

# A Bounded Confidence Approach to Understanding User Participation in Peer Production Systems

Giovanni Luca Ciampaglia\*  
ciampagg@usi.ch

September 29, 2018

## Abstract

Commons-based peer production does seem to rest upon a paradox. Although users produce all contents, at the same time participation is commonly on a voluntary basis, and largely incentivized by achievement of project's goals. This means that users have to coordinate their actions and goals, in order to keep themselves from leaving. While this situation is easily explainable for small groups of highly committed, like-minded individuals, little is known about large-scale, heterogeneous projects, such as Wikipedia.

In this contribution we present a model of peer production in a large online community. The model features a dynamic population of bounded confidence users, and an endogenous process of user departure. Using global sensitivity analysis, we identify the most important parameters affecting the lifespan of user participation. We find that the model presents two distinct regimes, and that the shift between them is governed by the bounded confidence parameter. For low values of this parameter, users depart almost immediately. For high values, however, the model produces a bimodal distribution of user lifespan. These results suggest that user participation to online communities could be explained in terms of group consensus, and provide a novel connection between models of opinion dynamics and commons-based peer production.

## 1 Introduction

In the past decade mass collaboration platforms have become common in several production contexts. The term *commons-based peer production* has been coined to refer to a broad range of collaborative systems, such as those used for producing software, sharing digital content, and organizing large knowledge repositories, however, seem to be based upon a paradox. In wikis, there is a link between quality and cooperation [30], but, at the same time, contribution is voluntary, based on non-monetary incentives

---

\*Accepted to Socinfo 2011. The final publication is available at [www.springerlink.com](http://www.springerlink.com)

[23, 26]. For small teams, this might not be a problem. In large scale wikis, where low access barriers are necessary to attract vast masses of contributors [8], and where expert users play a crucial role in maintenance and governance [2], user retention becomes instead crucial [10].

An established fact about participation to online groups is the *preferential behavior* of users, that is, a newcomer’s long-term participation can be predicted by the outcome of his or her early interactions [1, 21]. This could be explained in terms of Socialization theory [6], as users assess the willingness of the community to accept them and vice versa. It is also true, however, that quality assessment of the produced contents, and in particular comparison of the objectives of an individual with those of the community, is important in determining user participation [17]. This could be explained as a form of day-to-day coordination or group consensus taking place among editors [15].

In this paper we study user participation as a collective social phenomenon [4]. Other models of peer-production have been proposed already, for example for social information filtering platforms [13]. Here, we draw specifically from the modeling work on models of social influence under bounded confidence [9, 12].

Let us consider a community of users engaged in editing a collection of pages, e.g. Wikipedia. Pages are denoted by a certain number of features upon which users can find themselves in agreement or not. For example, let us consider the writing style of pages. Users try to modify pages according to their objectives, i.e. using their own style. At the same time, by interacting with contents, users can be also influenced by the style of other users. This reciprocal influence, however, happens only to a certain extent, that is, only when user and page (that is, their styles) are similar enough. Vandals, to illustrate with the same example, might not be interested in learning the encyclopedic writing style. In the context of social psychology this phenomenon is known as *bounded confidence*, and is regarded as a general feature of human communication within groups that try to reach consensus [12]. It can be also seen as a form of herding in that people are influenced by the social context they are in [22].

The population of users in our model is dynamic, with user departure determined endogenously by the social influence process. Although others have already studied Deffuant’s model to a dynamic population [3], here we explicitly link the process of social influence to user participation.

We implemented these ideas in an agent-based model of a peer production system. In this model, several factors affect the behavior of agents, such as user activity, content popularity, and community growth. To understand what factors are truly important for the resulting dynamics of user participation, we performed a factor screening using global sensitivity analysis.

## 1.1 Related work

The subject of user participation in mass collaboration systems has been already touched by several authors, for example on social networking sites [16], and knowledge sharing platforms [31]. A “momentum” law has been proposed for the distribution of user life edits of inactive users [29]. The distribution of user account lifespans has been shown to decay with a heavy tail, and a power-law model has been proposed after this obser-

vation [11]. Empirical data from Wikipedia, however, seem to support a super-position of different regimes [7]; a feature of the model we present here is indeed a bimodal distribution of user lifespans. In the context of wikis and other free open source initiatives some authors have used survival analysis to outline the differences between different communities, [20] but this modeling technique is not suited to understand the connection between social influence, group coordination, and user retention. We advocate the need to explicitly model such processes explicitly.

The paper is organized as follows: in Sec. 2 we introduce our model of peer production; in Sec. 3 we briefly describe global sensitivity analysis and Gaussian Processes, the two statistical techniques we used for the factor screening study; in Sec. 4 we present our main results and we discuss them in Sec. 5.

## 2 An agent-based model of commons-based peer production

In this section we introduce our model of peer production. While we make explicit use of the terminology of wiki platforms (e.g. “users” who “edit pages”) we stress that ours is a general model of consensus building in a dynamic bipartite population, and not merely a description of a wiki platform. We also stress that in our model the state of agents may not necessarily represent an opinion in the classic sense of other studies of opinion dynamics, i.e. extremes of the spectrum do not necessarily denote – say – political extremism, nor we speak of “moderates” to identify the center of the opinion space.

