Big from the beginning: Assessing online contributors' behavior by their first contribution

Sylvain Dejean, Nicolas Jullien

Big From the Beginning: Assessing Online Contributors' Behavior by Their First Contribution.

Sylvain Dejean, Nicolas Jullien.

Université de la Rochelle, Sylvain.Dejean@univ-lr.fr, and Institut Mines Télécom Bretagne, ICI-M@rsouin, Nicolas.Jullien@telecom-bretagne.eu

Abstract

This paper aims at investigating the process of involvement in open online communities producing knowledge, via the link between the first contribution and the level of contribution reached. While most studies look at the career of contribution after the first contribution, we focus on what happened before and during the first contribution. We challenge the fact that becoming a core member starts with peripheral contributive activities and results from a continuous learning process, as explained by the theory of community of practice. On the contrary, and coherent with epistemic community theory, our results, based on 13,000 answers to a survey on the use and contributions to Wikipédia, show that the future level of users' involvement depends on the time between the discovery of Wikipedia and the first contribution (negatively), and of the effort made in the first contribution (positively). Implications for management are also discussed.

JEL classification: L17, L86, H41

1. Introduction

Mobilizing hundreds (Linux) to thousands of contributors (Wikipedia), volunteer online open projects aiming at creating new knowledge, online "commu-

nities of creation", as named by Rullani and Haefliger (2013), are viewed as central in the generation of new, innovative knowledge by and for firms. But the path to successful community building is still risky and uncertain, and as for business building, most of the attempts fail, no matter how many hundreds of thousands of dollars were put into them (Worthen, 2008).

One of the key elements to develop a successful and sustainable community, as explained a quarter of century ago by Eric von Hippel (1986), is to attract enough highly competent and "committed/committing" contributors, being they named "lead users", "core", or "big" contributors (Mahr and Lievens, 2012; Fang and Neufeld, 2009), i.e. the most productive people, who are also those with more responsibility in the management of the project (Rullani and Haefliger, 2013). This focus is explained by the fact that, since both big and small contributors are needed (O'Mahony and Bechky, 2008; Hemetsberger and Reinhardt, 2009), as in most collective actions and projects (Oliver et al., 1985; Ostrom, 1990), the former are much rarer than the latter, in addition to being more productive (Lakhani and von Hippel, 2003; Maillart et al., 2008; Voss, 2005).

New big contributors are constantly needed as they are subject to disengagement after some years (Ortega and Izquierdo-Cortazar, 2009 for open online communities, Borzillo et al., 2011 for intra-organization communities), and difficult to recruit and retain (Von Krogh et al., 2003 in the case of open source software communities, Halfaker et al., 2013 for Wikipedia). This echoes more general findings about the efficiency of groups. As shown by Uzzi and Spiro (2005) in the case of musical comedies, and Uzzi (2008) in the case of a social network, for a creative group to be successful, it needs to fine tune the level of newcomers, for fresh ideas, in an already constituted group (Guimera et al., 2005; Defélix et al., 2005). Wikipedia, for instance, is known for its gender bias amongst editors, which would explain that certain topics are less well covered

by the encyclopedia[1]. Specific programs, targeting new contributors, or "newcomers", have been designed to facilitate the supposed learning curve leading to regular contribution, thus "mentoring" them (Wikipedia program term). This means, using Kram (1983)'s typology, assigning to a newcomer (new registered person or beginner contributor) a volunteer coach/counselor, who is a regular contributor, to guide him/her through the different contributing steps and rules (see Mateos-Garcia and Steinmueller, 2008 for open-source project Debian, and Musicant et al., 2011, for Wikipedia).

In this article we ask whether it is possible to identify the different contributors early enough in the process to adapt the mentoring to their profiles, and thus making it possible to decrease the high discouragement rate of both newcomers and mentors (Musicant et al., 2011). This question is based both on studies showing that those future big contributors may be identified from their very first contributions (Fang and Neufeld, 2009 for open source, Panciera et al., 2009 for Wikipedia), and on the theory of epistemic community (Cohendet et al., 2001; Edwards, 2001), which stresses that those communities are project-oriented communities of experts, whose expertise is acquired outside the community. According to this theory, entering a community is equivalent to starting to contribute, there is no peripheral participation.

In addition to providing the knowledge-community managers with results on how to better recruit future core contributors, this article aims at contributing to the studies of social practices in context and to the characterization of such online communities, leveraging on and discussing previous works such as Amin and Roberts (2008)'s on the different models of "knowing in action" (craft/task-based, professional, epistemic/creative, virtual), and Rullani and

[1]On that matter, the MIT Technology Review propose a good introduction of the recruiting problem and of its consequences: `http://www.technologyreview.com/featuredstory/520446/the-decline-of-wikipedia/`

Haefliger (2013)'s analysis of the dynamics of intra-organizational communities of creation. This contribution is based on econometric analyses of a survey of more than 13,000 Wikipedia users, and sometime contributors, of their first contribution and of their curent level of contribution.

The article is organized as follows: in section 2 a review of the literature used to construct our framework of investigation, in section 3 the formulation of our hypotheses, in section 4 the data collection strategy (choice of the community and definition of the questions), in section 5 the results. We discuss the consequences of this work, its limits and future research in section 6 before concluding.

2. A career in communities of creation: From peripheral to big contributor?

There is a consensus that big and small contributors do not have the same aims when contributing: in the case of open source software, Shah (2006) showed that regular contributors enjoy programming and interacting with the rest of the community (i.e., labeled as "hobbyists"), whereas new or sporadic contributors are typically driven by an immediate need for software (i.e., use value). For the most involved wikipedians, the recognition from their peers ('credit') is an important motivation (Forte and Bruckman, 2005; Bryant et al., 2005), as is the sense of mission (Liang et al., 2008; Prasarnphanich and Wagner, 2009); for most of the (small) contributors, the will to fix mistakes is the principal motivation, meaning that they are not strongly committed to the project (Kamata et al., 2010). According to Shah (2006), this echoes the more general sociological notion of "career" (Becker, 1960, 1963), which stresses that people's motivations and actions are curved by the social interactions they meet in their practice.

Considering this, the concept of legitimate peripheral participation, or LPP (Wenger, 2006, 1998), has been used to explain how this learning works: future

contributors start to get involved themselves by observing, "dipping their toes in to passively participate while learning more about a complex system" (Antin et al., 2012) before editing (for Wikipedia) or coding (for open-source), then interacting with the experienced member at the margin, and so on. There would appear to be a slow process of "decantation" in the group of readers and early contributors leading to the emergence of regular contributors.

This argument is, however, theoretically and practically disputable. First of all, most of the studies cited, even Antin et al. (2012), focus on what happens after the first contribution, that is when people have proved their capacity to produce new knowledge. If there is a learning process, it does not concern knowledge, nor how to propose it. Theoretically, LPP is seen in communities of practice, like professional/specialized forums (online or, as in Wenger (1998)'s studies, local, geographically situated), where people exchange primarily about their "practices" and build their knowledge of those practices. The participation is a process, where people first observe, then make minor contributions, and gradually increase their engagement and the complexity of their contributions. On the contrary, communities of creation are (virtual) epistemic communities, or task-oriented groups, which brings experts together around a common goal (Amin and Roberts, 2008)[2]: the building of (new) knowledge. People, even newcomers, are evaluated on their capacity to produce this knowledge[3], on

[2]On that matter, we follow Amin and Roberts (2008)'s analysis of the different forms of "knowing in action", but on one point: they distinct between virtual communities and epistemic ones, on the only basis of the nature of the communication between people (face-to-face or virtual, see table 2, page 257), when the other types communities are segmented by the type of knowledge and the competences of the people involved in. Actually, they are not followed on that distinction by Cohendet et al. (2001); Rullani and Haefliger (2013) who agree on the other parts of the analysis. Following their example we will rely on Amin and Robert's analysis of the knowledge production, and will not discuss the impact of the mean of communication on the exchange, which is out of the scope of this study.

[3]For Wikipedia, when projects have rules for running for administrator, they are about knowing the rules, but also about the number of edits of articles (more than 3,000 and more of one year of activity for the French Wikipedia, `https://fr.wikipedia.org/wiki/Wikipédia:Candidature_au_statut_d'administrateur`). O'Mahony and Ferraro (2007, part II), on Open Source project, showed that "developers who were making greater technical contributions (in

the basis of competencies acquired mainly outside the community and before starting to contribute (Edwards, 2001). In other words, from the beginning, some contributors are more able to propose major contributions than others, and make significant contributions (there is no gradual engagement).

The argument about the differences between big and small contributors being an indicator of a learning period can be overturned: those who are the most willing to become regular contributors have, from the beginning, different capacities and goals, which are just not statistically discernible because of the mass of the lurkers. The studies comparing current small contributors and current big contributors appear to miss the point, by including among the current (and future) small contributors some new future big contributors. As already said, studies seem to indicate that from the very first contributions, the future very active contributors behave differently from the others (Fang and Neufeld, 2009 for open source, Panciera et al., 2009 for Wikipedia).

There are other theoretical arguments in favor of a correlation between the level and nature of the very first contribution and the will to be very involved in the community. In a standard job market signaling strategy analysis (Spence, 1973), Lerner and Tirole (2002) argued that contributors to open source projects were driven by the will to prove their competencies. Consequently, the first contribution could be considered as a signal to the community managers of the will to be integrated in the production process, and of the contributor's value, whose position will be monetized on the job market later on. However, as far as we know, this has not been proven in the case of open source software, even though a majority of open-source contributors were computer professionals in 2004 (Lakhani and Wolf, 2005), and can be seen as more disputable in the case

terms of impact but not effort) and who were more engaged in organization building were more likely to become members of the leadership team". (p. 1096). Fleming and Waguespack (2007) found the same result in their study of the Internet Engineering Task Force community.

of Wikipedia, where the link with the job market is less obvious. Another economics argument may be more convincing. Regarding the editing complexity (Butler et al., 2008; Cardon, 2012), making the first contribution is an investment, which is lost if not followed by other contributions, in other words, a sunk cost. If the people are ready to pay this sunk cost: making the contribution and coping with the rebuffing rejection/revision process, it may indicate that they are, more or less consciously, disposed to amortize it over more than just one or two contributions.

