Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

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, 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* (Sumi et al. 2011) 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 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 (assign a probability for each of the predefined actions) using the Bush-Mostellar reinforcement learning algorithm (Bush and Mosteller 1953). 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 Section 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 an 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 the agents interact.

Discussions among members in online forums may not always reach a 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). Yasseri 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.

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 theory 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 (http://en.wikipedia.org/wiki/Wikipedia:Edit_warring).

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 1953) 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 adjacent 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 adjacent agent's credibility. Each agent i has a level of credibility identified by α_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 exponential families to express payoff as a function of two independent probabilities. The payoff P_i received by agent i due to an adjacent agent j is given by (1).

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

Where $f(\alpha_j)$ depends on the action of i and j. Table 1 gives the payoff matrix, however, for brevity we only show $f(\alpha_j)$ since the rest of the payoff function 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^+	R^-
Action of agent i	C^+	$e^{-(1-\alpha_j)}$	$e^{-\alpha_j}$	α_j	$1-\alpha_j$
	C^-	$e^{-\alpha_j}$	$e^{-(1-\alpha_j)}$	$1-\alpha_j$	α_j
	R^+	α_j	$1-\alpha_j$	$e^{-\frac{1}{\alpha_j}}$	$e^{-\frac{1}{1-\alpha_j}}$
	R^-	$1-\alpha_j$	α_j	$e^{-\frac{1}{1-\alpha_j}}$	$e^{-\frac{1}{\alpha_j}}$

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

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

3.2 Reinforcement Learning

The beliefs of agents are represented as a vector of probabilities of the above four actions such that the probabilities sum to one. In each turn, agents randomly select one of the four actions weighted by the belief vector. Suppose, for an agent i the probability of choosing an action k is denoted by $x_{i,k}$ and since our model has four actions, we have $\sum_{k=1}^{k=4} x_{i,k} = 1$. At the start of the simulation, each agent is initialized with a particular set of probabilities 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 1953). 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)

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

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

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

if $s_i(t) < 0$, then

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

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

α_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 payoff can only take values in $[0,1]$, hence (3) is reduced to (6) as shown below.

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

- λ 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.
- $x_{i,k}(t)$ 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)

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

4 EXPERIMENTS AND RESULTS

The goal of our experiments is to study the influence of various factors on the process of consensus formation in edit wars. There are two factors that play a critical role during edit wars in Wikipedia: a) The initial beliefs of agents and b) the credibility of agents. 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 comparing various real world scenarios that include evaluating different group compositions, where a group is a distribution of agents with different beliefs. Our experiments have two dependent variables: a) the duration of an edit war, and b) productivity. The shorter the duration and higher the productivity, 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 ε, typically 0.01.

$$\overline{KLD} = \frac{\sum_{\forall i,j | i<j} KLD(i,j)}{\binom{\#Agents}{2}.sup\{\forall_{i,j}(KLD(i,j))\}} \tag{8}$$

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.

$$|\overline{KLD}_{t-2} - \overline{KLD}_{t-1}| + |\overline{KLD}_{t-1} - \overline{KLD}_t| < 2\varepsilon \tag{9}$$

Productivity 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, productivity is given by (10). The reason we measure productivity is that edit wars can be either constructive or destructive, and productivity gives us a way to quantify this outcome. For example, productivity 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 productivity, the more constructive the edit war is.

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

The experiments are performed using an agent based simulation software written in Java. The complete source code is made available at https://github.com/arunk054/Modeling_Edit_Wars. 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 1000 replications per experimental condition since we found that there is less than one percent difference in both the outcome variables even with ten thousand replications. 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 λ, the variance between different conditions of our experiments remain insensitive to λ.

4.1 Experiment 1: Likelihood to Commit (L_C)

The goal of this experiment is to evaluate the effect of agent's initial beliefs on the duration of edit wars. Since belief is represented as a probability distribution of the four possible actions, we would need four different independent variables in this experiment. However, 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, an agent's initial belief only depends on the *commit* and *revert* probabilities of the side they support. Since the probabilities should sum to one we show that a single variable can be used to study the effect agent's initial beliefs under this real-world assumption.

Let L_C and L_R denote the likelihood to commit and revert respectively, then the following relations hold,

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

Suppose an agent supports the positive side then based on our assumption the value of L_{C^-} and L_{R^-} must be infinitesimally small (of the order 10^{-3}). Since both sides are weighted equally in our model, the side an agent supports is irrelevant in this experiment. Therefore, each agent is randomly assigned one of the two sides at the start of the experiment. We evaluate the experiment with four different values of L_C as shown in Table 2. For each replication of an experimental condition, we generate a normal distribution with L_C as the mean and a 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 $\{L_{C^+}, L_{R^+}, L_{C^-}, L_{R^-}\} = \{0.229, 0.001, 0.001, 0.769\}$. The value of standard deviation is chosen such that there is almost no overlap in the distributions of any 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 ± 0.1. In addition, our normal distribution generator truncates values less than zero and greater than one.