To keep things simple, we consider only the unidimensional case, i.e. the state of an agent is a scalar number in the interval  $[0, 1]$ . We denote with  $x(t)$  the state of a generic user at time  $t$  and with  $y(t)$  the state of a generic page.

The interaction rule between a user and a page captures the dynamics of social influence. Let us imagine that at time  $t$  a user edits a page. Let  $\mu \in [0, 1/2]$  be the speed (or uncertainty) parameter and  $\varepsilon \in [0, 1]$  the confidence [18]. If  $|x(t) - y(t)| < \varepsilon$  then:

$$x(t) \leftarrow x(t) + \mu(y(t) - x(t)) \quad (1)$$

$$y(t) \leftarrow y(t) + \mu(x(t) - y(t)) \quad (2)$$

else, if  $|x(t) - y(t)| \geq \varepsilon$ , we allow only Eq. (2) to take place with probability  $r$ . This addition to the bounded confidence averaging rule reflects the fact that, in peer production systems, users often deal with content they do not agree with without being influenced by it, as when a vandalized page is reverted to a previous, non-vandalized revision (also known as rollback).

Different pages can reflect different topics and hence receive attention from users based on their popularity. We employ a simple reinforcement mechanism to model this. Let  $c_p \geq 0$  be a constant. If  $m_t$  is the number of edits a page has received up to time  $t$ , then the probability of it being selected at that time will be proportional to  $m_t + c_p$ . When  $c_p \rightarrow \infty$ , pages will be chosen for editing with uniform distribution, regardless

of the number of edits they have received. Hence, we can study the impact of content popularity in user participation by setting  $c_p$  to a small or large value. Of course users do not always choose to edit an existing page. Sometime, a user can decide to create a new page. We model this by considering a rate of new page creations  $\rho_p$ . Whenever a new page is created, its state  $y$  is equal to the state  $x$  of creator. Creators are chosen at random among existing users.

In order to model user participation, the population of users is dynamic. First, we consider an input rate of new users  $\rho_u$ , whose state is chosen at random within the interval  $[0, 1]$ . Second, we consider an inhomogeneous departure rate that depends on the experience of users. Let us consider a generic user at time  $t$  and let us denote with  $n_t$  the number of edits he (or she) did up to  $t$ , and with  $s_t$  the number of these edits that resulted in the application of Eq. (1). Let  $c_s \geq 0$  be a constant and  $r(t)$  be the ratio

$$r(t) = \frac{s_t + c_s}{n_t + c_s} \quad (3)$$

The rate of departure  $\lambda_d(t)$  is then defined as:

$$\lambda_d(t) = \frac{r(t)}{\tau_0} + \frac{1 - r(t)}{\tau_1} \quad (4)$$

with  $\tau_0 \gg \tau_1$  time scale parameters. Depending on the value of  $r(t)$ , the expected lifetime  $\langle \tau \rangle$  will interpolate between two values:  $\langle \tau \rangle = \tau_0$  (long lifetime) for  $r(t) = 1$ ,  $\langle \tau \rangle = \tau_1$  if  $r(t) = 0$  (short lifetime). If  $c_s \rightarrow \infty$ , we recover a homogeneous process with rate  $\tau_0^{-1}$ , so we can set  $c_s$  to control how sensitive the departure rate is to unsuccessful interactions.

### 3 Evaluation Methods

#### 3.1 Computer Code Emulation via Gaussian Processes

Although we can perform the statistical evaluation of our peer production model using directly the computer simulator, this approach is not desirable, as evaluation of the computer code can be quite time consuming. We rely instead on emulation of the computer code output. We use a Gaussian Process (GP) as a surrogate model of the average lifetime  $\langle \tau \rangle$  of users in our peer production system. Gaussian processes (or Gaussian Random Functions, GRF) are a supervised learning technique used for functional approximation of smooth surfaces and for prediction purposes: see [25] for the application of GP to computer code evaluation.

Given input sites  $\Theta_{\text{obs}} = (\theta_1, \theta_2, \dots, \theta_N)$  we can evaluate our model as specified above, and obtain observations of the average user lifetime  $T_{\text{obs}} = (\tau_1, \tau_2, \dots, \tau_N)$ . Based on these observations, we wish to predict the value of  $\tau$  at an untested input site  $\theta$ , i.e.  $\tau(\theta)$ . With a GP, this value is  $\hat{\tau}(\theta) = \mathbb{E}[\tau(\theta) | \Theta_{\text{obs}}]$ ; the uncertainty in the prediction, that is,  $\text{Var}[\hat{\tau}(\theta)]$ , is equal to  $\text{Var}[\tau(\theta) | \Theta_{\text{obs}}]$ . With it we can compute a confidence interval that characterizes the uncertainty of the prediction of  $\tau$  based on training data  $(\Theta_{\text{obs}}, T_{\text{obs}})$ .

