Moderate (M). The goal of this experiment is to study the effect of group compositions on the duration of edit wars. A group here refers to a specific proportion of the three categories of editors at the start of an edit war. We choose a subset of group compositions such that it represents a few real world configurations and also captures a few extreme cases.



An agent is assigned one of these categories by simply modifying the probability distribution of actions  $\{L_C^+, L_R^+, L_C^-, L_R^-\}$  as follows, P:  $\{0.5, 0.5, 0, 0\}$ , N:  $\{0, 0, 0.5, 0.5\}$ , M:  $\{0.25, 0.25, 0.25, 0.25\}$ , where each value in this set maps to the corresponding action probability.

This experiment has six different group compositions and each one is a different experimental condition. Figure 3 shows the number of turns to reach consensus for all six conditions. The baseline condition has equal proportion of all three categories of editors. We compare the outcomes of all other conditions with this baseline. The second condition contains positive campaigners (P) and negative campaigners (N) in equal proportion but no moderates (M). The third condition has very high percentage positive campaigners (90%) and no moderates, therefore, the ratio of P:N is 9:1. The fourth condition is essentially the same as the third condition except the ratio of P:N is 1:9. Evaluating similar cases also allows us to validate our model. In fifth condition, we evaluate the effect of moderates, where half of the agent population are moderates and the rest half has equal proportion of P and N. In the final condition, we evaluate a hypothetical case where all editors are moderates. The credibility  $\alpha_i$  of agents is uniformly randomly distributed. We use a constant learning rate  $\lambda = 0.5$  and the number of agents is fixed at one hundred. The number of replications per condition is one thousand.

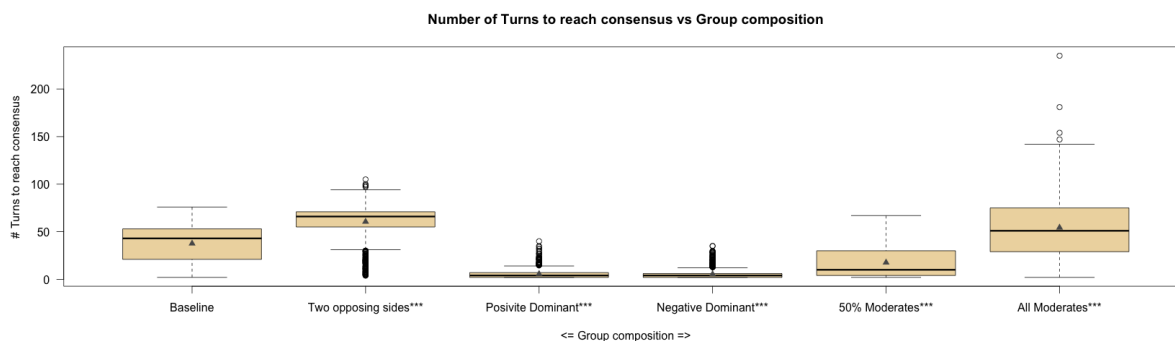


Figure 3: Number of turns to reach consensus for different group compositions.

We found that all group compositions had statistically significant difference compared to the baseline. In addition, a Tukey's HSD test showed statistically significant difference between all pairs of group compositions except between the positive and negative dominant groups. Since the two groups are essentially similar, the result bolsters the validity of our model. As shown in Figure 3, these two groups also take the shortest time to reach consensus, probably because it is easy to suppress opposition when the number of supporters on one side is highly skewed. The *Two opposing views* condition takes longer to reach consensus than the baseline, this shows that the presence of moderates can help reach consensus faster. Furthermore, this effect is more pronounced in the *50% moderates* condition, which takes about fifty-four percent less turns than the baseline. However, if the agent population consists of *All Moderates*, it takes longer than the baseline. Since a group made up of only moderates is an extremely unlikely scenario in an edit war, the result could mean that editors are just involved in a lengthy process of deliberation and not an edit war. We do not show the efficiency of these conditions since it follows a similar pattern as the number of turns.

