Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency

Kevin Crowston, Nicolas Jullien, and Felipe Ortega

> *"In the great mass of our people there are plenty individuals of intelligence from among whom leadership can be recruited."*
>
> Herbert Hoover
> 31st President of the United States

Extensive research has been conducted over the past years to improve our understanding of sustainability conditions for large-scale collaborative projects, especially from an economic and governance perspective. However, the influence of recruitment and retention of participants in these projects has received comparatively less attention from researchers. Nevertheless, these concerns are significant for practitioners, especially regarding the apparently decreasing ability of the main open online projects to attract and retain new contributors. A possible explanation for this decrease is that those projects have simply reached a mature state of development. Marwell and Oliver (1993; tinyurl.com/bapafxc) and Oliver, Marwell, and Teixeira (1985; tinyurl.com/bal2y5y) note that, at the initial stage in collective projects, participants are few and efforts are costly; in the diffusion phase, the number of participants grows, as their efforts are rewarding; and in the mature phase, some inefficiency may appear as the number of contributors is greater than required for the work.

In this article, we examine this possibility. We use original data from 36 Wikipedias in different languages to compare their efficiency in recruiting participants. We chose Wikipedia because the different language projects are at different states of development, but are quite comparable on the other aspects, providing a test of the impact of development on efficiency. Results confirm that most of the largest Wikipedias seem to be characterized by a reduced return to scale. As a result, we can draw interesting conclusions that can be useful for practitioners, facilitators, and managers of collaborative projects in order to identify key factors potentially influencing the adequate development of their communities over the medium-to-long term.

Introduction

Mobilizing hundreds of contributors, as in the case of Linux, to thousands of contributors, as in the case of Wikipedia, open online communities (tinyurl.com/bbhkeuc) are viewed as a central point for innovative generation of new knowledge (Chesbrough, 2003: tinyurl.com/ce6bsy8; Mahr and Lievens, 2012: tinyurl.com/akozt3b). Open source initiatives are numerous and they span various industries (Balka et al., 2009; tinyurl.com/yaxxs3a). The success of such projects provides a new perspective on a fundamental socio-economic question in today's society: the sustainability of participation in collective action. The Olson paradox (Olson, 1965; tinyurl.com/bdj4usa) suggests that large groups are less able than small ones to promote their common interest because individual incentives to contribute diminish with group size. However, many communities, of all sizes and in various contexts, have shown their ability to develop selective incentives and institutions, making them able to develop and protect their "commons" (Ostrom, 1990: tinyurl.com/b3neybk; Hess and Ostrom, 2006: tinyurl.com/bamczvb). The question then is if these communities will be able to sustain their activities over the long term.

Recruiting and retaining new members is a recurrent issue for open communities, a topic already stressed and

Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency

Kevin Crowston, Nicolas Jullien, and Felipe Ortega

studied by von Krogh, Spaeth, and Lakhani (2003; tinyurl.com/atwpr4k) in the case of open source software community inflow. A warning sign for the sustainability of open communities is the apparent increased difficulty of recruiting and retaining new members, which has been observed for the Wikipedia project in particular (Ortega, 2009; tinyurl.com/ahhvu55). However, the diversity of projects makes it difficult to assess whether these concerns are justified and how broadly they apply. To have better evidence of the situation, we need to understand better how production is organized in such projects. However, comparing open source projects is complicated because they use different techniques (e.g., languages), different technical systems to support cooperation (e.g., version control systems), different management structures, and so on. In this article, we focus on the methodological aspect of the measurement of the efficiency to propose and validate a methodology before applying it to complex-to-compare projects. We therefore sought a setting that would provide a greater degree of comparability between projects.