There are several strategies for selecting the input sites  $\Theta_{\text{obs}}$  at which we will run our computer simulator. Here we choose to employ a uniform, space-filling design generated via Latin Hypercube Sampling (LHS) because it yields better error bounds than those produced with uniform random sampling [19]. The space-filling requirement is attained using a *maximin* design. A maximin design is any collection of points  $\Theta$  that maximizes the minimum distance between points:  $\max_{\Theta} \min_{i < i'} \|\theta_i - \theta_{i'}\|$

### 3.2 Global Sensitivity Analysis

A computational or mathematical model is comprised usually of a number of parameters, or factors, which are meant to affect in some way its output, or response. Hence, in general, a model can be thought as a mapping between factors (input) and responses (output). One might be interested in the problem of quantifying how much output “variability” in this mapping can be apportioned to each of the inputs. Global Sensitivity Analysis (GSA) is a set of statistical techniques used to get an answer to this problem. See [24] for a primer on GSA.

One application of GSA is *factor screening*. The ranking of parameters is usually done by computing the sensitivity indices of each input parameter (factor). There are various techniques for computing the sensitivity indices, each with its own properties and assumptions. In this study we computed sensitivity indices by decomposing the output variance of our surrogate model. We used other techniques as well, namely partial correlation coefficients and standardized regression coefficients, and they gave concordant results. We choose to report here only the results of the decomposition of variance because it applies more naturally to non-linear models like ours.

The method we use was proposed by Sobol’ and is based on the analysis of variance (ANOVA) [28]. The idea is to decompose the variance of the output in several components that are attributable to independent factors, in our case the parameters of the model.

Let us assume that the space of parameters is  $[0, 1]^d$ , where  $d$  is the number of parameters. Sobol’ proposes to write the output  $Y$  as:

$$Y(\theta_1, \dots, \theta_d) = Y_0 + \sum_{i=1}^d Y_i(\theta_i) + \sum_{1 \leq i < j \leq d} Y_{i,j}(\theta_i, \theta_j) + \dots + Y_{1,2,\dots,d}(\theta_1, \theta_2, \dots, \theta_d) \quad (5)$$

and shows that this decomposition is unique under the assumption that components are orthogonal and have zero mean. In Eq. (5),  $Y_0 = E[Y]$ ,  $Y_i(\theta_i)$  is the *main* effect of parameter  $\theta_i$ ,  $Y_{i,j}(\theta_i, \theta_j)$  is the 2-way *interaction* effect between the  $i$ -th and  $j$ -th parameters ( $i \neq j$ ), and so on. Each summand is computable from suitable integrals. For example the main effect of  $Y_i$  is:

$$Y_i(\theta_i) = \int_0^1 \dots \int_0^1 Y(\theta_1, \dots, \theta_d) d\theta_{-i} - Y_0 \quad (6)$$

where with  $\theta_{-i}$  we mean the reduced parameter vector obtained by considering all parameters except  $\theta_i$ . Similar formulas can be obtained for higher order effects. Let

Table 1: Parameters settings for global sensitivity analysis.

| Parameter            | Variable name | Symbol        | Value(s)  | Unit  | Distribution    |
|----------------------|---------------|---------------|-----------|-------|-----------------|
| Const. popularity    | const_pop     | $c_p$         | (0, 100)  |       | uniform         |
| Const. successes     | const_succ    | $c_s$         | (0, 100)  |       | uniform         |
| Confidence           | confidence    | $\varepsilon$ | (0, 1/2)  |       | uniform         |
| Daily edit rate      | daily_users   | $\lambda_e$   | (1, 20)   | day   | uniform         |
| Daily rate of pages  | daily_pages   | $\rho_p$      | (1, 20)   | 1/day | uniform         |
| Daily rate of users  | daily_edits   | $\rho_u$      | (1, 20)   | 1/day | uniform         |
| Initial no. of pages |               | $N_p$         |           |       | see Subsec. 4.2 |
| Initial no. of users |               | $N_u$         |           |       | see Subsec. 4.2 |
| Long lifetime        | long_life     | $\tau_0$      | (10, 100) | day   | uniform         |
| Rollback probability | rollback_prob | $r$           | (0, 1)    |       | uniform         |
| Short lifetime       | short_life    | $\tau_1$      | (1/24, 1) | day   | uniform         |
| Simulation time      |               | $T$           | 1         | year  |                 |
| Speed                | speed         | $\mu$         | (0, 1/2)  |       | uniform         |
| Transient time       |               | $T_0$         | 2         | year  |                 |

us now consider the variances of the summands of Eq. (5). We can decompose  $\sigma^2$ , the total variance of  $Y$ , as:

$$\sigma^2 = \sum_{i=1}^d \sigma_i^2 + \sum_{1 \leq i < j \leq d} \sigma_{i,j}^2 + \cdots + \sigma_{1,2,\dots,d}^2 \quad (7)$$

The sensitivity indices proposed by Sobol' are obtained by standardizing all summands of Eq. (7), obtaining:

$$1 = \sum_{i=1}^d M_i + \sum_{1 \leq i < j \leq d} C_{i,j} + \cdots + C_{1,2,\dots,d} \quad (8)$$