After this rapid review of the literature, it seems that there is a relative consensus that there are different levels of involvement, from user to core/big contributor (again, Von Krogh et al., 2003; Hertel et al., 2003, and more recently Crowston, 2011): "non-contributors"; small/peripheral contributor; regular; and very involved contributor. But there is no consensus on the existence and on the characteristics of the journey between those different levels. This paper looks at a specific part of this journey, wondering if regular and/or big contributors present any specific behavior when they enter the community, i.e. during the non-contributing phase and when they do their first contribution, thus making them identifiable very early by the managers of those communities.

3. Hypotheses

As explained in the preceding section, in a community defined by its goal, which is the production of knowledge, contributing is viewed as the action of providing new knowledge. This does not mean that the readers/users do not participate in the project (Antin and Cheshire, 2010), but, in addition to not creating new knowledge, they are very hard to identify (the fact of simply observing or using is very hard to discern). The first contribution, defined as the first time somebody proposes new, original knowledge (new for the community, not necessarily for the humankind), is a clear milestone, which makes it

possible to define the observation and possibly learning period (before the first contribution), the active period (after), and helps to refine our questions: does this first contribution carry specific information about a correlation between the level of expertise in the beginning and the future behavior of the contributor (the level of contribution reached) (hypothesis 1)? What happened before this contribution, in terms of community and knowledge learning (hypothesis 2)?

3.1. Hypothesis 1: The first contribution: a measure of expertise and a signal for the future

The question about the link between the first contribution and the level of involvement can be phrased this way: is this first contribution part of the learning process (how to make a contribution), as the possibility to experiment is at the core of the learning process, and supposed to be facilitated in an online project (Bryant et al., 2005), or is it more a signal of expertise, carrying with it information about who is going to contribute strongly and who will stay at the edge?

If we follow the argument of legitimate participation process, the first contribution will be minor (e.g., correcting a misspelled word, reporting a bug), to learn how to contribute. And there is no reason to think that it is correlated to the future level of contribution. If we consider the fact that we are looking at epistemic communities, considering the argument by Edwards, the first contribution can be major, and is not necessarily a way to learn what contributing means. The signaling argument (i.e., newcomers signal their expertise by making a strong first contribution) and the sunk cost argument (i.e., that this important investment is legitimated if users amortize it in the future) add to Edwards' view to defend the idea that the first contributionis a signal of how active the contributor will be in the future. In other words:

Hypothesis 1.1. The investment in the first contribution. The level of involvement reached may be positively correlated with the "magnitude" of the first contribution in terms of new knowledge.

Hypothesis 1.2. The learning motivation of the first contribution. The first contribution is a signal of the new status of the contributor, and the reason for so doing is not to learn something about how to contribute, for the future regular contributors.

3.2. Hypothesis 2: Approaching the community

Identifying the very first contributions as a key signal regarding the future level of contribution raises the question of the period before this contribution: in LPP theory, learning is argued to happen before the phase of active contribution, 1) participating at the periphery, reading the articles (Wikipedia), participating in user group lists (open-source) or/and 2) acquiring some knowledge about the contributing process. Learning is also said to continue after this initial phase, as contributions become more complex. According to Edwards, and to the literature on epistemic communities, people who are going to contribute new knowledge acquire this knowledge outside the community. Their learning period would appear to be more about how to propose new knowledge and about the organization of the community, i.e., getting to know the people involved. On that point, both theories agree that if people come for knowledge, they stay because of the social connections they develop (Butler et al., 2007). Considering this, knowing insiders may reduce the learning period and favor involvement (Mateos-Garcia and Steinmueller, 2008; Musicant et al., 2011). This is because contributors may be "mentored", which, to our understanding, involves making the rules more explicit and/or accelerating the social process. To test the value of these human connections, we asked people how they get information about how to contribute, and compared "codified" information, (i.e. online tutorials),

with straightforward human interaction. This may be somewhat compensated for by the passive, observatory period, if passive learning is a way to integrate into the community. So the length of time between starting to use Wikipedia and starting to contribute may mitigate the advantage of human help. This leads to two joined hypotheses:

Hypothesis 2.1. The source of help. If people have been helped to take their first step in the project, they will be more involved in the project, more than if they have only accessed the written guidelines.

Hypothesis 2.2. Time of observatory period. The more time people spend observing the community before contributing, being simple readers, the better they understand it, and thus the greater their chances of becoming at least regular contributors.

4. Data collection strategy

We chose to verify the hypotheses via a survey of Wikipedia users. We first explain why we selected Wikipedia, before defending the need for a survey to collect the data needed.

4.1. Choice of Wikipedia

Although Open Source initiatives are numerous, in various industries (Balka et al., 2009), the main open knowledge project outside the computer industry is to be found in the encyclopedia editing project known as Wikipedia. It has become one of the most successful knowledge production projects ever, with more than 4 million articles for the English version and more than one million visits per day, and is seen as a model for knowledge management theory (McAfee, 2006; Hasan and Pfaff, 2006). But even this successful project has recruiting problems, as already mentioned. Secondly, Hess and Ostrom (2006a) pointed out is that online communities lowered the boundary between those who are

in and those who are out. Wikipedia, which does not require programming competences from its contributors, seems to be one of the communities where the boundary is the lowest. If anyone can enter, i.e. try to produce knowledge, this should facilitate various trajectories of contribution. If the very involved contributors may be detected their first contribution to Wikipedia, open source communities where barriers to access are higher should present the same characteristics.

Finally, there is already a discussion about the fact that Wikipedia may or may not be a place of apprenticeship (Antin et al., 2012 vs Panciera et al., 2009, even if both agree that the first contributions provide information about the future profile of contributor).

4.2. Data collection

Online communities produce complete and available data on contributions, but, unfortunately little information about the contributors. Wikipedia is a good example of this, as information about contributors regarding their skills, their sociological background or their motivations are poorly documented: Lam et al. (2011), using users' page gender boxes and preference settings, for gender studies, reported a gender information rate of only 6.5% for the editors of the English Wikipedia. In addition to this, anonymous contributors are, by definition, not registered, when representing more than 90% of the contributions to the French Wikipédia, according to Auray et al. (2007), even if the regular contributors are said to be all registered (ibid). Thus, following Amichai-Hamburger et al. (2008); Yang and Lai (2010), and, in order to collect a complete set of information about the various contributors to the project, we chose to make a survey of these users (and amongst them, contributors to Wikipedia) regarding their contribution in Wikipédia, i.e. the French Wikipedia, and to link this contribution level to socio-demographic variables and the variables describing

the first contribution[4].

To reach the users, the version published online was announced in the "Bistro"[5] and finally, thanks to members of Wikimédia France and the site administrators, announced as a banner on the Wikipédia home page, from mid-January to mid-February 2011. This meant that every user of the site could see the banner during this period. This data collection method allowed us to construct a non-probability-based sample of French Wikipédia users (and contributors).

4.3. Construction of the variables

4.3.1. Independent variables

We present first the four constructed variables corresponding to our four hypotheses: the investment in the first contribution (H 1.1) and the motivation to do it (H 1.2), the source of mentoring (H 2.1) and the measure of the observation period (H 2.2), and the the control variables included in the model.

The definition of the first contributions: the very first contribution. When people register in Wikipedia, they may have spent a lot of time contributing to this community editing articles, anonymously. Our results suggest that: only 39% of the contributors were registered when they made their first contribution. So we defined the first contribution as the first time a person edited the system and made a change, whether anonymously or not.

H 1.1. The investment in the first contribution. Regarding the types of contribution, Antin et al. (2012) used a 10 revision types typology: "adding citations, adding content, changing Wiki markup (meta-information), creating articles,

[4] As a convention, in this article, when we speak about Wikipedia in general, we name it Wikipedia, and when it concerns the French project, we use the French spelling, Wikipédia.

[5] Discussion space in and of Wikipédia, http://fr.wikipedia.org/wiki/Wikipédia:LeBistro

deleting content, fixing typos, reorganizing text, rephrasing existing text, vandalism, deleting vandalism, and "unsure" if coders felt that the nature of the revision was ambiguous." We did not expect many answers for vandalism, and as it seems more a way to test the system than an editing action, nor for meta-information contribution, the latter being done more by experienced contributors. We ended up with the following categories: fixing a spelling mistake/typo, changing or adding a reference, or the creation, the translation or the modification of an article (reorganizing text, rephrasing existing text). As our goal was to discriminate among the contributions by their effort, we grouped them the following way: fixing a spelling mistake, changing or adding a reference are defined as minor contribution (MINOR_CONTRIB) to the encyclopedia (as they are at sentence level), and the others, which are at article creation level, as major contribution (MAJOR_CONTRIB).

H 1.2. The learning motivation of the first contribution. Aligned with hypothesis 1.2, we asked people if they made a first contribution to test the system (MOTIV_TEST), to explore the contribution process, (MOTIV_CURIOSITY), or to contribute by improving an article (MOTIV_IMPROVE).

H 2.1. The sources of help. As explained before, having certain connections may increase future contributors' comprehension of the rules and of the system, but there are also online tutorials, even if reading them is not an obligation. Does human contact, or mentoring, provide added value to the the tutorial, putting people on the "regular contributor" track? We thus asked the Wikipedians if they made their first contribution without seeking explanations (HOW_EASY), using tutorials available on the Wikipedia website (HOW_TUTO) or being helped by those around them (HOW_PEER). This is directly inspired from the "Facilitating Conditions", as listed by Venkatesh et al. (2003, Table 12): "1. Guidance was available to me in the selection of the system. 2. Specialized

instruction concerning system was available to me. 3. A specific person (or group) is available for assistance with system difficulties."