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 since they 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 $L_C > 0.4$. The figure also shows the level of statistical significance between any two consecutive conditions. For example, in Figure 1(a), the difference in the number of turns between $L_C = 0.4$ 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, change in productivity 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 $L_C = 0.6$ of 0.048 with a 95% confidence interval of $[0.045, 0.051]$. Although the increase in productivity with increase in L_C is obvious, the results

Table 2: Independent and Control variables in Experiment 1

Variable	**Description**	**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
α_i	Credibility of agent i, uniformly randomly distributed in $[0,1]$	*Random*	-
λ	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	-

help in the following two ways; first, it suggests that the relationship is not linear and that there is a greater increase in productivity for higher values of L_C. Second, the simplicity of the experiment allows us to verify whether the model behaves as expected under predictable conditions.

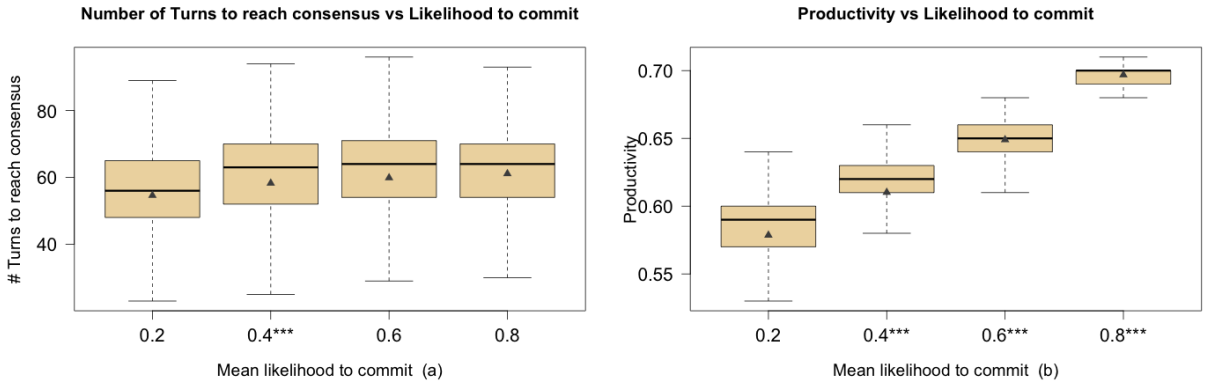

Figure 1: (a) Number of Turns to reach consensus and (b) Productivity for different values of L_C. (Solid triangle marks the mean.) ***p < 0.001

4.2 Experiment 2: Credibility (α)

The goal of this experiment is to evaluate the effect of agent's credibility on the duration of edit wars. Although we could intuitively argue that an edit war among highly credible agents will last shorter and be more productive than among agents with lower credibility, our experiment shows a more complex relationship. We study the effect of α in the range $[0.2, 0.8]$ with a 0.1 increment. For a given value of α, each agent i is assigned an α_i from a normal distribution with a mean α and a standard deviation 0.02. This ensures that the distributions of two consecutive values of α have minimal overlap since 95% of values will lie within ± 0.04 with a standard deviation of 0,02. The other independent variables are: a) L_C is uniformly randomly distributed in $[0,1]$, b) λ is 0.5, c) agent population size is 100.

Figure 2(a) shows that the number of turns to reach consensus increases with increase in credibility until it reaches a peak at $\alpha = 0.5$ and then continues to decrease. There are two results here that we found are surprising, first, edit wars are longer when α is about half and second, edit wars are shorter among

agents with low α. This behavior 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 behavior in Figure 2(a). When an agent has a credibility of around 0.5, it is almost a random toss of a coin to predict whether an action performed by the agent is accurate or not. Therefore this uncertainty increases the number of turns to reach consensus. On the other hand, Figure 2(b) shows that productivity increases monotonically with increase in credibility, and shows a sub-linear relationship. Hence even though agents with high credibility take about the same time to reach consensus as agents with low credibility, they are however, considerably more productive. We also observe that the slope of the curve is maximum for α in the range $[0.4, 0.6]$, which is because this is the region of uncertainty as explained before and therefore, involves a spike in the number of discussions. This also complements our results observed in Figure 2(a) that edit wars last longer when α is in this region of uncertainty. In both Figures 2(a) and 2(b) we use LOESS regression to fit a curve with the data.