$M_i$  is the *main sensitivity index* of parameter  $\theta_i$ ,  $C_{i,j}$  is the *two-way interaction index* between  $\theta_i$  and  $\theta_j$ , etc. Two quantities are of interest for assessing the importance of a parameter: the already cited main sensitivity index  $M_i$ ; and the *total interaction index*  $T_i$ , which is defined as the sum of all terms that involve parameter  $\theta_i$ :

$$T_i = \sum_{j \neq i} C_{i,j} + \sum_{\substack{1 \leq j < k \leq d \\ j, k \neq i}} C_{i,j,k} + \cdots + C_{1,2,\dots,d} \quad (9)$$

## 4 Results

### 4.1 Simulation scenario

Table 1 lists all parameters of the model, together with simulation settings. Two quantities have been held fixed: simulation time, and transient time. Two other parameters,

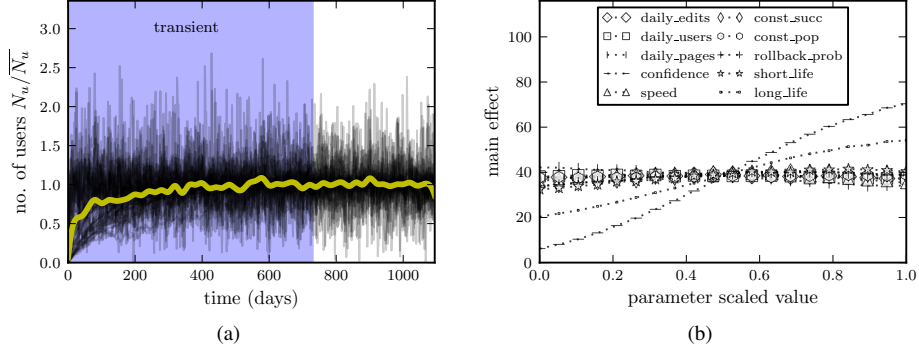


Figure 1: (a): transient time determination. (b): main effects plot.

the initial number of users  $N_u$  and the initial number of pages  $N_p$ , are determined after a transient, see Subsec. 4.2 below. All remaining parameters, instead, were assigned a range of values. To sum up, we had an input space of 10 independent dimensions.

We chose the long ( $\tau_0$ ) and short ( $\tau_1$ ) user lifetimes to range in non-overlapping intervals corresponding to different time scales, consistently with empirical observations of user participation from Wikipedia [7]. The value of the simulation time  $T$  was chosen so that a simulation would comprise more than one generation of long-term users.

Intervals for event rates such as the daily rate of edits ( $\lambda_e$ ), of new user arrivals ( $\rho_u$ ), and of new page creations ( $\rho_p$ ), were chosen looking at plausible values from the public statistics on the Wikipedia project.<sup>1</sup> These parameters have a strong influence on simulation time, therefore ranges for them were set trying to strike a balance between exhaustiveness of the sensitivity analysis and simulation wall clock time.

The choice of ranges for the constant popularity term ( $c_p$ ) and for the constant successes term ( $c_s$ ) was a bit more problematic. To our knowledge, none of them has ever been studied before in the context of peer production communities. We settled for ranges we deemed would be large enough for our purposes.

Finally, the opinion dynamics parameters. It is clear that  $\mu < 1/2$ . Regarding the confidence  $\varepsilon$ , the literature on bounded confidence models in one dimension suggests that for  $\varepsilon > 1/2$  the dynamics of consensus does not change noticeably. This should apply also to the dynamics of user participation in our model. We ran some simulations of the average lifetime, and found confirmation to this intuition. We thus restricted  $\varepsilon$  to the interval  $(0, 1/2)$ .

## 4.2 Transient

Transient duration  $T_0$  was determined empirically: we plotted the daily number of users  $N_u(d; \theta)$ ,  $d = 1, 2, \dots$ , for various values of the parameters  $\theta$  and chose  $T_0$

<sup>1</sup>These statistics are freely available on <http://stats.wikimedia.org>.

as the time after which all curves look stationary. Figure 1a reports the results of this exercise. In the figure, the shaded region corresponds to the transient interval  $(0, T_0)$ . The value of  $T_0$  is 730 days. The values of  $\theta$  were taken from a maximin LHD with 50 points. Each curve is scaled by its average value  $\bar{N}_u(\theta)$  computed over the interval  $d \in [731, 1095]$ . The yellow solid line is a B-spline fit of 50 evenly spaced observations of the expected scaled number of users  $N_u/\bar{N}_u$ , and serves as a guide for the eye.

During the transient phase we did not record any data, so that the estimation of  $\tau$ , on which the sensitivity our analysis is based, did not reflect the dynamics of opinion formation during the transient.

### 4.3 Factor screening via global sensitivity analysis

We sampled a maximin Latin Hypercube Design (LHD) with 50 points using the intervals listed in Tab. 1. To sample a decent maximin design, we generated  $10^4$  hypercubes at random and selected the one that maximized Eq. (3.1). We computed the average user lifetime  $\langle \tau(\theta) \rangle$  by running 10 replications for any  $\theta$  and averaging the values obtained.