H 2.2. Length of observation period. To evaluate the period between the discovery of Wikipedia and the contribution, we asked people if they contributed for the first time during the first month after discovering Wikipedia (FIRST_CONT1); the first year after discovering Wikipedia (FIRST_CONT2) or more than a year after (FIRST_CONT3). This measures only a part of the level of involvement, as somebody may spend a lot of time observing a community during a short period, or vice-versa[6]. We discuss this limitation in the discussion section.

Control variables. We controlled for a potential "lassitude/exhaustion effect", which is usual in a community based on voluntary contribution, implementing a variable which represents the number of years passed since the first contribution.

To say that anyone can edit does not mean that everybody is able to, even if the barriers are lower than for open source software production. The same sociological differences found in the adoption of new technologies and especially in information technologies (Atkin et al., 1998; van Dijk and Hacker, 2003) seem to matter regarding the use and the building of Wikipedia: users are younger, of a higher level of education, and more often male, than the population using the Internet[7]. The contributors are of a higher level of education, mostly male, older in mean than Wikipedia users, and have better computer skills (Glott et al., 2010). But when the gap is bridged, the socio-demographic variables seem to count for little or nothing to explain the different levels of contribution: there is, for example, no significant gender difference in the level of edition between registered Wikipedians (in the English Wikipedia, Antin et al., 2011). So, as

[6] We thank one of the anonymous referees for pointing out this limitation.

[7] `http://www.pewinternet.org/Reports/2011/Wikipedia/Report.aspx`, survey of 852 people representative of US Internet users.

control variables, we introduced the sociological description of the contributors (age, gender, level of education and whether the person is employed or not).

Table 1 describes all the independent variables used in this article.

[Insert Table 1 here]

4.3.2. The dependent variable: the level of contribution

As our goal was firstly to distinguish between contributors and readers, and the people who just tried to contribute from those who are regular contributors, we needed to categorize people on these three levels of contribution (reader, occasional contributor, regular contributor).

Secondly, as we wanted to see if the first contribution could also explain a level of involvement we tried to define two levels of contribution within the regular contributors, based on the literature, which shows the existence of a group of very involved contributors. But the characterization of those big contributors is far from being simple and various definitions co-exist[8].

Considering the difficulty to obtain a robust measurement, both because the authors do not agree on what this measure should be, and because it is very hard for people to precisely evaluate the number of edits they make, in mean, in a month, we decided to propose a simplified subjective evaluation of their level of contribution to the people surveyed, and to ask them to identify themselves with one of these categories. To the question "Have you ever contributed to Wikipedia?", the possible answers were: "Never"; "It has happened once or twice, but no more"; "You contribute regularly"; "You consider yourself a big contributor". This avoided the definition of a precise border, but increased the

[8] Panciera et al. (2009) defined "Wikipedians as [registered] editors who have made at least 250 edits over their lifetime" but showed that this barrier doesn't matter beside the existence of a population of "super-elite" editors "who made more than 5000 edits" and seem to share specific behavior. Wikipedia Statistics pages propose another level of measurement, naming "active Wikipedians" those "who contributed 5 times or more in the month", "very active Wikipedians" those "who contributed 100 times or more in the month", and "contributors", "Wikipedians who edited at least 10 times since they arrived".

risk of bad self assignment, especially regarding the two categories of regular and big contributors, which are purely subjective. To check this point, we asked people to estimate the time spent per week contributing to Wikipedia as did Nov (2007), knowing that, according to Glott et al. (ibid), the mean time spent creating content is 6.4 hours per week, but that 75% spend less than 4 hours per week).

Table 2 describes the relationship between the self-evaluation of regular and big contributors and the time spent per week on Wikipédia.

[Insert Table 2 here]

We observed a clear segmentation as 69% (resp. 14%) of the regular (resp. the big) contributors spend less than 5 hours per week contributing to Wikipédia. Symmetrically, 64% (resp. 13%) of the big (resp. the regular) contributors spend more than 10 hours per week contributing to Wikipédia[9]. Thus our self-assignment strategy controlled by the hours spent appeared solid enough to be used in our econometric model.

5. The results

5.1. Descriptive statistics

About 16,000 people responded to the survey and after cleaning the file of duplicates and incomplete answers, 13,386 responses were used. Among the people who answered, about two-thirds were non-contributors, and just over 12% regular or big contributors (see details in Table 3).

[Insert Table 3 here]

[9]However, the strong contributors who spend less than five hours per week on Wikipédia did not necessarily make a wrong estimation of their involvement, first because importance of the involvement and quantity of contribution are not perfectly correlated, and second because contributors can alternate periods of high contribution with more sporadic involvement, while considering themselves to be important contributors in total.

If it is hard to evaluate a response rate, the number of pages viewed during this period is between 650 and 690 million, the number of contributors around 65,000 and the number of active Wikipedians (Wikipedians who contributed 5 times or more in this month), around 5,000. As we captured a little over 1,500 answers from regular and big contributors, we estimated that we captured approximately one third of the active Wikipedians on Wikipédia. The very active Wikipedians (Wikipedians who contributed 100 times or more in this month) were around 700 during that period (14% of the active Wikipedians), where, in our sample, the big contributors represent 22% of the regular or big contributors. As previously discussed, these two categories do not overlap, but the same order of magnitude indicates that, if not representative, we have a significant number of very involved contributors.

Table 4 details the main descriptive statistics of the variables used in this paper, and Tables 5 and 6 the correlation between these variables.

[Insert Table 4 here]

[Insert Tables 5 and 6 here]

5.2. The econometric model

Our explanatory variables based on the first contribution can be observed only if the Wikipedians have already made a contribution (37.5% of our respondents, see Table 2). This makes possible an over (under) estimation of the dependent variable, due to the fact that some unobserved effects, which have a positive (negative) impact on the probability of having already made a contribution, may have the same impact on the probability of being involved at a more intense level of contribution. To overcome this potential bias, we used a two-stage Heckman procedure which first estimated the probability of being a contributor according to socio-demographic characteristics and a variable representing computer skills in managing complex documents (this variable, which

is excluded from the second equation, ensures the identification of the model[10]) . Our second step was to estimate an ordered probit model based on the increasing involvement in contribution, considering only those who had already made a contribution. The dependent variable is INT_CONTRIB ranging from 1 to 3 with 1 for an occasional contributor, 2 a regular contributor and 3 a big contributor. The convergence to the maximum likelihood in this system of equations can be complicated and computationally demanding. Roodman (2011) proposed a general tool implemented on Stata software which uses GHK algorithm to estimate a full-information maximum likelihood. The procedure models the errors of the two equations (selection equation and ordered probit on contribution) as jointly normally distributed, to control from the unobserved effects described above.

Table 7 displays the estimates of this ordered probit model with Heckman correction. Column (2) and (3) shows the result of the two-stage Heckman procedure with column (3) as the selection equation. Column (1) is the baseline model of the estimation of involvement in the contribution without taking into account the potential selection bias. To ensure that the self-evaluation of respondents about their contribution did not bias the results, we used the time spent on Wikipedia as an alternative to INT_CONTRIB. This information is available for those who have already made at least one contribution. One can see from Table 2 that respondents' self-evaluation and time spent on Wikipedia are strongly correlated, as almost 70% of regular contributors declared they spent less than 5 hours per week on Wikipedia while 86% of big contributors declared more than 5 hours. Column (4) describes the result of the ordered probit with Heckman correction (with column (5) as the selection equation) for

[10]We tried alternative variables to identify the Heckman selection model, like the skill "managing online identity" which may have a lower impact on the contribution. These changes didn't affect the results.

a dependent variable which ranges from 1 to 5, 1 representing less than 1 hour per week and 5 more than 20 hours per week.

The ordered probit specification previously used has some disadvantages, one of the most important being the parallel slope assumption which implies that the effect of an explanatory variable is the same whatever the dependent variable's category. In practice, this assumption is rarely true, which is why we also processed a regression considering a dichotomous variable for the level of contribution to Wikipedia. The variable REG_CONTRIB takes the value of 1 if the respondent is at least a regular contributor and 0 otherwise. The two-stage Heckman procedure is repeated with this binary variable in column (6) and (7).

The results of this estimate are consistent with those based on the self-evaluation of the contribution level.

[Insert Table 7 here]

5.3. Results

Our core objective was to test if the contributors were enrolled since the beginning, which means that the future involvement in the community can be observed in the characteristics of the first contribution, and to study the existence of a learning period before this first contribution. We first present the results regarding our hypotheses, the control variables, and the predicative capacity of our model. We discuss the robustness of our model regarding collinearity and endogeneity in Annexe 2.

5.3.1. Hypothesis 1.1. The significance of a major first contribution

While a minor contribution is weakly or not positively associated with the intensity of contribution, a first contribution which is significant in terms of investment, like writing or rearranging an article, is strongly positively associated with major involvement. This validates H 1.1.

5.3.2. Hypothesis 1.2. Making a first contribution is not a learning action

The will to improve an article and to bring new knowledge to the online encyclopedia is not statistically significant. However, those who made their first contribution to "test" the editing process or "out of curiosity" have a decreasing probability of becoming a big contributor. Considering the fact that the learning actions are associated with the decreasing probability of becoming a big contributor, we estimates that H 1.2. is also validated.

5.3.3. Hypothesis 2.1. The impact of mentoring

Neither the use of tutorials, nor the fact that the first contribution was made alone can be associated with stronger future involvement in contribution. More interesting is the result which strongly links the level of contribution with the fact of having been helped by a person during the first contribution. This result partially validates H 2.1., as the only positive impact is mentoring, while we hypothesized an impact of the tutorials too.

5.3.4. Hypothesis 2.2. A short learning process

The probability of getting involved in the contribution increases when the first contribution is made one month after having discovered the encyclopedia. This effect remains when the contribution is made in the first year. The result is robust for all the different specifications of the models. The marginal effect calculation shows that having contributed the first month increases the probability of becoming a regular contributor (respectively a big contributor) by 20% (respectively 21%). This contradicts what was expected in H 2.2.

5.3.5. Control. A socially determined contribution

The selection model used in the Heckman correction in Table 7 helps to define the profile of the contributors. It confirms, in line with the results found in the literature, that being a male, young or middle aged, educated and active, strongly increases the probability of being a contributor.