Wikipedia (wikipedia.org) is an interesting setting for our research, for several reasons. First and foremost, Wikipedia is a large and successful online community, comparable in many ways to open source software projects. However, as noted above, it is also a project that has experienced an apparent slow-down in the recruitment of new editors, raising the question of sustainability. Second, the structure of Wikipedia lends itself to a comparison of efficiency. Wikipedia maintains a separate version of the encyclopedia in different languages. Each version has an independent collection of articles maintained by its own community of editors (though nothing other than language fluency prevents an editor from contributing to more than one Wikipedia language). Importantly, these communities of editors have reached different levels of maturity. Some communities are quite mature, whereas others are still getting started and yet others fall somewhere in between. However, they all share the same tool for collaborative editing (MediaWiki; mediawiki.org) and the same basic rules to guide this cooperative editing: the "five pillars" of Wikipedia (tinyurl.com/bhs3m). As well, if we measure the global structure of these communities as a network in which the articles are the nodes and the hyperlinks to other articles connect these nodes, it seems to be approximately similar, at least for the largest Wikipedias (Zlatic and Stefancic, 2011; tinyurl.com/andcqo7). In contrast to studies on open source software (e.g., Crowston et al., 2006: tinyurl.com/a3wec75; Koch, 2009: tinyurl.com/a74aqj7) that compare projects that use various technologies, programming languages or collaborative tools, this uniformity may help us to better understand which differences can be correlated with process evolution.

The article is organized as follows: first, we define the inputs and the outputs to be evaluated in our analysis of efficiency and our analysis approach, multiple-input multiple-output efficiency techniques (specifically data envelopment analysis). Then, we present the data and our current results. We discuss these results in the last section and present some conclusions that can be useful for practitioners, managers, and facilitators in these kind of open communities to assess their current evolution and prevent negative factors that could influence proper development of these communities in due course.

Theory Development

This analysis focuses on a comparison of the 39 largest Wikipedias (according to the official article count provided by Wikimedia Foundation, which is displayed on the home page of each version). The unit of analysis for our study is a Wikipedia community writing in a specific language (e.g., French, German, Japanese). However, we decided not to include the English Wikipedia in this analysis, because it is a prominent outlier regarding many aspects. It is by far the largest Wikipedia by number of articles, with four times more entries than the next language (the German Wikipedia), so we were concerned that it would have too great an influence on our results. It also exhibits a much broader community, attracting editors from the five continents, as it has become the default language version for many contributors and readers. As such, it is difficult to define the population from which English-language editors are drawn, which is a necessary step in our analysis.

Given this sample of projects, we assess the efficiency of the different Wikipedia communities in each language to turn their readers (inputs) into contributors (outputs). Research has shown that a mix of experienced editors and fresh newcomers increases the likelihood for an article to reach the top quality, or "Feature Article", level in Wikipedia (Ransbotham and Kane; 2011: tinyurl.com/azbxulp; Bryant et al., 2005: tinyurl.com/a3h3d6x; Arazy et al., 2011: tinyurl.com/allx4j8). Thus, the output of the recruitment process is the number of editors (of different types, described below) contributing to the project. We take as input the number of potential contributors, also described below.

Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency

Kevin Crowston, Nicolas Jullien, and Felipe Ortega

Economists formalize the link between inputs and outputs as a production function, literally a mathematical function giving the amount of outputs of a process for a given amount of inputs. Efficiency is outputs divided by inputs; to optimize efficiency means to obtain the maximum possible outputs for a given amount of inputs. In our case, the form of this production function is unknown, as are the coefficients relating its components. However, we are not trying to propose a characterization of the Wikipedia production function, but rather to evaluate if communities in different languages are more (or less) efficient than others. Following Farell (1957; tinyurl.com/b7apobr), the relative efficiency of different producers can be compared by examining the "frontier production function". This function describes, for various combinations of inputs and outputs, which producers are efficient. In other words, efficiency refers to the members of a sample of producers who have the highest outputs for a particular mix of inputs. Note that this definition of efficiency is relative rather than absolute; there is not some theoretical sense behind the term "efficiency". An additional consideration in analyzing the efficiency of production is the question of "return to scale", that is, whether a big project may be more efficient because of its size (e.g., in a larger and better known project, it is easier to attract new producers) or perhaps less efficient because of the overhead of coordinating more participants.