#### 4.4 Discussion

Results of our experiments show that the mean agent credibility  $\alpha$  has a greater influence on the time taken to reach consensus than mean likelihood to commit  $L_C$ , whereas efficiency is influenced by  $L_C$  more than  $\alpha$ . This suggests that an agent's actions are not as important as the credibility of the agent performing that action. On the other hand, efficiency appears to depend on both  $L_C$  and  $\alpha$ . However, efficiency increases linearly with increase in  $L_C$ , but follows a sigmoidal growth curve with increase in  $\alpha$ . This indicates that

having agents with high credibility increases their productivity but beyond a certain level of credibility there is no noticeable increase in efficiency.

We found that even when  $L_C$  and  $\alpha$  remain same across different populations of agents, there is a significant difference in the outcome measures depending on the fraction of agents that support a given side. It turns out, when this fraction is close to one, edit wars take an extremely short time to reach consensus, almost as if an edit war never existed. This is also what one would expect in the real world, when a particular decision is made unanimously because an overwhelming majority support it. In such cases an issue no longer remains contentious. However, we found that the duration of edit wars is maximum when both sides are equally represented. This happens in the real world when an issue is highly contentious that there is often an equal mix of editors from both sides of the issue. Wikipedia administrators should therefore, monitor the composition of agents in an edit war and watch out for such signs. On the other hand, when the agent population contains moderates, which are agents with a neutral opinion, the duration of edit wars decreases with increase in the fraction of moderates. This is also the reason why Wikipedia's guidelines for contribution encourage its editors to have a neutral point of view when editing an article.

## 5 VERIFICATION AND VALIDATION

Based on the validation techniques proposed by Sargent (2005), we discuss the following approaches as applied to our model: a) conceptual validity, b) model verification and c) operational validity. Conceptual validation implies that the theories or concepts the model is based on accurately characterizes real world. Our model is based on the ideas that consecutive actions supporting a similar viewpoint should be incentivized and that commits receive higher incentives than reverts. Therefore, the conceptual validity of our model can be ascertained if these concepts hold in real world. Wikipedia's guidelines<sup>5</sup> on achieving consensus state that consensus is achieved by incorporating the legitimate concerns of all editors and is not the result of voting or unanimity. This is in line with our approach of the use of autonomous agents and collectively evaluating consensus. In addition, Wikipedia's 3RR (three-revert rule)<sup>6</sup> policy prevents editors from making more than three reverts to a page in a given time period, which substantiates our notion of incentivizing commits higher than reverts.

Model verification refers to the accuracy of the model's implementation. We used software engineering principles such as object oriented design, modularity and unit testing to develop our Java based implementation. We use internal consistency checks, which is to show that certain invariants identified within the model are satisfied by our implementation. For instance, in experiment 3 we did not find a statistically significant difference between the two similar conditions, which is one form of internal consistency. Moreover, the amount of variability across different replications of a condition is below the acceptable limits.

The operational validity of a model is primarily done using plots or statistical tests showing expected model behavior. For example, when an agent's mean likelihood to commit ( $L_C$ ) is increased, then according to (10) the efficiency of the edit war should increase irrespective of other conditions. This can be observed from the results of our experiment 1 as shown in Figure 1(b), where the mean efficiency is a strictly increasing function of  $L_C$ .