Considering the intensity of contribution, Table 7 enables to specify the social determinants for Wikipedians to make the choice to be more deeply involved in the community. As expected from previous studies, they are male, over 20, with a master's degree level or more but do not differ in the other social aspects mentioned above. The positive and significant coefficient associated with the time constraint variable in the selection equation shows that the Wikipedians who are engaged in a professional activity are more able to become contributors, but this time constraint has no effect on the level of involvement in the contribution. The stronger the involvement in the Wikipedia community, the more demanding these characteristics are: computer skills, plus experience and knowledge, are required to learn how to contribute to a collaborative project such as Wikipedia. Being a male, in his thirties, with a master's degree or more, are all reasons to expect the person to become an important contributor. However, these results challenge the previous findings that socio-demographic characteristics do not explain the level of investment in the project.

These results, notwithstanding their interest, are not at the core of the article's discussion and will not by analyzed further. The reader interested in the discussion between social capital and involvement in open online communities may find it interesting, in addition to Atkin et al. (1998); van Dijk and Hacker (2003), to read the discussion proposed by Jeppesen and Lakhani (2010), and Sue Gardner's blog[11] on the gender gap in Wikipedia.

We also have to note that, as expected, the number of years since the first contribution is negatively correlated with the contribution, suggesting that motivation decreases over time, confirming previous studies (Ortega and Izquierdo-Cortazar, 2009; Borzillo et al., 2011).

[11] http://suegardner.org/2011/02/19/nine-reasons-why-women-dont-edit-wikipedia-in-their-own-words/

5.3.6. Is the first contribution a good predictor of future involvement?

In this section we address the ability of the first contribution to predict future involvement.

The significance of the coefficient associated with the first contribution shows that early involvement can explain the level of future contribution, but doesn't necessarily improve the prediction. For simplicity, we tested the ability of our estimation to predict the outcome "being at least a regular contributor" (columns (6) and (7) in Table 7). If the predicted probability was above 0.5 we considered that the outcome was "being at least a regular contributor" and "not being a regular contributor" otherwise. We compared the model which considers the importance of the first contribution with a "naive" model where the predicted outcome is always 0, and with a model which only takes into account the control variables (age, gender, education, and professional activity variables). The different models and their results are presented in Table 8.

[Insert Table 8 here]

The model which considers the importance of the first contribution correctly predicts 72.6% of the outcomes, the naïve model 66.7%, and the model taking into account only the control variables 66.9%.There is a reduction of incorrect prediction by 18% for our model. The sensitivity and flexibility indicators enable to respectively predict the true positive and true negative outcomes. 36.6% of the regular contributors are correctly predicted by our model while they are only 14.2% in the control variables model. Even if there is still a high number of "false positives", introducing variables related to the first contribution behavior significantly increases the ability of the model to predict who will become big contributors. Finally we exhibit in table 8 the area under the ROC (Receiver Operating Characteristic) curve which measure the true positive rate over the false positive rate at different threshold value, a value larger than 0.5 suggest

a good predictive ability. The area under the ROC curve is 0.73 for our model and 0.65 for the control variables model.

This leads us to the discussion.

6. Discussion

6.1. Contribution to the literature

The standard theory based on the existence of a learning process inside the community would suggest that Wikipedians observe the community to learn the process and then start with a minor contribution before learning to improve their contribution in frequency or complexity. Conversely, our results suggest that the biggest contributors start to contribute with a major contribution (H1.1), which is not seen as part of a learning process (H1.2). They do not observe the community during a long period of time before proposing new knowledge (H.2.1), although we did not measure the effort put forth in observing the community during this (short) period. All these results go against the argument of a gradual contribution process, increasing in complexity toward regular contribution amongst the newcomers, which is at the core of the communities of practice theory and is illustrated by legitimate peripheral participation. Big contributors start big. These results remind that a prerequisite for contributing to an epistemic community is to be able to bring new knowledge (what Fleming and Waguespack, 2007 called the "culture of engineer"). And it seems that this capacity (to write a Wikipedia article) is learned outside the community (maybe in another community), maybe by observing it, and maybe intensively, but without interacting as in communities of practice. This gives credits to Edwards' proposition about a private sphere of learning (2001). They do not contradict the studies on social learning and intra-community dynamic of inclusion, but rather help to pinpoint the frontiers of this dynamic. Core contributors, even regular ones, are contributors from the beginning.

The result concerning mentoring (hypothesis 2.1) indicates that contributors have contributors in their social networks. We already know the importance of social interaction in usages of the Internet (DiMaggio et al., 2001), and in the context of Wikipedia's editing process, this result suggests also that a community of big contributors may pre-exist involvement in the Wikipedia project. It could be because contributors share common social characteristics but also because they know each other, are already part of (other online) common projects, or simply because the recruitment of new editors is spatially and/or socially a constraint, as contributions are geographically situated (Lieberman and Lin, 2009; Hecht and Gergle, 2010). It can also be due to the fact that integrating a group means learning tacit knowledge about its functioning, which is easier to transfer face-to-face. We will argue (without being able to demonstrate it here) that with the idea of mentoring comes the idea of co-optation, and that in Wikipedia, as in other epistemic communities such as the open source community (Fang and Neufeld, 2009), big contributors recruit their future peers amongst their acquaintances.

6.2. Managerial implications

Promoting and encouraging high value and sustainable contribution is a major issue for the convenors of communities, being online or offline. Our results suggest it is important to pay particular attention to the first contribution of a newcomer, the moment of this contribution and its intensity. However, in a practical way it depends on the type of contribution needed and thus on the type of community.

For Wikipedia managers, and more generally for knowledge production-oriented communities (epistemic communities), our results have a very practical consequence, and make the recommendations by Halfaker et al. (2011) more precise: the modification of a piece of knowledge (in this case, an article),

the creation of a new piece of knowledge (here, an article) by a newcomer (i.e. an anonymous or a newly registered person) must be closely monitored and encouraged, even if the new contributor does not publish by the rules. This contribution has to be viewed by the content project manager as a signal of willingness to contribute, and thus accepted or positively amended, so that the new contributor feels welcome. These mentoring programs should be carried out where the day-to-day management is conducted (Forte et al., 2009; Mateos-Garcia and Steinmueller, 2008), and knowledge contributions are received and evaluated. This would be the domain of the content project leader more than that of the global core managers, who would intervene later on. As Forte et al. (2012, table 1, p. 2) pointed out, this is exactly what these nested organizations are made for, in addition to production activity support: maintaining the group's well-being and providing support to members. Secondly, we propose that having big contributors present their work in high schools or colleges would be a good strategy to create social links with contributors amongst the more promising potential newcomers, as knowledge producers are of higher education and are recruited because they can provide knowledge [12].

On the other hand, when the transmission of coordination know-how matters more than knowledge production, like in a charitable project of people giving their time to distribute food to the homeless, socialization and learning by doing are key for the community. In these communities of practice and coordination, a gradual learning process (peripheral learning) matters. Therefore, convenors should pay attention to giving every newcomer the opportunity to socialize and to interact with the core members, facilitating these interactions and letting the future core members "emerge" by taking an increasing part in the collective

[12] In line with the Foundation's strategy regarding the Wikipedia Education Program, `http://en.wikipedia.org/wiki/Wikipedia:Education_program` or with the Google Summer of Code program.

action and becoming a major link in the community.

Open-source communities seem in between, as they do require knowledge contribution (the code), but seem to favor peripheral participation, too (Rullani and Haefliger, 2013). We would argue that the mechanisms are similar to those for Wikipedia, if one considers that there are two communities there: one being a community of practice, the user support groups (user forum, mailing list), dedicated to the practicing of the product; and the other, the development community, being an epistemic community. Again, apprenticeship takes place outside the community of developers and is used as a meeting forum to recruit peers, as described by (Rullani and Haefliger, 2013). At the same time, while in Wikipedia the articles are more or less independent, (big) open-source projects are made up of several interacting modules. For this reason, learning about the organization of the project is a type of technical (knowledge) learning too, which favors the process of knowledge learning after the first contribution and may mitigate the "big contributor from the beginning" process. This proposal should be tested in future research.

6.3. Limitations and future research

There are obvious limitations to our work. First, our survey addresses only one project (French Wikipédia). Taking into account the cultural variety on the practice of collective intelligence, especially when dealing with Wikipedia production, these results have to be confirmed in other language projects. The representativity of the sample may also be seen as a limitation. However, we are not trying to make a complete census of Wikipedia users' behavior, but to study a link between their first contribution and their level of involvement in the community. According to the existing literature (Kittur et al., 2007; Javanmardi et al., 2009), the 16,000 answers to our survey, and the distribution of the people according to their involvement (cf. Table 3), seem to be consistent

and a sufficiently large sample to lead to econometric analyses of such a link. Another limitation of our work may be that the level of involvement relies on declarative variables. Crowston et al. (2006); Lee et al. (2009), studying FLOSS and relying on Hackman (1987), showed the importance, as an output, of taking into account the producers' (or contributors') feedback and perception to have a global view of the output of such open online projects. It seemed to us that the level of involvement is also partly subjective and that a declarative variable may be as accurate as the (actually also declarative) number of edits. The advantage of asking for a level of involvement is that it does not require complex equations to aggregate the various ways people can contribute to the project. Its correlation with the (also declared) number of hours spent on Wikipédia proves we were right, at least for that case. Finally, our main focus was on the difference between occasional contributors and regular (or more than regular) contributors and our main results are not impacted by the uncertainty about the precision of our measurement of the level of contribution. However, questioning users about their first contribution, especially if it happened a long time ago, is not as accurate as observing this first contribution and may be subject to reporting bias. Monitoring anonymous or newly registered editors' contributions may be an answer in future research, to improve the key result of our work, the fact that it is possible to identify the main contributors from the beginning.