We first plotted the values of the response variable  $\tau$  versus each input parameter to check visually for any linear trend. Scatter plots are shown in Fig. 2. A multiple linear regression gave a coefficient of determination  $R^2 = 0.83$ . However, no clear trend emerges from the plots for all parameters except for the confidence  $\varepsilon$  and the long lifetime  $\tau_0$ . For the latter, something similar to a linear trend can be seen, whereas for the other the relationship looks more of sigmoidal type. We tried fitting a sigmoid function to  $\tau$  as a function of  $\varepsilon$ . The result of a K-S test (p-value  $< 3.5 \times 10^{-4}$ ) rejected the normality of the residuals, and therefore led us to exclude a sigmoid model as a possible functional form of  $\tau(\varepsilon)$ .

Next, we fitted a GP emulator to the average user lifetime data, using the open source machine learning toolkit from the SciKits collection<sup>2</sup>. We then discarded the simulator and used  $\hat{\tau}(\theta)$  in lieu of it. To compute the sensitivity indices we used the Winding Stairs (WS) method, a resampling technique proposed in [14]. We computed main ( $M_i$ ) and total interaction ( $T_i$ ) effect indices for each parameter ( $i = 1 \dots 10$ ) using a WS matrix with  $10^4$  rows. The results are shown in Tab. 2.

The total variance  $\hat{\sigma}^2$  was also computed from  $\mathbf{W}$  (each column of a WS matrix is an independent sample). The WS method yields better estimates of the total interaction effects than other methods [5], so we impute the presence of some slightly negative values of  $M_i$  to the uncertainty in the estimation of the total output variance  $\sigma^2$  and to the presence of factors with almost null total effect.

Only two factors have a  $T_i > 3\%$ . These are the confidence  $\varepsilon$ , and the long term lifetime  $\tau_0$ . We explored further the individual contribution of each parameter in the output variance by looking at the main effect plots. These are plots of  $Y(\theta_i)$  as a function of  $\theta_i$ , and can be obtained evaluating Eq. (6) using Monte Carlo averaging and the GP emulator. To facilitate comparison of the different parameter ranges, in Fig. 1b we plotted the main effect as a function of the scaled parameter value. Figure

<sup>2</sup>Home page: <http://scikit-learn.sourceforge.net/>.



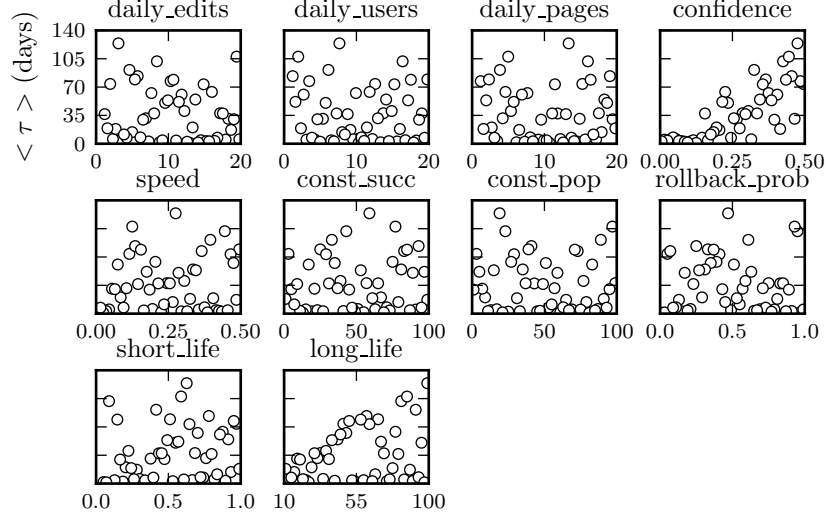


Figure 2: Scatter plots of  $\langle \tau \rangle$  versus  $\theta = (\lambda_e, \rho_u, \rho_p, \varepsilon, \mu, c_s, c_p, r, \tau_0, \tau_1)$ . Error bars (standard error of the mean lifetime computed over 10 realization) are all smaller than the data points.

1b shows that  $\rho_p$  and  $\tau_1$  have a slight effect on user lifetime too, the first negative and the second positive.

The difference between  $T_i$  and  $M_i$  is the fraction of variance that is only due to interactions between  $\theta_i$  and any other parameter or groups of parameters. For  $\varepsilon$  this difference is 0.08 and for  $\tau_0$  it is 0.05. Summed up together, this residual interaction effect amounts to almost three quarters (77%) of the total interaction effects from all remaining parameters. Thus we expect  $\varepsilon$  and  $\tau_0$  to have some interesting interactions with other parameters. We explored two-way interactions systematically using two-way interaction plots, which are the 3D counterparts of the curves of Fig. 1b.

Given two parameters  $\theta_i$  and  $\theta_j$ , with  $i \neq j$ , we computed  $Y_{i,j}(\theta_i, \theta_j)$ : we evaluated Eq. (6) in a similar way, this time holding fixed the values of two parameters instead of one. Here we report the results on the interaction between  $\varepsilon$  and other parameters, included  $\tau_0$ . The plots are shown in Fig. 3 and 4.