Furthermore, the external validity of our model can be achieved by matching the results of the model with the real world, which often requires calibrating the parameters of the model. In this case we would like to use the data from an edit war in Wikipedia and use our model to accurately predict the time taken for the edit war to reach a consensus. We choose an edit war that has already reached consensus, and use the revision history to identify all editors and their actions up to half way through the edit war. This will be assumed as the start of the edit war and our goal is to use our model to predict the time taken for the second half. Each action in the first half of the edit war is annotated as one of the four possible actions of our model. Each editor is assigned a likelihood to commit ( $L_C(i)$ ) by finding the ratio of commits to the sum of commits and reverts made by the editor in the past anywhere on Wikipedia. The probability

<sup>5</sup><http://en.wikipedia.org/wiki/Wikipedia:Consensus>

<sup>6</sup>[http://en.wikipedia.org/wiki/Wikipedia:Edit\\_warring#The\\_three-revert\\_rule](http://en.wikipedia.org/wiki/Wikipedia:Edit_warring#The_three-revert_rule)

distribution of actions is computed based on the annotated actions for that editor. Given this input, we use our model to compute a 95% confidence interval of the mean number of turns and efficiency. There are a number of ways by which the time taken on Wikipedia can be converted to the number of turns our model outputs. One approach is by devising a threshold for the number of actions on Wikipedia to be counted as one turn. We evaluated our model's accuracy with data from the edit war on planet Pluto's reclassification as a dwarf planet that took place between 23<sup>rd</sup> and 27<sup>th</sup> August 2006 and involved 197 non-anonymous editors. However, this required considerable calibration of the parameters of our model, and therefore, a scope for future work is to improve the data conversion techniques between our model and the real world.

## 6 CONCLUSION

It is human tendency to have strong opinions, which often leads to heated arguments over certain issues. When such arguments take place on online platforms where user generated content is the only form of content creation, these arguments turn into what are known as *edit wars*. These edit wars pose a serious challenge to the growth and maintenance of online communities such as Wikipedia. Therefore, it is important to develop methods to automatically monitor edit wars and decide when to intervene. We showed that our model can be used to predict the time taken for a given edit war to reach consensus. We used our model to study the various factors that influence the duration and efficiency of edit wars. Our model can therefore, be used by administrators of Wikipedia and maintainers of other online communities to make policy decisions on participation and contribution.

Although our model can explain many of the real world phenomena associated with an edit war, it makes a few assumptions for brevity and simplicity. We do not consider an agent's ability to learn across multiple pages simultaneously and we assume an agent's rate of learning to be constant throughout the process. This requires sophisticated models of human cognition, a topic of interest in an emerging field known as cognitive social simulation. Another assumption our model makes is to treat all commit actions the same, however, the content of a commit could offer additional insights. For example, we could leverage techniques from areas like Natural Language Processing (NLP) to take into account the semantics of the content. A limitation of our model is that unlike other related works (Yasseri et al. 2012), it does not detect edit wars instead it assumes a given condition as an edit war and predicts the desired outcomes. We believe the core concepts of our approach can be broadly applied to the study of member behavior in online communities and aid in the design of improved guidelines and policies.

## REFERENCES

- Arazy, O., O. Nov, R. Patterson, and L. Yeo. 2011, April. "Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict". *J. Manage. Inf. Syst.* 27 (4): 71–98.
- Bush, R. R., and F. Mosteller. 1955. "Stochastic models for learning".
- Carley, K. 1991. "A theory of group stability". *American Journal of Sociology*:331–354.
- Chiclana, F., J. T. Garca, M. del Moral, and E. Herrera-Viedma. 2013. "A statistical comparative study of different similarity measures of consensus in group decision making". *Information Sciences* 221 (0): 110 – 123.
- Cialdini, R. B., and N. J. Goldstein. 2004. "Social Influence: Compliance and Conformity". *Annual Review of Psychology* 55 (1): 591–621. PMID: 14744228.
- Erdamar, B., J. L. García-Lapresta, D. Pérez-Román, and M. Remzi Sanver. 2014, May. "Measuring Consensus in a Preference-approval Context". *Inf. Fusion* 17:14–21.
- Friedkin, N., and E. Johnsen. 1990. "Social-influence and Opinions". *Journal of Mathematical Sociology* 15 (3-4): 193–205.
- Kittur, A., B. Suh, B. A. Pendleton, and E. H. Chi. 2007. "He Says, She Says: Conflict and Coordination in Wikipedia". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07*, 453–462. New York, NY, USA: ACM.