This result has to be examined further, especially outside Wikipedia. As there is no language barrier, it is easier to migrate from one open-source project to another than to migrate from the French Wikipedia to the German one, for instance. An experienced developer can be "hired" to contribute to a project, as some editors are "hired" to contribute to a thematic project in Wikipedia. Herraiz et al.'s work (2006) supports this hypothesis, showing that "volunteers tend to follow a step-by-step joining process [in open source projects], while

hired developers usually experience a "sudden" integration". We would advocate survey studies on open source project contributors to check if the same link exists between the very first contribution to an open-source project and the level of involvement in open-source in general. This should be done by controlling for the experience contributors acquired in other communities before joining the community under study, and by taking into account the hypothesis that, around the epistemic community which produces the code, there is the community of practice (i.e. product users).

7. Conclusion

Although an early important first contribution does not explain why contributors will become big editors, this behavior is certainly a proxy of their motivation and their willingness to get involved in the production of the online encyclopedia, and some of its characteristics are very observable: contributing in the first month after the discovery of Wikipedia, to modify, create or improve an article after asking somebody how this can be done, are all reasons to expect the person to become an important contributor.

There is little peripheral learning before the contribution because this is not a community of practice, but rather an online epistemic community, or a community of creation, where people self-select and signal their willingness to actively contribute, from the beginning.

After this commitment, people enter the community of regular contributors and start to socialize. This socialization process may be what differentiates big contributors from average ones. Pentzold (2011), for Wikipedia, and von Krogh et al. (2012), for open source, defend the idea that becoming a big contributor may be an additional step from being a regular contributor. This additional commitment would occur for reasons developed during the participation in the project, as the development of this sense of "community", i.e. understanding

and accepting the rules of the organization (Butler et al., 2008; Cardon, 2012). In that respect, our results (Table 9) show that the feeling of belonging to a community is strongly correlated with the level of contribution.

[Insert Table 9 here]

This seems rather standard. What people learn once hired in an organization is not they job but how to do it, the processes and the social organization of the group/organization they joined, in a word the organizational culture. This is what makes organization more than an addition of competences, the process that aligns people's interest with that of the collective (Hernandez, 2012), in an organization which seems rather close to Mintzberg and McHugh (1985)'s view of an adhocratic organization.

In a word, our results, and their limitations, in addition to confirming the fact that there is little peripheral learning in communities of creation, call for more research on the links between the different roles played by the regular contributors, the paths to reach these roles, and the characteristics of their contributions (production of knowledge, and also discussions, etc.), after the beginning. They also plead for longitudinal studies of the people who signal themselves proposing major contribution from the beginning, to see if their motivations change during their journey to regular and core contributors.

References

Amichai-Hamburger, Y., Lamdan, N., Madiel, R., Hayat, T., 2008. Personality characteristics of wikipedia members. CyberPsychology & Behavior 11, 679–681.

Amin, A., Roberts, J., 2008. Knowing in action: Beyond communities of practice. Research Policy 37, 353–369.

Antin, J., Cheshire, C., 2010. Readers are not free-riders: Reading as a form of participation on Wikipedia, in: Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, pp. 127–130.

Antin, J., Cheshire, C., Nov, O., 2012. Technology-mediated contributions: Editing behaviors among new wikipedians, in: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, ACM, New York, NY, USA. pp. 373–382.

Antin, J., Yee, R., Cheshire, C., Nov, O., 2011. Gender differences in Wikipedia editing, in: WikiSym 2011 Conference Proceedings - 7th Annual International Symposium on Wikis and Open Collaboration, ACM. pp. 11–14.

Atkin, D.J., Jeffres, L.W., Neuendorf, K.A., 1998. Understanding Internet Adoption as Telecommunications Behavior. Journal of Broadcasting & Electronic Media 42, 475–490.

Auray, N., Poudat, C., Pons, P., 2007. Democratizing Scientific Vulgarization. The Balance between Cooperation and Conflict in French Wikipedia. Observatorio (OBS*) Journal 3, 185–199.

Balka, K., Raasch, C., Herstatt, C., 2009. Open source enters the world of atoms: A statistical analysis of open design. First Monday 14.

Becker, H.S., 1960. Notes on the concept of commitment. American Journal of Sociology 66, 32–40.

Becker, H.S., 1963. Outsiders. Studies in the Sociology of Deviance. Free Press.

Borzillo, S., Aznar, S., Schmitt, A., 2011. A journey through communities of practice: How and why members move from the periphery to the core. European Management Journal 29, 25–42.

Bryant, S.L., Forte, A., Bruckman, A., 2005. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia, in: proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work, ACM.

Butler, B., Joyce, E., Pike, J., 2008. Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia, in: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM, New York, NY, USA. pp. 1101–1110. doi:http://doi.acm.org/10.1145/1357054.1357227.

Butler, B., Sproull, L., Kiesler, S., Kraut, R., 2007. Community effort in online groups: Who does the work and why?, in: Weisband, S. (Ed.), Leadership at a distance: Research in Technologically Supported Work. Lawrence Erlbaum, Mahwah, NJ.

Cardon, D., 2012. Discipline but not punish. The governance of Wikipedia, in: Massit-Folléat, F., Méadel, C., Monnoyer-Smith, L. (Eds.), Normative Experience in Internet Politics. Presses des Mines, Paris.

Cohendet, P., Créplet, F., Dupouet, O., 2001. Interactions between epistemic communities and communities of practice as a mechanism of creation and diffusion of knowledge, in: Zimmermann, J.B., Kirman, A. (Eds.), Interaction and Market Structure. Springer, Londres.

Crowston, K., 2011. Lessons from volunteering and free/libre open source software development for the future of work, in: Proceeedings of Researching The Future in Information Systems: IFIP Working Group 8.2 Working Conference, Future IS, Springer, Turku, Finland. p. 215. URL: http://crowston.syr.edu/system/files/ifipwg82paper_final.pdf.

Crowston, K., Howison, J., Annabi, H., 2006. Information system Success in Free and Open Source Software Development: Theory and Measures. Software Process Improvement and Practice 11, 123–148.

Defélix, C., Chanal, V., Galey, B., 2005. Les personnes innovantes dans les entreprises doivent-elles faire l'objet d'une GRH spécifique ? Une étude exploratoire. Gestion 2000 mars-avril, 99–113.

van Dijk, J., Hacker, K., 2003. The digital divide as a complex and dynamic phenomenon. The Information Society 19, 315–326.

DiMaggio, P., Hargittai, E., Neuman, W.R., Robinson, J.P., 2001. Social implications of the Internet. Annual Review of Sociology 27, 307–336.

Edwards, K., 2001. Epistemic Communities, Situated Learning and Open Source Software Development, in: Cultures and the Practice of Interdisciplinarity' Workshop at NTNU, p. 24.

Fang, Y., Neufeld, D., 2009. Understanding Sustained Participation in Open Source Software Projects. Journal on Management Information Systems 25, 9–50.

Feller, J., Fitzgerald, R., Hissam, S., Lakhani, R.K. (Eds.), 2005. Perspectives on free and open source software. MIT Press, New York.

Fleming, L., Waguespack, D.M., 2007. Brokerage, Boundary Spanning, and Leadership in Open Innovation Communities. Organization Science 18, 165–180.

Forte, A., Bruckman, A., 2005. Why do people write for Wikipedia? Incentives to contribute to open-content publishing. working paper .

Forte, A., Kittur, N., Larco, V., Zhu, H., Bruckman, A., Kraut, R.E., 2012. Coordination and beyond: social functions of groups in open content production, in: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, ACM, New York, NY, USA. pp. 417–426.

Forte, A., Larco, V., Bruckman, A., 2009. Decentralization in wikipedia governance. Journal of Management Information Systems 26, 49–72.

Glott, R., Schmidt, P., Ghosh, R., 2010. Wikipedia Survey – Overview of Results. Technical Report. UNU-MERIT. URL: http://www.wikipediasurvey.org/docs/Wikipedia_Overview_15March2010-FINAL.pdf.

Guimera, R., Uzzi, B., Spiro, J., Amaral, L.A.N., 2005. Team assembly mechanisms determine collaboration network structure and team performance. Science 308, 697–702.

Hackman, J.R., 1987. The design of work teams, in: Lorsch, J.W. (Ed.), Handbook of Organizational Behavior. Prentice Hall, Englewood Cliffs, NJ, pp. 315–342.

Halfaker, A., Geiger, R.S., Morgan, J.T., Riedl, J., 2013. The rise and decline of an open collaboration system how Wikipedia's reaction to popularity is causing its decline. American Behavioral Scientist 57, 664–688.

Halfaker, A., Kittur, A., Riedl, J., 2011. Don't bite the newbies: How reverts affect the quantity and quality of wikipedia work, in: Proceedings of the 7th International Symposium on Wikis and Open Collaboration, ACM, New York, NY, USA.

Hasan, H., Pfaff, C., 2006. The wiki: an environment to revolutionise employees'

interaction with corporate knowledge, in: Proceedings of the Australasian Computer-Human Interaction Conference, OZCHI 2006, Sydney.

Hecht, B.J., Gergle, D., 2010. On the "localness" of user-generated content, in: Proceedings of the 2010 ACM conference on Computer supported cooperative work, ACM, New York, NY, USA. pp. 229–232.

Hemetsberger, A., Reinhardt, C., 2009. Collective Development in Open-Source Communities: An Activity Theoretical Perspective on Successful Online Collaboration. Organization Studies 30, 987–1008.

Hernandez, M., 2012. Toward an Understanding of the Psychology of Stewardship. The Academy of Management Review (AMR) 37, 172–193.

Herraiz, I., Robles, G., Amor, J.J., Romera, T., González-Barahona, J.M., 2006. The processes of joining in global distributed software projects, in: GSD '06: Proceedings of the 2006 international workshop on Global software development for the practitioner, ACM, New York, NY, USA. pp. 27–33. doi:http://doi.acm.org/10.1145/1138506.1138513.

Hertel, G., Niedner, S., Herrmann, S., 2003. Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. Research Policy 32, 1159–1177.