There are several techniques for estimating the frontier production function. A detailed comparison is out of the scope of this paper, but interested readers are referred to (Kitchenham, 2002; tinyurl.com/bgd2z4j) for a more complete discussion of these techniques regarding software production. We used the data envelopment analysis (DEA) models originally proposed by Charnes, Cooper, and Rhodes (1978; tinyurl.com/b2tuxpz), following Koch's (2009; tinyurl.com/a74aqj7) use in the case of open source software. Koch noted that "these models were developed to measure the efficiency of non-profit units, in which neither clear market prices for their inputs and outputs exist, nor a clear evaluation for their relations" (p. 403). In addition, "DEA can account for economies or diseconomies of scale, and is able to deal with multi-input, multi-output systems in which factors have different scales" (p. 398). These characteristics made DEA an appropriate technique for our comparison of different Wikipedia projects.

Data

External data (inputs)

To estimate the input to the recruitment process, we need data on the number of potential editors for each Wikipedia in a different language. We consider this group as the number of people with a tertiary education, who speak that language and have access to the Internet. The rationale for this choice can be found in Glott, Schmidt, and Ghosh (2010; tinyurl.com/66zazh5), as well as in a survey on the French Wikipedia (Dejean and Jullien, 2012; tinyurl.com/bz6x7zn), showing that Wikipedia contributors are significantly more educated than readers. To estimate the Internet population, we retrieved data from Internet World Stats (internetworldstats.com). This site aggregates Internet usage data from several sources, including "data published by Nielsen Online, by the International Telecommunications Union, by GfK, local Regulators and other reliable sources". Data are available at the language level for Chinese, Spanish, Japanese, Portuguese, German, Arabic, French, Russia, and Korean. For other cases (Dutch, Hungarian, Persian, Romanian, Bulgarian, Croatian, and Greek), we calculated the total number of users by multiplying the Internet rate in the main countries speaking the language by the population of these countries plus the population of minorities speaking the language by the Internet rate in the other countries where the language is spoken. A similar procedure has been conducted for the number of people with a tertiary-level education by language. The primary data for this measure comes from UNESCO (tinyurl.com/blx7n6f) for most of the countries in the study and the OECD for Russia (tinyurl.com/ahogygv) and China (tinyurl.com/b3ubalt). Of course, these sources provide only an approximation of desired input variables, but they are our best estimates. However, drastic inaccuracy in these estimates would in turn affect our productivity estimation.

Wikipedia data collection (outputs)

As in prior studies of Wikipedia (e.g., Wilkinson and Huberman, 2007: tinyurl.com/bjgge7x; Ortega et al., 2007: tinyurl.com/auwcneq; Ortega et al., 2009: tinyurl.com/bfb3spm), we relied on the database dumps published by the Wikimedia foundation. These databases contain complete records (date and time, author, etc.) of every single contribution that comes in the form of a "revision" to any page in any of the 39 Wikipedias under study. Thus, it is possible to count the number of active editors per month and break them down in three groups, following the definitions offered by Wikimedia Foundation (stats.wikimedia.org/EN/): very active Wikipedians (those with 100 or more revisions in a given month); active Wikipedians (between 5 and 100 revisions in a given month), and other contributors (those with fewer numbers of edits in a certain month). For this step, data extraction has been implemented as a software program that is part of WikiDAT (Wikipedia Data Analysis Toolkit; tinyurl.com/aykvdbt).

Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency

Kevin Crowston, Nicolas Jullien, and Felipe Ortega

Analysis approach: DEA modelling

We must contemplate two main criteria regarding the choice of a DEA model: its orientation (input-oriented or output-oriented) and the return to scale in the production process. Regarding the first criteria, as in (Koch, 2009; tinyurl.com/a74aqj7), an output-orientation seems to be more appropriate given that, for a certain period of time, the inputs (the population of volunteers potentially joining a Wikipedia in a certain language) are more or less fixed and the goal is to maximize the output. As for the second criteria, considering the study on collective action (Marwell and Oliver, 1993; tinyurl.com/bapafxc), the analysis of software projects, and our previous discussion, it seems rather difficult to assume a constant return to scale. Instead, these projects seem to have an increasing return to scale in a first phase, and then a decreasing one. Hence, we use the BCC-O (output-oriented) model (Banker, Charnes, and Cooper, 1984; tinyurl.com/bxv62wy) that lets us assess the return to scale. For the data analysis, we adapted Sadiq's (2011; tinyurl.com/bxadd9r) macro under SAS.