- Kriplean, T., I. Beschastnikh, D. W. McDonald, and S. A. Golder. 2007. "Community, Consensus, Coercion, Control: Cs\*W or How Policy Mediates Mass Participation". In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, GROUP '07, 167–176. New York, NY, USA: ACM.
- Kullback, S., and R. A. Leibler. 1951, 03. "On Information and Sufficiency". *Ann. Math. Statist.* 22 (1): 79–86.
- Ono, K., M. Harao, and K. Hirata. 2005, July. "Multi-agent based modeling and simulation of consensus formations in arguments". In *Information Technology and Applications, 2005. ICITA 2005. Third International Conference on*, Volume 1, 264–267 vol.1.
- Ren, W., R. Beard, and E. Atkins. 2005, June. "A survey of consensus problems in multi-agent coordination". In *American Control Conference, 2005. Proceedings of the 2005*, 1859–1864 vol. 3.
- Robu, V., H. Halpin, and H. Shepherd. 2009, September. "Emergence of Consensus and Shared Vocabularies in Collaborative Tagging Systems". *ACM Trans. Web* 3 (4): 14:1–14:34.
- Sargent, R. G. 2005. "Verification and Validation of Simulation Models". In *Proceedings of the 37th Conference on Winter Simulation*, WSC '05, 130–143: Winter Simulation Conference.
- Savkin, A. V. 2006. "The problem of coordination and consensus achievement in groups of autonomous mobile robots with limited communication". *Nonlinear Analysis: Theory, Methods and Applications* 65 (5): 1094 – 1102. Hybrid Systems and Applications (4) Hybrid Systems and Applications (4) Hybrid Systems and Applications.
- Sobkowicz, P. 2013, 12. "Quantitative Agent Based Model of User Behavior in an Internet Discussion Forum". *PLoS ONE* 8 (12): e80524.
- Sumi, R., T. Yasserli, A. Rung, A. Kornai, and J. Kertész. 2011, July. "Edit wars in Wikipedia". *ArXiv e-prints*.
- Troitzsch, K. G. 2009. "Perspectives and Challenges of Agent-based Simulation As a Tool for Economics and Other Social Sciences". In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '09, 35–42. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
- Xie, J., S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski. 2011, July. "Social consensus through the influence of committed minorities". *Physical Review E* 84 (1): 011130.
- Yasserli, T., R. Sumi, A. Rung, A. Kornai, and J. Kertész. 2012, 06. "Dynamics of Conflicts in Wikipedia". *PLoS ONE* 7 (6): e38869.

## AUTHOR BIOGRAPHIES

**ARUN KALYANASUNDARAM** is a Ph.D. student in the Institute for Software Research at Carnegie Mellon University. His research interests are large scale distributed collaborations, agent-based modeling, social network analysis, and machine learning. His email address is [arunkaly@cs.cmu.edu](mailto:arunkaly@cs.cmu.edu).

**WEI WEI** is a Ph.D. candidate in the Institute for Software Research at Carnegie Mellon University. His research interests include multi-agent systems, dynamic network analysis, data mining, and geo-temporal network dynamics. His email address is [weiwei@cs.cmu.edu](mailto:weiwei@cs.cmu.edu).

**KATHLEEN M. CARLEY** is a professor in the Institute for Software Research at Carnegie Mellon University. Her research areas are dynamic network analysis, computational social and organization theory, and information diffusion. Her email address is [kathleen.carley@cs.cmu.edu](mailto:kathleen.carley@cs.cmu.edu).

**JAMES D. HERBSLEB** is a professor in the Institute for Software Research at Carnegie Mellon University. His research interests lie in the intersection of software engineering, computer-supported cooperative work, and socio-technical systems. His email address is [jdh@cs.cmu.edu](mailto:jdh@cs.cmu.edu).