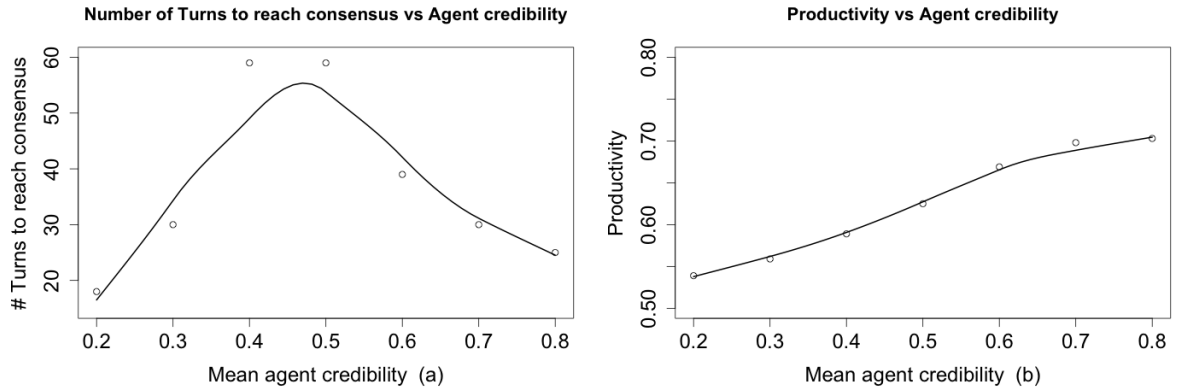

Figure 2: (a) Number of turns to reach consensus, (b) Productivity - for different values of α.

4.3 Experiment 3: Group Composition

Our experiments so far have ignored one key aspect of edit wars in real world; 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 some of the real world scenarios. 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: $\{0.5, 0, 0.5, 0\}$, where each value in this set maps to the corresponding action probability. A moderate (M) is defined as someone with a neutral point of view and a tendency to promote productivity.

Figures 3(a) and 3(b) show the number of turns to reach consensus and productivity respectively for the seven different group compositions. The whiskers on the box plots are not shown for clarity. The *baseline* has equal proportion of all three categories of editors. We compare the outcomes of all other groups with this baseline. The second group, *opposing sides* contains positive campaigners (P) and negative campaigners

(N) in equal proportion but no moderates (M). The third (*positive dominant*) and fourth (*negative dominant*) groups have a very high percentage (90%) of P and N respectively with no M. The last three groups contain 50%, 75% and 90% M respectively and are used to evaluate the effect of M in the agent population. In these three groups, the remaining agent population is comprised of equal proportion of P and N. We do not evaluate a group composition with all moderates because that is a hypothetical scenario since edit wars happen when at least some agents have strong opinions towards a particular side of the issue. The credibility α_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.

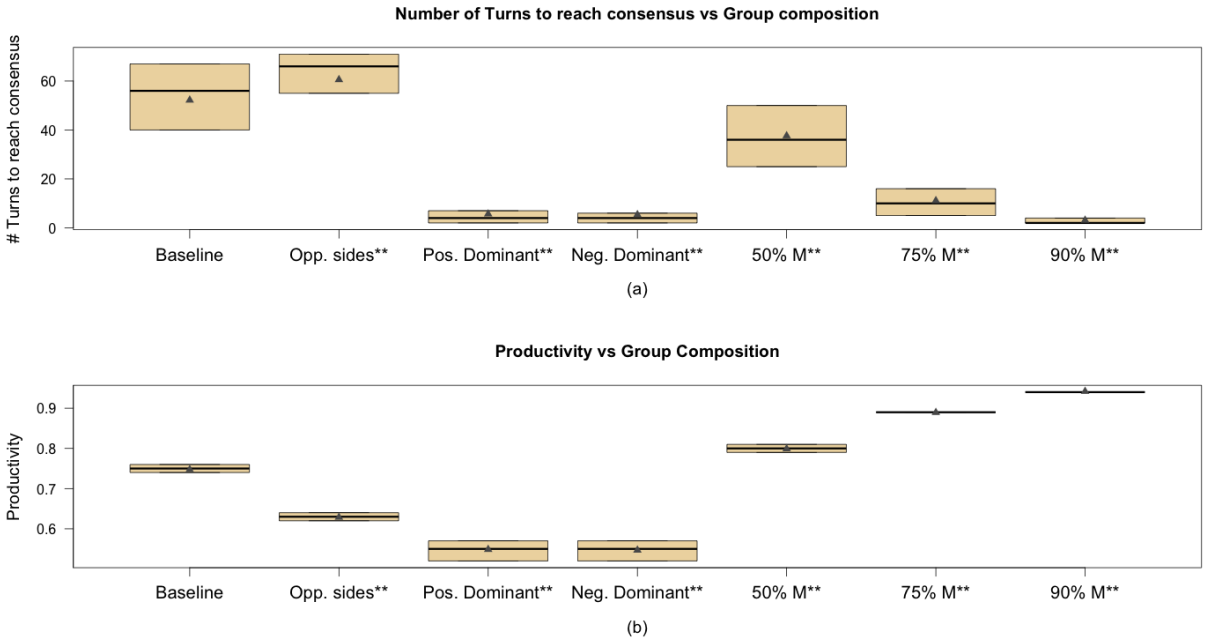

Figure 3: Number of turns to reach consensus for different group compositions.