Findings

An exploratory plot of our datasets shows a strong (but not perfect) correlation between the total number of Wikipedia contributors, the Internet population (Figure 1), and total tertiary-educated members of the population (Figure 2). Using the DEA model, we can identify different levels of efficiency in the conversion of these inputs to the Wikipedia community of contributors of different kinds. We first apply a constant return to scale model, then we introduce the possibility of a variation in return to scale. The results for this analysis are shown in Figure 3. The projects are listed in decreasing order of size. The bars indicate the relative efficiency. The longest bars, representing 100% efficiency, correspond to projects on the efficient frontier, that is, those that create the most outputs from their particular combination of inputs. Shorter bars represent projects that use a similar mix of inputs but produce comparatively fewer outputs than other projects. Specifically, certain Wikipedias, such as Malaysian (ms), Arabic (ar), and Chinese (zh), have many fewer editors than would be suggested by the population of Internet users who could become editors, whereas Estonian (et), Hungarian (hu), Norsk (no), and Finnish (fi) show high efficiency in recruiting editors. As far as the return to scale is concerned, Table 1 presents the sign of the return to scale variable. It seems that the largest and most efficient projects exhibit decreasing return to scale, suggesting increased difficulty in recruiting new Wikipedians. On the other hand, when they are efficient, the smaller Wikipedias seem to be still in an increasing return to scale phase.