Hess, C., Ostrom, E., 2006a. Introduction: An Overview of the Knowledge Commons, in: (Hess and Ostrom, 2006b) (Ed.), Understanding Knowledge as a Commons. From Theory to Practice, pp. 3–26.

Hess, C., Ostrom, E. (Eds.), 2006b. Understanding Knowledge as a Commons. From Theory to Practice. MIT Press.

Javanmardi, S., Ganjisaffar, Y., Lopes, C., Baldi, P., 2009. User contribution

and trust in Wikipedia, in: 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing.

Jeppesen, L.B., Lakhani, K.R., 2010. Marginality and problem-solving effectiveness in broadcast search. Organization science 21, 1016–1033.

Kamata, M., Kato, D., Kunieda, K., Yamada, K., 2010. Web community contributor's motivation: Japanese wikipedia case study, in: Proc. of the IADIS Int. Conf. Collaborative Technologies 2010, Proc. of the IADIS Int. Conf. Web Based Communities 2010, Part of the MCCSIS 2010, pp. 29–33.

Kittur, A., Chi, E.H., Pendleton, B.A., Suh, B., Mytkowicz, T., 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie, in: CHI 2007, San Jose, CA, USA.

Kram, K.E., 1983. Phases of the mentor relationship. The Academy of Management Journal 26, pp. 608–625.

von Krogh, G., Haefliger, S., Spaeth, S., Wallin, M.W., 2012. Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development. MIS Quarterly 36, 649–676.

Lakhani, K., von Hippel, E., 2003. How open source software works: Free user to user assistance. Research Policy 32, 923–943. URL: `http://opensource.mit.edu/papers/lakhanivonhippelusersupport.pdf`.

Lakhani, K., Wolf, R., 2005. Why hackers do what they do: Understanding motivation and effort in free/open source software projects, in: Perspectives on free and open source software. in Feller et al. (2005) ed., pp. 3–22.

Lam, S.K., Uduwage, A., Dong, Z., Sen, S., Musicant, D.R., Terveen, L., Riedl, J., 2011. WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance,

in: Proceedings of the 7th International Symposium on Wikis and Open Collaboration, ACM, New York, NY, USA.

Lee, S.Y.T., Kim, H.W., Gupta, S., 2009. Measuring open source software success. Omega 37, 426 – 438.

Lerner, J., Tirole, J., 2002. Some simple economics of open source. Journal of Industrial Economics 50, 197–234.

Liang, C., Chen, C., Hsu, Y., 2008. The participation motivation and work styles of the administrators for chinese wikipedia. Journal of Educational Media and Library Science 46, 81–110.

Lieberman, M., Lin, J., 2009. You are where you edit: Locating Wikipedia users through edit histories, in: ICWSM '09, pp. 106–113.

Mahr, D., Lievens, A., 2012. Virtual lead user communities: Drivers of knowledge creation for innovation. Research Policy 41, 167 – 177.

Maillart, T., Sornette, D., Spaeth, S., von Krogh, G., 2008. Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution. Physical Review Letters 101, 218701.

Mateos-Garcia, J., Steinmueller, W.E., 2008. The institutions of open source software: Examining the Debian community. Information Economics and Policy 20, 333–344.

McAfee, A.P., 2006. Enterprise 2.0: The Dawn of Emergent Collaboration. Management of Technology and Innovation 47.

Mintzberg, H., McHugh, A., 1985. Strategy formation in an adhocracy. Administrative Science Quarterly 30, 160–197.

Musicant, D.R., Ren, Y., Johnson, J.A., Riedl, J., 2011. Mentoring in Wikipedia: A Clash of Cultures, in: Proceedings of the 7th International Symposium on Wikis and Open Collaboration, ACM, New York, NY, USA.

Nov, O., 2007. What motivates wikipedians? Communications of the ACM 50, 60–64.

Oliver, P., Marwell, G., Teixeira, R., 1985. A theory of critical mass interdependence, group heterogeneity, and the production of collective action. American Journal of Sociology 91, 522–556.

O'Mahony, S., Bechky, B.A., 2008. Boundary organizations: Enabling collaboration among unexpected allies. Administrative Science Quarterly 53, 422–459.

O'Mahony, S., Ferraro, F., 2007. The emergence of governance in an open source community. Academy of Management Journal 50, 1059–1106.

Ortega, F., Izquierdo-Cortazar, D., 2009. 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, IEEE Computer Society, Washington, DC, USA. pp. 7–12.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press.

Panciera, K., Halfaker, A., Terveen, L., 2009. Wikipedians are born, not made: a study of power editors on Wikipedia, in: Proceedings of the ACM 2009 international conference on Supporting group work, ACM, New York, NY, USA. pp. 51–60.

Pentzold, C., 2011. Imagining the Wikipedia community: What do Wikipedia

authors mean when they write about their "community"? New Media & Society 13, 704–721.

Prasarnphanich, P., Wagner, C., 2009. The role of wiki technology and altruism in collaborative knowledge creation. Journal of Computer Information Systems 49, 33–41.

Roodman, D., 2011. Fitting fully observed recursive mixed-process models with CMP. The Stata Journal 11, 159–206.

Rullani, F., Haefliger, S., 2013. The periphery on stage: The intra-organizational dynamics in online communities of creation. Research Policy 42, 941–953.

Shah, S.K., 2006. Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development. Management Science 52, 1000–1014.

Spence, M., 1973. Job Market Signaling. Quarterly Journal of Economics 87, 355–374.

Uzzi, B., 2008. A social network's changing statistical properties and the quality of human innovation. Journal of Physics A: Mathematical and Theoretical 41, 224023, 12pgs.

Uzzi, B., Spiro, J., 2005. Collaboration and creativity: The small world problem. American Journal of Sociology 111, 447–504.

Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F., 2003. User acceptance of information technology: Toward a unified view. MIS Quarterly 27, 425–478.

von Hippel, E., 1986. Lead users: a source of novel product concepts. Management Science 32, 791–805.

Von Krogh, G., Spaeth, S., Lakhani, K.R., 2003. Community, joining, and

specialization in open source software innovation: A case study. Research Policy 32, 1217–1241.

Voss, J., 2005. Measuring Wikipedia, in: Proceedings of the International Conference of the International Society for Scientometrics and Informetrics (ISSI), Stockholm.

Wenger, E., 1998. Communities of practice: Learning, Meaning, and Identity. Cambridge University Press, Cambridge, Mass.

Wenger, E., 2006. Communities of practice, a brief introduction. Technical Report. URL: http://www.ewenger.com/theory/communities_of_practice_intro.htm.

Worthen, B., 2008. Why most online communities fail. The Wall Street Journal July 16th.

Yang, H.L., Lai, C.Y., 2010. Motivations of Wikipedia content contributors. Computers in Human Behavior 26, 1377 – 1383.

Annexe 1. Tables

Table 1: Description of the variables.

VARIABLES	Details of building
CONTRIB	1 if ever made a contribution on Wikipedia, 0 otherwise
INT_CONTRIB	Ranging from 1 to 3 with 1 for an occasional contributor, 2 a regular contributor and 3 a big contributor.
HOURS	Ranging from 1 to 6 with 1 representing 1 hour or less on Wikipedia and 6 representing more than 20 hours per week on Wikipedia.
GENDER	1 being a male, 0 otherwise
AGE15	1 if age less than 16, 0 otherwise
AGE20	1 if age [16-20[, 0 otherwise
AGE30	1 if age [21-30[, 0 otherwise
AGE40	1 if age [31-40[, 0 otherwise
AGE50	1 if age [41-50[, 0 otherwise
AGE+	1 if more than 50 years old, 0 otherwise
ACTIVITY	1 if has a professional activity, 0 otherwise
ACTIVITY_FIRST	1 if had a professional activity when doing the first contribution, 0 otherwise
EDUCATION1	1 if High school level, 0 otherwise
EDUCATION2	1 if Mid-undergraduate high school level, 0 otherwise
EDUCATION3	1 if Undergraduate, 0 otherwise
EDUCATION4	1 if Graduate or more, 0 otherwise
EDUCATION5	1 if Professional diploma, 0 otherwise
MINOR_CONTRIB	1 if made a minor first contribution (spelling mistake fixing addition of a reference).
MAJOR_CONTRIB	1 if made a major first contribution (writing, translation, modification of an article).
MOTIV_CURIOSITY	1 if first contribution motivated by « curiosity », 0 otherwise
MOTIV_TEST	1 if first contribution made to « test » the editing process, 0 otherwise
MOTIV_IMPROVE	1 if first contribution made to improve an article, 0 otherwise
FIRST_CONT1	1 if time period between discovery of Wikipedia and first contribution is one month or less, 0 otherwise
FIRST_CONT2	1 if time period between discovery of Wikipedia and first contribution is one year or less, 0 otherwise
FIRST_CONT3	1 if time period between discovery of Wikipedia and first contribution is more than one year, 0 otherwise
HOW_ALONE	1 if first contribution made without looking for explanation, 0 otherwise
HOW_TUTO	1 if first contribution made reading tutorial on Wikipedia, 0 otherwise
HOW_PEER	1 if first contribution made asking somebody for help, 0 otherwise
YEAR	Number of year since the first contribution
COMPLEX_DOC	1 if Ability to manage complex documents

Table 2: Relationship between the self-evaluation of contribution and the time spent on Wikipedia.

Time spend on Wikipédia (in hour per week)	Share of regular contributors	Share of big contributors
< 1	0.31	0.02
[1 ;5[	0.38	0.12
[5 ;10[	0.18	0.22
[10 ;20[	0.09	0.32
≥ 20	0.04	0.32

Table 3: Distribution of the surveyed according to their level of contribution to Wikipédia.

Non contributors	Occasional contributors	Regular contributors	Big contributors
62 %	25%	9%	3%

Table 4: Descriptive statistics.