We found that all groups were statistically significantly different compared to baseline for both the outcome variables. In addition, a Tukey's HSD test showed statistically significant difference between all pairs of group compositions except between the *positive dominant* and *negative dominant* groups, since the two compositions are essentially the same with a different polarity. These two groups also take the shortest time to reach consensus because the number of agents supporting one side is highly skewed, it is easy to suppress opposition. On the other hand, the *opposing sides* group takes the longest to reach consensus because intuitively we know that there is more deliberation involved when both sides are equally represented. The results of the last three groups show that higher the proportion of moderates, the shorter and more productive the edit wars are. We further verify this claim by incrementally varying the proportion of moderates from zero to ninety percent and observe a consistent trend.

4.4 Discussion

Results of our experiments show that the mean agent credibility α has a greater influence on the time taken to reach consensus than the mean likelihood to commit L_C. On the other hand, productivity appears to depend on both L_C and α. However, we observe that productivity has a super-linear growth with L_C, whereas it has a sub-linear growth with α. This indicates that having agents with high credibility increases the productivity of an edit war but the increase is not noticeable beyond a certain level of credibility.

We found that even when L_C and α 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 that 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 and enforce policies to curb reverting.

5 VERIFICATION AND VALIDATION

Based on the validation techniques proposed by Sargent (2007), 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 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) (http://en.wikipedia.org/wiki/Wikipedia:Edit_warring#The_three-revert_rule) 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 productivity 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 productivity 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 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 productivity. 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^{rd} and 27^{th} August 2006 and involved 197 non-anonymous editors. However, since this required considerable calibration of the parameters of our model, improving the parameter conversion techniques between our model and the real world is a scope for future work .

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 productivity 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 work (Yasseri et al. 2012), it does not detect edit wars and instead assumes a given condition as an edit war and predicts the two outcomes - duration and productivity. We believe 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. "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. 1953. "A Stochastic Model with Applications to Learning". *The Annals of Mathematical Statistics* 24 (4): 559–585.

Carley, K. 1991. "A Theory of Group Stability". *American Journal of Sociology* 56 (3): 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. "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. "On Information and Sufficiency". *Ann. Math. Statist.* 22 (1): 79–86.

Ono, K., M. Harao, and K. Hirata. 2005. "Multi-agent Based Modeling and Simulation of Consensus Formations in Arguments". In *Proceedings of the Third International Conference on Information Technology and Applications*, Volume 1, 264–267.

Ren, W., R. Beard, and E. Atkins. 2005. "A Survey of Consensus Problems in Multi-Agent Coordination". In *Proceedings of American Control Conference*, Volume 3, 1859–1864.

Robu, V., H. Halpin, and H. Shepherd. 2009. "Emergence of Consensus and Shared Vocabularies in Collaborative Tagging Systems". *ACM Trans. Web* 3 (4): 14:1–14:34.

Sargent, R. G. 2007. "Verification and Validation of Simulation Models". In *Proceedings of the 39th Conference on Winter Simulation*, edited by J. Tew, R. Barton, S. Henderson, B. Biller, M. Hsieh, and J. Shortle, 124–137. Piscataway, New Jersey: IEEE Press.

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.

Sobkowicz, P. 2013. "Quantitative Agent Based Model of User Behavior in an Internet Discussion Forum". *PLoS ONE* 8 (12): e80524.

Sumi, R., T. Yasseri, A. Rung, A. Kornai, and J. Kertesz. 2011. "Edit Wars in Wikipedia". In *Privacy, Security, Risk and Trust (PASSAT) and IEEE Third Inernational Conference on Social Computing (SocialCom)*, 724–727: IEEE.

Troitzsch, K. G. 2009. "Perspectives and Challenges of Agent-based Simulation As a Tool for Economics and Other Social Sciences". In *Proceedings of The Eighth International Conference on Autonomous Agents and Multiagent Systems*, 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. "Social Consensus Through the Influence of Committed Minorities". *Physical Review E* 84 (1): 011130.

Yasseri, T., R. Sumi, A. Rung, A. Kornai, and J. Kertsz. 2012. "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.

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.

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.

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.