Almost all parameters show just a weak interaction with  $\varepsilon$ , which occurs at low ( $\varepsilon < 0.1$ ) and high ( $\varepsilon > 0.4$ ) values of it. Only the pair  $\{\varepsilon, \tau_0\}$  shows a significant degree of interaction.

#### 4.4 User lifetime distribution

Previous studies on continuous opinion dynamics under bounded confidence show that, as  $\varepsilon$  grows, the population of agents undergoes a gradual change from a regime with no consensus, to a regime of total consensus with a single cluster [9, 12]. In our model this

Table 2: Variance decomposition. Winding Stairs sample size  $10^4$  rows, total variance 635.365 days<sup>2</sup>.

| Parameter     | $M_i$  | $T_i$ |
|---------------|--------|-------|
| $\lambda_e$   | -0.002 | 0.014 |
| $\rho_u$      | -0.003 | 0.02  |
| $\rho_p$      | 0.003  | 0.027 |
| $\varepsilon$ | 0.65   | 0.73  |
| $\mu$         | -0.004 | 0.03  |
| $c_s$         | 0.004  | 0.03  |
| $c_p$         | -0.005 | 0.016 |
| $r$           | -0.005 | 0.026 |
| $\tau_1$      | 0.002  | 0.03  |
| $\tau_0$      | 0.18   | 0.23  |

shift must reflect somehow in the average user lifetime, but what shape the user lifetime distribution takes during it? The findings from the previous section let us restrict the field of study to just two parameters of the original ten, namely  $\varepsilon$  and  $\tau_0$ . In this section we focus only on them, and try to understand what is the actual distribution of user lifetimes, by simulating from our model.

We performed simulations holding fixed the user lifetime parameters ( $\tau_0 = 100$  days and  $\tau_1 = 1$  hour), while changing the value of  $\varepsilon$ . The values of all other parameters were fixed to the midpoints of the respective ranges listed in Tab. 1. We computed the log-lifetime  $u = \log(\tau)$  and fitted a 2-components Gaussian Mixture Model (GMM) to  $u$ . Figure 5 reports the result of the fitting, showing the densities of the individual components using stacked area plots. We report here only two values of  $\varepsilon$ ,  $\varepsilon = 0$  and  $\varepsilon = 0.3$ , which is a value greater than the threshold for consensus in Deffuant’s model, to show the difference between the two regimes.

## 5 Discussion

In this section we discuss the main findings of the present study. We presented an agent-based model of user participation in a peer-production community. We model participation as a bounded confidence consensus process, where users modify content according to their objectives and skills (represented by a continuous state  $x$ ), and are in turn indirectly influenced by the rest of the community. We use global sensitivity analysis to study the importance of the model’s parameters in explaining the average user lifetime. The first interesting – and rather surprising – finding is that, as shown in Tab. 2, of the overall ten parameters of the model, only two affect the average user lifetime in a considerable way. This is interesting because it suggests that several other factors like content popularity, user community growth, and user activity rate, are not as important as the general level of “tolerance” of the community (given by the confidence  $\varepsilon$ ) in affecting the process of group consensus. Moreover, interaction plots show that

relevant interactions occur between  $\varepsilon$  and  $\tau_0$ : this confirms the intuition that the role of  $\tau_0$  is to set the support of the distribution of  $\tau$ , and that  $\varepsilon$  acts as a switch, controlling the transition from a regime where only short-term forms of participation are possible, due to the low rate of successful user-page interactions, to a consensus regime where a cluster of long-term users is able to emerge.

Of course, the results from the factor screening should be also viewed in light of our simulation setup. We decided to focus on a stable community, where the number of users  $N_u$  is stationary, and not on the initial phase of community formation. Plausibly, during this transient phase other parameters, such as the speed  $\mu$ , and the rollback probability  $r$ , might have more importance in determining the span of user participation.

The second interesting finding is about the actual distribution of user participation, which is markedly bimodal. From Fig. 5 it is possible to appreciate, for  $\varepsilon = 0.3$ , a clear subdivision in two groups of users based on their participation span. We can see also a subdivision for  $\varepsilon = 0$ , which is probably related to the fact that  $c_s = 50$  in that setup. Although we did not perform a proper model calibration, this finding is encouraging, as previous studies on the distribution of user accounts lifetime in Wikipedia have shown a similar bimodal pattern [7].

In general, both findings show that agent-based model can be studied through the systematic use of simulations and computer code emulation, and provide a novel connection between model of opinion dynamics, whose study has been so far notoriously lacking on the empirical side [4, 27], and peer production.

### 5.0.1 Acknowledgments.

Alberto Vancheri and Paolo Giordano for the insightful discussions; the anonymous reviewers, for the suggestions on how to improve the manuscript; the conference organization, for their generous financial support.