Name of the variable	Non contributors (8,363 obs, Mean)	Occasional Contributors (3,354 obs, Mean)	Regular contributors (1,288 obs, Mean)	Big Contributors (381 obs, Mean)
GENDER	0.62	0.79	0.87	0.88
AGE15	0.12	0.1	0.05	0.04
AGE20	0.22	0.25	0.14	0.09
AGE30	0.29	0.36	0.36	0.31
AGE40	0.1	0.12	0.15	0.25
AGE50	0.08	0.07	0.11	0.13
AGE+	0.18	0.1	0.18	0.18
ACTIVITY	0.82	0.89	0.85	0.8
ACTIVITY_FIRST	n.a.	0.93	0.91	0.87
EDUCATION1	0.3	0.24	0.14	0.09
EDUCATION2	0.13	0.12	0.1	0.07
EDUCATION3	0.18	0.17	0.15	0.15
EDUCATION4	0.17	0.18	0.2	0.19
EDUCATION5	0.21	0.29	0.41	0.5
FIRST_CONT1	n.a.	0.05	0.18	0.36
FIRST_CONT2	n.a.	0.24	0.36	0.35
FIRST_CONT3	n.a.	0.71	0.46	0.28
MINOR_CONTRIB	n.a.	0.76	0.81	0.68
MAJOR_CONTRIB	n.a.	0.69	0.79	0.8
MOTIV_CURIOSITY	n.a.	0.39	0.31	0.31
MOTIV_TEST	n.a.	0.26	0.17	0.13
MOTIV_IMPROVE	n.a.	0.93	0.95	0.9
HOW_TUTO	n.a.	0.44	0.56	0.48
HOW_PEER	n.a.	0.05	0.07	0.09
HOW_ALONE	n.a.	0.79	0.73	0.75
YEAR	n.a.	6.15	6.25	6.1

Table 5: Correlations for the socio-demographic variables.

	GENDER	AGE15	AGE20	AGE30	AGE40	AGE50	AGE+	ACTIVITY	EDUCATION1	EDUCATION2	EDUCATION3	EDUCATION4	EDUCATION5
AGE16	-0.0912	1											
AGE20	-0.0124	-0.1605	1										
AGE30	0.0449	-0.2339	-0.3923	1									
AGE40	0.0249	-0.1221	-0.2047	-0.2985	1								
AGE50	0.0059	-0.0921	-0.1544	-0.2251	-0.1175	1							
AGE+	-0.0028	-0.1146	-0.1921	-0.2801	-0.1462	-0.1102	1						
ACTIVITY	0.0187	0.1072	0.1754	0.0695	-0.001	-0.042	-0.3794	1					
EDUCATION1	-0.0674	0.5614	0.288	-0.3073	-0.1759	-0.101	-0.1249	0.0806	1				
EDUCATION2	-0.0064	-0.0772	0.2936	-0.074	-0.0959	-0.041	-0.0571	0.0197	-0.1819	1			
EDUCATION3	0.0188	-0.1344	0.0859	0.0866	-0.0273	-0.0131	-0.0795	0.0213	-0.2263	-0.1562	1		
EDUCATION4	0.0107	-0.1483	-0.2036	0.1839	0.0345	0.0269	0.0538	-0.0147	-0.2462	-0.1699	-0.2114	1	
EDUCATION5	0.0389	-0.2048	-0.3439	0.0947	0.2091	0.1027	0.1639	-0.0874	-0.3604	-0.2488	-0.3095	-0.3367	1
FIRST_CONT1	0.0405	0.0043	-0.0927	-0.038	0.0903	0.0341	0.0454	-0.0498	-0.0406	-0.033	-0.0097	-0.008	0.0715
FIRST_CONT2	0.0636	-0.0296	-0.0427	0.0155	-0.0248	0.0282	0.0587	-0.0217	-0.0569	0.0276	-0.0017	0.0233	0.0126
FIRST_CONT3	-0.0838	0.022	0.0998	0.0108	-0.035	-0.0486	-0.0828	0.0501	0.0759	-0.0038	0.0093	-0.0167	-0.0565
MINOR_CONTRIB	0.0134	0.0079	0.0148	0.012	-0.0069	-0.0109	-0.0267	0.0196	-0.0102	0.0154	-0.0122	0.0026	0.006
MAJOR_CONTRIB	0.0234	0.0091	0.0192	-0.0251	-0.0122	-0.004	0.0212	-0.0094	-0.0063	0.0087	0.0053	0.0196	-0.0208
MOTIV_CURIOSITY	0.0058	0.0394	0.0661	-0.0298	-0.0246	-0.0163	-0.0336	0.017	0.0803	0.0167	-0.0115	-0.0007	-0.071
MOTIV_TEST	-0.0033	0.1487	0.1173	-0.0702	-0.0674	-0.0547	-0.0559	0.0214	0.1829	0.0261	-0.0068	-0.0277	-0.1472
MOTIV_IMPROVE	0.03	-0.0156	0.039	0.0265	0.0048	-0.0338	-0.0511	0.0206	-0.0337	0.0201	-0.0026	0.0005	0.0173
HOW_TUTO	0.0009	-0.1253	-0.1209	-0.0396	0.0775	0.0989	0.1545	-0.0993	-0.1248	-0.0163	-0.0009	0.0629	0.0673
HOW_PEER	-0.0655	0.0071	-0.032	-0.0435	0.0182	0.0059	0.0743	-0.0843	0.0103	-0.0206	-0.0135	0.0075	0.0094
HOW_ALONE	0.0221	0.0832	0.0779	0.0567	-0.0464	-0.0277	-0.1818	0.1142	0.0569	0.0037	0.017	-0.0286	-0.0413
YEAR	-0.0291	0.2261	0.1257	-0.0912	-0.1391	-0.0222	-0.0546	0.0469	0.2403	0.0506	0.0224	-0.0548	-0.2137

Table 6: Correlations for the other variables.

	FIRST_ CONT1	FIRST_ CONT2	FIRST_ CONT3	MINOR_ CONTRIB	MAJOR_ CONTRIB	MOTIV_ CURIOSITY	MOTIV_ TEST	MOTIV_ IMPROVE	HOW_ TUTO	HOW_ PEER	HOW_ ALONE
FIRST_ CONT2	-0.2115	1									
FIRST_ CONT3	-0.4303	-0.7842	1								
MINOR_ CONTRIB	-0.0297	0.0555	-0.033	1							
MAJOR_ CONTRIB	0.0374	0.0292	-0.0523	-0.208	1						
MOTIV_ CURIOSITY	-0.0158	-0.0319	0.0406	0.0172	-0.0378	1					
MOTIV_ TEST	-0.046	-0.0138	0.041	-0.0218	0.042	0.2797	1				
MOTIV_ IMPROVE	-0.0671	0.0179	0.0271	0.1491	0.0755	-0.1915	-0.1426	1			
HOW_ TUTO	-0.0032	0.0367	-0.0327	-0.0021	0.0732	0.027	-0.077	-0.0134	1		
HOW_ PEER	0.0295	0.0028	-0.0233	-0.042	0.0155	0.0009	-0.0058	-0.0399	0.0496	1	
HOW_ ALONE	0.0141	0.0018	-0.01	0.0511	-0.02	-0.0272	0.0279	0.0461	-0.4111	-0.2353	1
YEAR	-0.0024	0.0712	-0.0635	-0.0478	-0.0185	0.0221	0.0815	-0.0247	-0.0692	0.0343	-0.0054

Table 7: Ordered probit with Heckman selection.

Dependant variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	INT_CONTRIB	INT_CONTRIB	CONTRIB	HOURS	CONTRIB	REG_CONTRIB	CONTRIB
GENDER	0.294***	0.445***	0.593***	0.371***	0.592***	0.464***	0.593***
	(0.0503)	(0.0697)	(0.0258)	(0.0678)	(0.0258)	(0.0708)	(0.0258)
AGE15	-0.373***	-0.266**	0.283***	-0.316***	0.283***	-0.345***	0.282***
	(0.110)	(0.117)	(0.0602)	(0.111)	(0.0602)	(0.125)	(0.0602)
AGE 20	-0.310***	-0.188**	0.316***	-0.295***	0.316***	-0.266***	0.315***
	(0.0820)	(0.0951)	(0.0485)	(0.0899)	(0.0485)	(0.102)	(0.0485)
AGE 30	-0.178***	-0.0926	0.244***	-0.165**	0.244***	-0.145*	0.244***
	(0.0615)	(0.0697)	(0.0395)	(0.0662)	(0.0395)	(0.0753)	(0.0395)
AGE 40	0.0173	0.0913	0.249***	0.0248	0.248***	-0.0129	0.247***
	(0.0696)	(0.0723)	(0.0467)	(0.0706)	(0.0466)	(0.0805)	(0.0466)
AGE 50	0.0586	0.116	0.191***	0.0952	0.191***	0.0790	0.190***
	(0.0772)	(0.0769)	(0.0516)	(0.0751)	(0.0516)	(0.0847)	(0.0516)
AGE+	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
ACTIVITY	-0.0581	-0.0252	0.0724**	-0.0740	0.0722**	0.0175	0.0730**
	(0.0566)	(0.0562)	(0.0367)	(0.0551)	(0.0367)	(0.0603)	(0.0367)
EDUCATION1	0.00756	0.0430	0.111**	0.0848	0.112**	0.0450	0.111**
	(0.0792)	(0.0768)	(0.0436)	(0.0758)	(0.0436)	(0.0807)	(0.0436)
EDUCATION2	0.0709	0.114	0.127***	0.104	0.127***	0.0841	0.126***
	(0.0756)	(0.0736)	(0.0424)	(0.0732)	(0.0424)	(0.0780)	(0.0424)
EDUCATION3	0.133*	0.218***	0.264***	0.175**	0.264***	0.192**	0.264***
	(0.0788)	(0.0809)	(0.0458)	(0.0799)	(0.0458)	(0.0857)	(0.0458)
EDUCATION4	0.278***	0.416***	0.459***	0.356***	0.459***	0.391***	0.460***
	(0.0745)	(0.0851)	(0.0437)	(0.0834)	(0.0437)	(0.0897)	(0.0437)
EDUCATION5	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
FIRST_CONT1	1.110***	1.053***		0.999***		1.114***	
	(0.0562)	(0.0721)		(0.0592)		(0.0786)	
FIRST_CONT2	0.466***	0.441***		0.429***		0.447***	
	(0.0411)	(0.0441)		(0.0408)		(0.0455)	
FIRST_CONT3	Ref.	Ref.		Ref.		Ref.	
MINOR_CONTRIB	0.0738	0.0730*		0.0452		0.159***	
	(0.0451)	(0.0425)		(0.0423)		(0.0464)	
MAJOR_CONTRIB	0.287***	0.278***		0.269***		0.315***	
	(0.0438)	(0.0424)		(0.0416)		(0.0452)	
MOTIV_CURIOSITY	-0.0844**	-0.0782**		-0.0603		-0.0951**	
	(0.0405)	(0.0385)		(0.0383)		(0.0412)	
MOTIV_TEST	-0.231***	-0.220***		-0.214***		-0.218***	
	(0.0489)	(0.0470)		(0.0465)		(0.0494)	
MOTIV_IMPROVE	-0.0110	0.00527		0.0327		0.0807	
	(0.0760)	(0.0721)		(0.0719)		(0.0774)	
HOW_TUTO	0.0504	0.0488		0.0440		0.0993**	
	(0.0404)	(0.0381)		(0.0380)		(0.0413)	
HOW_PEER	0.237***	0.221***		0.290***		0.235***	
	(0.0782)	(0.0745)		(0.0735)		(0.0824)	
HOW_EASY	-0.0575	-0.0481		-0.0444		-0.0557	
	(0.0473)	(0.0449)		(0.0445)		(0.0486)	
YEAR	-0.0224**	-0.0188**		-0.0256***		-0.0160	
	(0.00962)	(0.00902)		(0.00249)		(0.00990)	
COMPLEX_DOC			0.179***		0.180***		0.179***
			(0.0120)		(0.0120)		(0.0120)
Constant			-1.979***		-1.983***	-33.91*	-1.980***
			(0.0654)		(0.0653)	(19.82)	(0.0653)
Log Likelihood	-3655	-11886		-13543		-11002	
Observations	13,353	13,353		13,353		13,353	