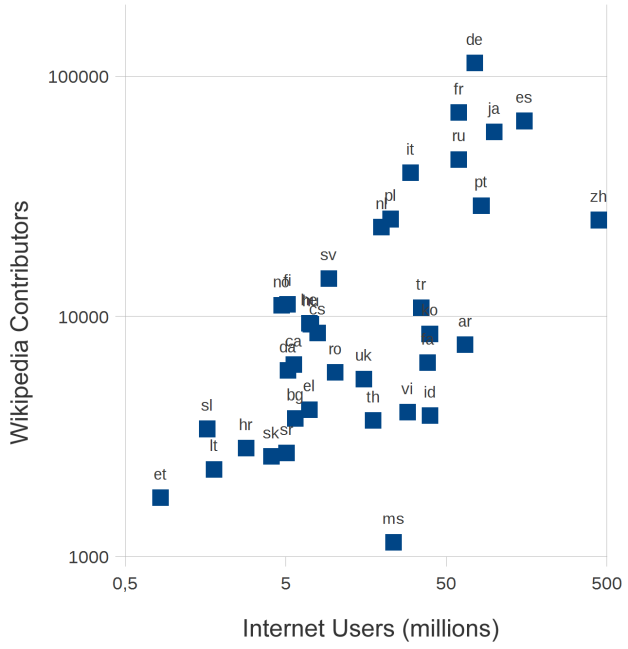

Figure 1. Number of contributors versus Internet population

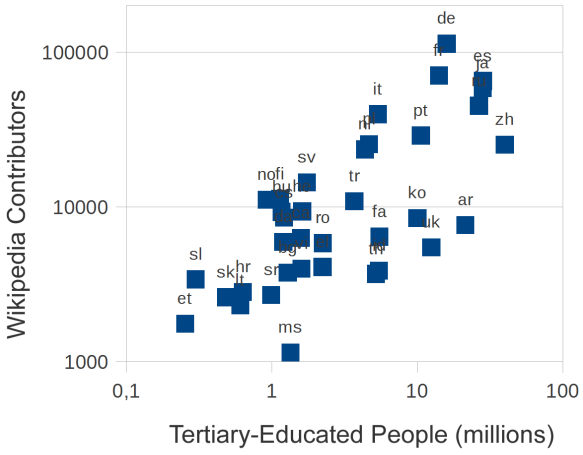

Figure 2. Number of contributors versus population with a tertiary education

Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency

Kevin Crowston, Nicolas Jullien, and Felipe Ortega

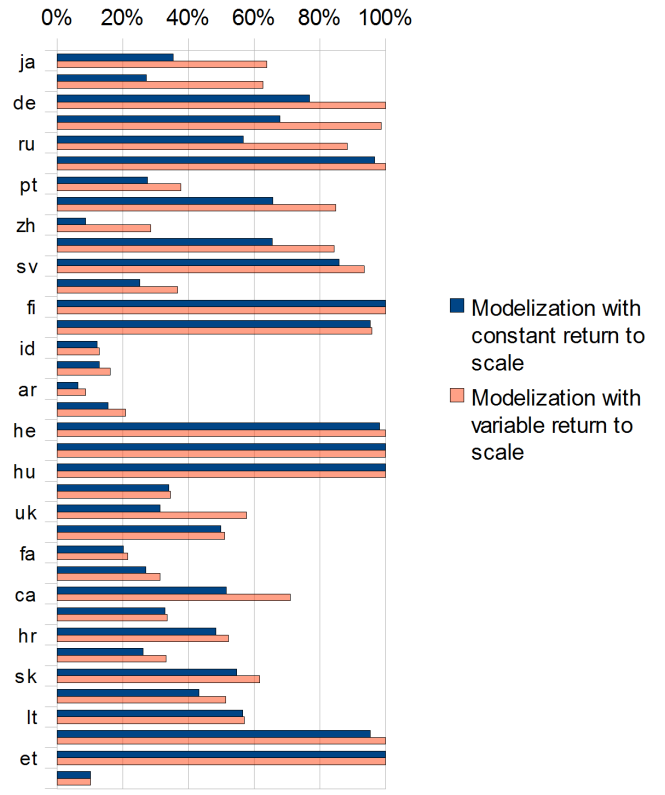

Figure 3. Efficiency in recruitment of contributors. (Projects are listed in decreasing order of size.)

Conclusion

The work presented here provides an initial step to identifying differences in the work practices of the various Wikipedia projects, shedding light on the sustainability of such collective intelligence projects, and it proposes a way to extend the work initiated by Stvilia, Al-Faraj and Yi (2009; tinyurl.com/bdlounp), Hara, Shachaf, and Hew (2010; tinyurl.com/aajhf93), and Callahan and Herring (2011; tinyurl.com/b3pjrff). Our analysis indicates that the size and maturity level of the project matters, because the largest Wikipedias are assessed by this model as being inefficient (that is, recruiting proportionally fewer new editors for a given mix of potential participants than other projects with a comparable mix). If we add a factor to control for return to scale, the largest projects increase their performance, but display a negative return to scale. In other words, the larger projects are demonstrably in a phase where they are less able to recruit new members. Furthermore, the analysis reveals striking differences in efficiency among the smaller projects, which presumably are otherwise at similar states of development.

Table 1. Return to scale for the recruitment of contributors. Efficient projects are highlighted in bold-italics and red font.

Project		
Japanese	ja	-1.57
Spanish	es	-1.60
German	de	***-0.04***
French	fr	-0.11
Russian	ru	-0.12
Italian	it	***-0.12***
Portuguese	pt	-0.17
Polish	pl	-0.14
Chinese	zh	-0.15
Dutch	nl	-0.10
Swedish	sv	-0.45
Turkish	tr	-0.29
Finnish	fi	***-0.03***
Czech	cs	-2.19
Indonesian	id	-0.65
Thai	th	-0.38
Arabic	ar	-0.73
Korean	ko	-0.09
Hebrew	he	***-0.08***
Norwegian	no	***0.02***
Hungarian	hu	***-0.14***
Vietnamese	vi	-0.36
Ukrainian	uk	-0.64
Danish	da	-1.31
Farsi	fa	-0.62
Romanian	ro	-0.15
Catalan	ca	-0.78
Bulgarian	bg	-0.49
Croatian	hr	0.34
Greek	el	-0.75
Slovak	sk	0.42
Serbian	sr	-0.16
Lithuanian	lt	0.13
Slovenian	sl	***0.15***
Estonian	et	***-0.19***
Malaysian	ms	-0.56

Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency

Kevin Crowston, Nicolas Jullien, and Felipe Ortega

The results of our analysis are suggestive, but clearly represent just a first step. While we have shown differences in efficiency, we do not yet fully understand why these differences arise. The next step of the research will be to find better explanations for these differences. There can be many possible explanations for difficulties in recruitment, but research literature on open source software projects (Koch, 2008; tinyurl.com/b6hcrll) and on collective action more generally (Marwell and Oliver, 1993; tinyurl.com/bapafxc) suggests that such a slow-down may simply happen as a result of the project entering a mature phase in which fewer additions, and thus fewer contributors, are required. Nevertheless, a more troubling possibility is that the evolution of the projects has led to the development of working patterns that make contributing to these projects more difficult. This scenario could make participants' work less rewarding (Ransbotham and Kane; 2011; tinyurl.com/azbxulp), raising invisible barriers to contributions from outsiders and new members (to take Ostrom's perspective) and thus threatening the long-term sustainability of the project. Distinguishing these possibilities for the larger projects is important to understanding their prospects.

However, explaining the differences among the smaller projects requires more nuanced explanation. While the current data do not provide an answer, we hypothesize two possible explanations. First, many of the less efficient projects have a lower level of tertiary-educated people compared to the efficient group. This difference could be a key to explaining the low efficiency of recruitment. A second speculation regards the effects of control of information: many of the low-efficiency projects are tied to countries where the Internet and the production of information is more closely controlled by the authorities than in the efficient group. It may be that freedom of expression is pre-requisite for efficient recruitment of editors. Zhang and Zhu's (2011; tinyurl.com/afqraut) recent study on the Chinese Wikipedia gives arguments for this hypothesis.

Better understanding these differences should provide insight for the long-term sustainability of both Wikipedia as well as other open knowledge-creation projects. In particular, the first hypothesis suggests that these projects are dependent on the investments made in education by the countries in which the projects are situated. Given the importance of the tertiary education variable, universities seem to be appropriate places to promote Wikipedia, which is in line with the Foundation's strategy regarding Wikipedia Education Program (tinyurl.com/9cqrh3r).

Another topic for future research is to address the limitations in the current study. A main limitation is that the validity of our analysis is dependent on the quality of the data used. In particular, the external data used for the inputs to the recruitment process are only best estimates. Systematic errors in these data would affect our measure of the relative efficiency of recruitment for the affected languages. On the other hand, while we are quite confident in the data extracted from the Wikipedia dumps, a limitation of the work presented here is that we evaluated the projects only for a single month, August 2011. Having only one month of data could lead to misinterpretations, especially taking into account that August is a vacation month in some countries. We are working on extending the analysis to twelve months and doing a mean estimation of the efficiency of the various projects.

Future research might also examine the transferability of the proposed methodology to open source software. In characterizing the open source production function, characteristics from the software engineering perspective such as the time to close bugs, the number of issue reports submitted, or activity in the mailing lists, may be of equal importance to the total number of contributors for evaluating the sustainability of a project. It is also important to consider the age of the project, as reflected in the return to scale effect. A complication we noted in the introduction to the article is that the diversity of open source projects makes it hard to compare them. One possible approach would be to compare different sub-projects within a larger project, which might control for variability in tools and processes.

Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency

Kevin Crowston, Nicolas Jullien, and Felipe Ortega

About the Authors

Kevin Crowston is a Distinguished Professor of Information Science at the Syracuse University School of Information Studies (aka the iSchool). He is currently on a temporary rotation as a Program Director for the Human-Centered Computing Program at the US National Science Foundation in the Information and Intelligent Systems Division of the Computer and Information Science and Engineering Directorate. His research examines new ways of organizing made possible by the extensive use of information technology.

Nicolas Jullien is an Associate Professor at the LUSSI Department of Telecom Bretagne (Brest, France). His research interests are on the organization and the attractiveness of open, online communities (Linux, Wikipedia). Most of his papers are available at: tinyurl.com/asfqzsm

Felipe Ortega is a Researcher in the Department of Statistics and Operations Research at University Rey Juan Carlos in Madrid, Spain. He is also a part-time Associate Professor at University Alfonso X El Sabio, teaching courses in the Information and Communication Technologies Department. His research is focused on open online communities, with emphasis on data retrieval, replicability, and data analysis.

Citation: Crowston, K., N. Jullien, and F. Ortega 2013. Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency. *Technology Innovation Management Review*. January 2013: 20-26.