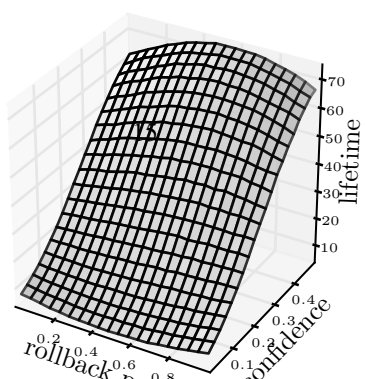
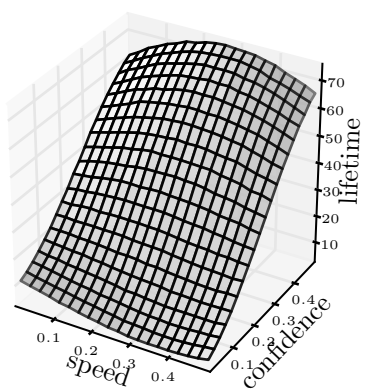
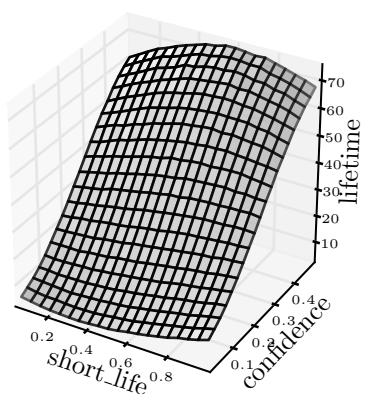
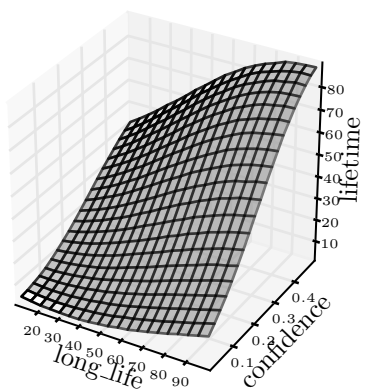
## References

- [1] Backstrom, L., Kumar, R., Marlow, C., Novak, J., Tomkins, A.: Preferential behavior in online groups. WSDM'08 pp. 1–11 (Dec 2007)
- [2] Beschastnikh, I., Kriplean, T., McDonald, D.W.: Wikipedian self-governance in action: Motivating the policy lens. In: Proceedings of the second ICWSM conference (2008)
- [3] Carletti, T., Fanelli, D., Guarino, A., Bagnoli, F., Guazzini, A.: Birth and death in a continuous opinion dynamics model. Eur. Phys. J. B 64(2), 285–292 (Jul 2008)
- [4] Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. 81(2), 591–646 (May 2009)
- [5] Chan, K., Saltelli, A., Tarantola, S.: Winding stairs: A sampling tool to compute sensitivity indices. Statistics and Computing 10, 187–196 (2000), <http://dx.doi.org/10.1023/A:1008950625967>

- [6] Choi, B., Alexander, K., Kraut, R.E., Levine, J.M.: Socialization tactics in wikipedia and their effects. In: CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work. pp. 107–116. ACM, New York, NY, USA (2010)
- [7] Ciampaglia, G.L., Vancheri, A.: Empirical analysis of user participation in online communities: the case of wikipedia. In: Proceedings of ICWSM (2010)
- [8] Cifforilli, A.: Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of wikipedia. *First Monday* 8(12) (Dec 2008), [http://firstmonday.org/issues/issue8\\_12/cifforilli/index.html](http://firstmonday.org/issues/issue8_12/cifforilli/index.html)
- [9] Deffuant, G., Neau, D., Amblard, F., Weisbuch, G.: Mixing beliefs among interacting agents. *Adv. Comp. Sys.* 3, 87–98 (2001)
- [10] Goldman, E.: Wikipedia's labor squeeze and its consequences. *Telecomm. and High Tech. Law* 8, 157–184 (August 2009)
- [11] Grabowski, A., Kosiński, R.A.: Life span in online communities. *Phys. Rev. E* 82(6), 066108 (Dec 2010)
- [12] Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence–models, analysis, and simulation. *J. Art. Soc. Soc. Sim.* 5(3), paper 2 (2002), <http://jasss.soc.surrey.ac.uk/5/3/2.html>
- [13] Hogg, T., Lerman, K.: Stochastic models of user-contributory web sites. pp. 50–57 (2009)
- [14] Jansen, M., Rossing, W., Daamen, R.: Monte-Carlo Estimation Of Uncertainty Contributions From Several Independent Multivariate Sources. In: Grasman, J., van Straten, G. (eds.) *Predictability And Nonlinear Modelling In Natural Sciences And Economics*. pp. 334–343 (1994)
- [15] Kittur, A., Kraut, R.E.: Beyond wikipedia: Coordination and conflict in online production groups. pp. 215–224 (2010), <http://dx.doi.org/10.1145/1718918.1718959>
- [16] Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 462–470. ACM, New York, NY, USA (2008)
- [17] Lin, H.F., Lee, G.G.: Determinants of success for online communities: an empirical study. *Behaviour & Information Technology* 25(6), 479–488 (Nov-Dec 2006)
- [18] Lorenz, J.: Continuous opinion dynamics under bounded confidence: A survey. *Intl J. Mod. Phys. C* 18, 1819–1838 (2007)