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table 8: Predicted outcomes

	Prediction if outcome is always 0	Model without first contribution variables	Model with first contribution variables
Sensitivity	0%	14.2%	36.6%
Flexibility	100%	93.3%	90.4%
Correctly classified	66.7%	66.9%	72.6%
Area under ROC curve	-	0.65	0.73

Table 9: Correlation between a feeling of belonging to the community and level of contribution.

	Non-contributors		Regular Contributors		Big Contributors	
You feel you belong to the	Raw	Partial	Raw	Partial	Raw	Partial
"Wikipedia community"	-0.116**	-0.108**	0.135**	0.114**	0.161**	0.188**

** significant at 1 percent.

Table 8 displays correlation coefficients between categories of (non) contributors and a dummy variable which stands for the feeling of being part of a "Wikipedia community". For each "raw" column the result is a simple coefficient correlation. The partial correlation enables us to check for the effect of other variables in the relationship between the level of contribution and the sense of belonging to the Wikipedia community. For the non contributors, control variables are sex, age, level of education and time constraint, for the regular and big contributors all the variables displayed in Table 1 are used in the partial correlation.

Annexe 2. Robustness check. Dealing with endogeneity and multicollinearity

The different regression in table 7 showed that our result is robust to different specification of the dependent variable, but we also need to pay particular attention to classical problems of biased estimators. First we can suspect multicollinearity of variables dedicated to the first contribution. Tables 5 & 6 show no sgnificant correlation between these variables. In addition, we reestimate the two-stage Heckman procedure sequentially introducing information on the form (major or minor contribution), the reason for the first contribution and how it was made. Results are presented in Table 10, bellow. The sequential introduction of additional information concerning the first contribution neither changes

the main estimates, nor significantly increases the standard error of estimators, supporting the assumption that our explicative variables are independent enough.

Table 10: sequential introduction of addition information in the model.

Dependant variable	(1)	(2)	(3)	(4)	(5)
	INT_CONTRIB	INT_CONTRIB	CONTRIB	HOURS	CONTRIB
GENDER	0.408***	0.440***	0.459***	0.445***	0.593***
	(0.0706)	(0.0679)	(0.0655)	(0.0697)	(0.0258)
AGE15	-0.343***	-0.314***	-0.257**	-0.266**	0.283***
	(0.114)	(0.115)	(0.114)	(0.117)	(0.0602)
AGE 20	-0.266***	-0.230**	-0.190**	-0.188**	0.316***
	(0.0931)	(0.0956)	(0.0946)	(0.0951)	(0.0485)
AGE 30	-0.162**	-0.122*	-0.102	-0.0926	0.244***
	(0.0688)	(0.0700)	(0.0696)	(0.0697)	(0.0395)
AGE 40	0.0380	0.0732	0.0849	0.0913	0.249***
	(0.0729)	(0.0719)	(0.0710)	(0.0723)	(0.0467)
AGE 50	0.0739	0.101	0.112	0.116	0.191***
	(0.0774)	(0.0760)	(0.0750)	(0.0769)	(0.0516)
AGE+	Ref.	Ref.	Ref.	Ref.	Ref.
ACTIVITY	-0.0369	-0.0246	-0.0290	-0.0252	0.0724**
	(0.0562)	(0.0555)	(0.0553)	(0.0562)	(0.0367)
EDUCATION1	0.0604	0.0588	0.0449	0.0430	0.111**
	(0.0771)	(0.0755)	(0.0750)	(0.0768)	(0.0436)
EDUCATION2	0.137*	0.130*	0.113	0.114	0.127***
	(0.0740)	(0.0723)	(0.0717)	(0.0736)	(0.0424)
EDUCATION3	0.231***	0.232***	0.220***	0.218***	0.264***
	(0.0812)	(0.0795)	(0.0788)	(0.0809)	(0.0458)
EDUCATION4	0.414***	0.439***	0.416***	0.416***	0.459***
	(0.0856)	(0.0832)	(0.0825)	(0.0851)	(0.0437)
EDUCATION5	Ref.	Ref.	Ref.	Ref.	Ref.
FIRST_CONT1	1.087***	1.043***	1.023***	1.053***	
	(0.0657)	(0.0758)	(0.0789)	(0.0721)	
FIRST_CONT2	0.470***	0.445***	0.435***	0.441***	
	(0.0425)	(0.0450)	(0.0458)	(0.0441)	
FIRST_CONT3	Ref.	Ref.	Ref.	Ref.	
MINOR_CONTRIB		0.0639	0.0631	0.0730*	
		(0.0409)	(0.0411)	(0.0425)	
MAJOR_CONTRIB		0.275***	0.275***	0.278***	
		(0.0413)	(0.0416)	(0.0424)	
MOTIV_CURIOSITY			-0.0717*	-0.0782**	
			(0.0373)	(0.0385)	
MOTIV_TEST			-0.218***	-0.220***	
			(0.0461)	(0.0470)	
MOTIV_IMPROVE			0.000589	0.00527	
			(0.0701)	(0.0721)	
HOW_TUTO				0.0488	
				(0.0381)	
HOW_PEER				0.221***	
				(0.0745)	
HOW_EASY				-0.0481	
				(0.0449)	
YEAR				-0.0188**	
				(0.00902)	
COMPLEX_DOC					0.179***
					(0.0120)
Constant					-1.979***
					(0.0654)
Log Likelihood	-11928	-11904	-11886	-11876	
Observations	13,353	13,353	13,353	13,353	

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

The second potential issue is the endogeneity of our econometric specifica-

tion. Reverse causality and omitted variables are typically the main sources of endogeneity. If the former cannot be a major concern in our estimation, for obvious timing reasons, the latter has to be considered. To address this endogeneity problem we use an instrumental variable approach (IV) which first estimates the early first contribution (FIRST_CONTRIB) which is suspected of endogeneity, and then estimates the involvement in contribution (INT_CONTRIB) using the estimated values of FIRST_CONTRIB. Identification conditions require that the instrument variable present in the estimation of endogenous variable is correlated with FIRST_CONTRIB (relevance condition) but not with the INT_CONTRIB (exogeneity condition). We use the binary variable ACTIVITY_FIRST as an instrument, which equals 1 when the respondent had a professional activity at the time of the first contribution. We consider that this instrument satisfies both identification conditions. Being employed at the time of the first contribution should be correlated with an early first contribution and should be independent of the current contribution. The result of the regression with the instrumental variable is available in Table 11, the main conclusion is that the variable suspected of endogeneity (FIRST_CONTRIB) is still positive and strongly significant, suggesting that endogeneity should not be a major concern in our estimation. We used FIRST_CONTRIB as a potential source of endogeneity but we could have used any variable related to first contribution behavior without changing the result.

Table 11: Ordered probit with Heckman selection (3).

Dependant variable	(1)	(2)
	INT_CONTRIB	INT_CONTRIB
GENDER	0.523***	0.406***
	(0.0385)	(0.0531)
AGE15	-0.161*	0.154
	(0.0842)	(0.108)
AGE 20	-0.0765	-0.197**
	(0.0628)	(0.0907)
AGE 30	-0.0302	-0.0342
	(0.0484)	(0.0683)
AGE 40	0.00406	0.254***
	(0.0569)	(0.0738)
AGE 50	0.106*	0.120
	(0.0621)	(0.0844)
AGE+	ref.	ref.
ACTIVITY	0.0712	0.182***
	(0.0460)	(0.0675)
EDUCATION1	0.126**	0.121
	(0.0622)	(0.0904)
EDUCATION2	0.156***	0.230***
	(0.0590)	(0.0830)
EDUCATION3	0.302***	0.184**
	(0.0615)	(0.0865)
EDUCATION4	0.477***	0.389***
	(0.0582)	(0.0791)
EDUCATION5	ref.	ref.
FIRST_CONT1	2.622***	
	(0.167)	
Activity_First		-0.971***
		(0.0883)
Constant	-2.206***	-1.534***
	(0.0821)	(0.114)
Log Likelihood	-6329	
Observations	13353	

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1