- [19] McKay, M.D.: Latin hypercube sampling as a tool in uncertainty analysis of computer models. In: Proceedings of the 24th conference on Winter simulation. pp. 557–564. WSC '92, ACM, New York, NY, USA (1992), <http://doi.acm.org/10.1145/167293.167637>
- [20] Ortega, F., Izquierdo-Cortazar, D.: Survival analysis in open development projects. In: Proceedings of the 2009 ICSE Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development. pp. 7–12. FLOSS '09, IEEE Computer Society, Washington, DC, USA (2009), <http://dx.doi.org/10.1109/FLOSS.2009.5071353>
- [21] Panciera, K., Halfaker, A., Terveen, L.: Wikipedians are born, not made. In: Proceedings of GROUP'09 (2009)
- [22] Raafat, R.M., Chater, N., Frith, C.: Herding in humans. Trends in Cognitive Sciences 13(10), 420 – 428 (2009), <http://www.sciencedirect.com/science/article/B6VH9-4X6PPCY-1/2/38f26f1994570f7a58d587bb5a7a0569>
- [23] Rafaeli, S., Ariel, Y.: Psychological aspects of cyberspace: Theory, research, applications, chap. Online Motivational Factors: Incentives for Participation and Contribution in Wikipedia, pp. 243–267. Cambridge University Press (2008)
- [24] Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M.: Sensitivity Analysis in Practice—A guide to Assessing Scientific Models. John Wiley & Sons, Ltd. (2004)
- [25] Santner, T., Williams, B., Notz, W.: The Design and Analysis of Computer Experiments. Springer-Verlag, New York (2003)
- [26] Schroer, J., Hertel, G.: Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. Media Psychology 12(1), 96–120 (2009)
- [27] Sobkowicz, P.: Modelling opinion formation with physics tools: Call for closer link with reality. Journal Artificial Societies and Social Simulation 12(1), 11 (2009), <http://jasss.soc.surrey.ac.uk/12/1/11.html>
- [28] Sobol', I.M.: Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and Computers in Simulation 55(1-3), 271–280 (February 2001), <http://www.sciencedirect.com/science/article/B6V0T-42DX509-11/2/7992ee21d186afc323213675e8547d6f>
- [29] Wilkinson, D.M.: Strong regularities in online peer production. In: Proceedings of the 9th ACM conference on Electronic commerce. Chicago, Illinois USA (2008)
- [30] Wilkinson, D.M., Huberman, B.A.: Cooperation and quality in wikipedia. In: Proceedings of WikiSym '07, 3rd Intl Symposium on Wikis. Montréal, Québec, Canada (October, 21–23 2007)

- [31] Yang, J., Wei, X., Ackerman, M., Adamic, L.: Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (2010), <http://aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1466/1856>



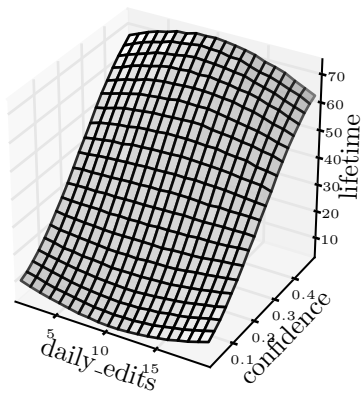
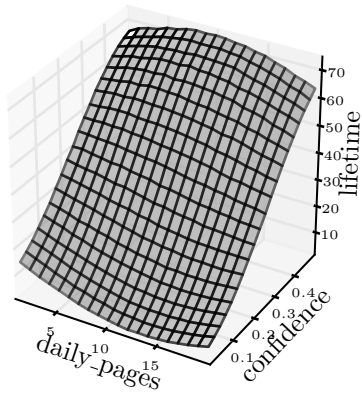
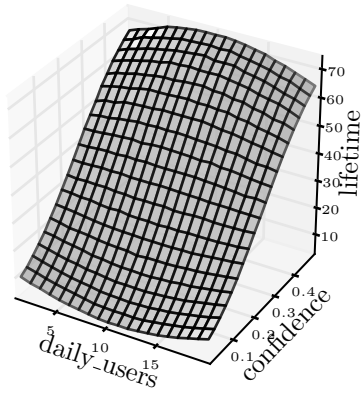


Figure 4: Two-way interaction plots (cont'd)



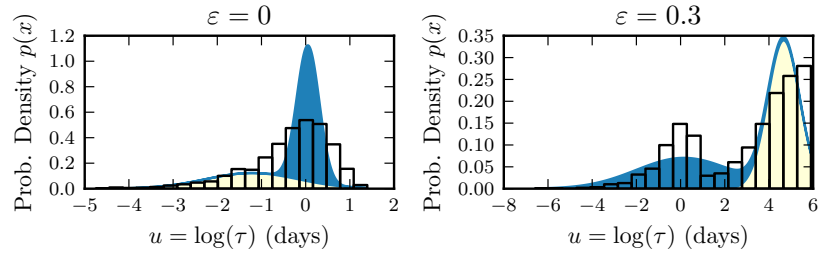


Figure 5: GMM fit of log-lifetime of user accounts in two different runs of the model. For  $\varepsilon > \varepsilon_c$  a bi-modal pattern is a clear feature of user participation.