4-1-2007

The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia

Denise Anthony
Dartmouth College

Sean W. Smith
Dartmouth College

Tim Williamson
Ning, Inc.

The Quality of Open Source Production: Zealots and Good Samaritans in the Case of *Wikipedia*

Denise Anthony,[1*] Sean W. Smith,[2] Tim Williamson [3]

Technical Report TR2007-606
Department of Computer Science
Dartmouth College

Version of April 2007

(A preliminary version of this paper was released online in November 2005)

KEY WORDS: collective goods, public goods, group identity, reputation, open source production

WORD COUNT: 6,016 (main text, notes, references)

TABLES: 6

FIGURES: 1

1 Department of Sociology, Dartmouth College, Hanover NH 03755
2 Department of Computer Science, Dartmouth College, Hanover, NH 03755
3 Ning, Inc., Palo Alto CA. This paper reports work done while a student at Dartmouth.

* To whom correspondence should be addressed. Denise Anthony, Department of Sociology, HB6104, Dartmouth College, Hanover, NH 03755; email: danthony@dartmouth.edu

The Quality of Open Source Production:
Zealots and Good Samaritans in the Case of *Wikipedia*

Abstract

New forms of production based in electronic technology, such as open-source and open-content production, convert private commodities (typically software) into essentially public goods. A number of studies find that, like in other collective goods, incentives for reputation and group identity motivate contributions to open source goods, thereby overcoming the social dilemma inherent in producing such goods. In this paper we examine how contributor motivations affect the *quality* of contributions to the open-content online encyclopedia *Wikipedia*. We find that quality is associated with contributor motivations, but in a surprisingly inconsistent way. Registered users' quality increases with more contributions, consistent with the idea of participants motivated by reputation and commitment to the *Wikipedia* community. Surprisingly, however, we find the highest quality from the vast numbers of anonymous "Good Samaritans" who contribute only once. Our findings that Good Samaritans as well as committed "zealots" contribute high quality content to *Wikipedia* suggest that it is the *quantity* as well as the *quality* of contributors that positively affects the quality of open source production.

Word count: 169

The Quality of Open Source Production:

Zealots and Good Samaritans in the Case of *Wikipedia*

I. Introduction

When we think about the revolution in information and communication technologies over the past decade we might fail to recognize some of the amazing organizational innovations that have also emerged (*cf.* Neff and Stark 2003; O'Mahony 2003). One of the most important of these organizational innovations is the emergence of "open source" production (also known as "open content"), defined as the free and open creation, alteration and distribution of goods, typically software, via the contributions from vast numbers of widely distributed and uncoordinated actors (Lakhani and Wolf 2005; Open Source Initiative 2005). Open source production is remarkable because it converts a private commodity (typically software) into essentially a public good (Kollock 1999; Kogut and Metiu 2001; O'Mahony 2003).[1] Indeed, advocates of open source software often describe it as a "movement," similar to social movements for other public goods (Raymond 2001; Stallman 1999; Torvalds and Diamond 2001).

Early studies of open source suggest that production is fueled by a small number of experts who contribute much of the content (Ghosh and Prakash 2000; Mockus et al 2005; Lerner and Tirole 2002; Lakhani and von Hippel 2002). According to this research, these experts are motivated by factors such as reputation and group identity, mechanisms identified by social scientists as capable of overcoming the social dilemma inherent in collective goods production. Here we move beyond examinations of what motivates contributors to ask, how are contributor motivations related to the *quality* of open source goods?

In general public goods are chronically under produced in society (Olson 1965). Given the inherent social dilemma in producing public goods (Olson 1965; Hardin 1968; Kollock 1998), open source production would seem to be based on a problematic and inefficient model. Some argue, however, that open source production can be not only efficient (Kogut and Metiu 2001), but even superior (e.g., von Hippel 2001; Weber 2005) to other forms of production. Indeed the success of open source software implies that open source production may be of superior quality to privately produced software (e.g., Mockus et al 2005; *cf.* Neumann 2005). Do the collective action mechanisms that motivate contributions to open source goods also explain the quality of those goods? In seeking to answer this question, this paper makes three contributions. First, we theorize the relation between contributor motivations in open source goods and quality using the case of the online, open-content encyclopedia, *Wikipedia.org*. Second, we use data from 7,058 contributors to *Wikipedia.org* to test hypotheses about contributor motivations and quality. Finally, we consider the implications for organizing collective action given our findings that suggest that it is both the *quantity* and *quality* of contributors that positively affects the *quality* of open source production goods.

II. The Case of Wikipedia

Wikipedia, the online, open content encyclopedia (www.wikipedia.org) is a compelling example of open source production. According to its Main Page, *Wikipedia* is "the free-content encyclopedia that anyone can edit." The English language version, started in 2001, currently has the most content with over 1.75 million articles (as of April 2007). *Wikipedia* describes itself as "a multilingual, web-based, free content

encyclopedia project. Wikipedia is written collaboratively by volunteers; its articles can be edited by anyone with access to the Internet" (http://en.wikipedia.org/wiki/Wikipedia). It has editions in roughly 200 different languages and contains entries both on traditional encyclopedic topics and on almanac, gazetteer, and current events topics.

Not only is *Wikipedia* content open access, but the creation *and* revision of the content is also entirely open such that anyone can add to or edit any entry. The precursor to *Wikipedia* was conceived by developers Jimmy Wales and Larry Sanger as a freely accessible encyclopedia, but the quality was to be ensured by seeking expert contributions evaluated by peer review (see Lih 2004; http://en.wikipedia.org/wiki/Wikipedia#History). In contrast, *Wikipedia* as it now exists succeeded by replacing professional contributions and expert peer review with their most democratic extremes: no proof of identity or qualifications is necessary in order to contribute or edit content.

As with any encyclopedia, the value of *Wikipedia* is the quality of its content, yet its overall quality is a much debated issue. In the few systematic studies comparing quality of content between *Wikipedia* and professionally produced encyclopedias, *Wikipedia* is found to be comparable in quality (Giles 2005; Lih 2004; *cf.* Encyclopedia Britannica 2006). Yet questions about quality persist. The concerns about quality in *Wikipedia*, in both popular press and scholarly accounts, focus on the nature and skills of the contributors and editors (Giles 2005; Encyclopedia Britannica 2006; Nature 2006; Orlowski 2005; Terdiman 2005; Wagstaff 2004). Given that the creation of its content is completely open, quality depends entirely on who contributes to *Wikipedia*. Yet, as noted by critics, why would any actor, let alone an expert, contribute?

One simple factor encouraging contributions to *Wikipedia* and other open source goods is the low cost of contributing (Lerner and Tirole 2002). The very 'wiki' technology used by *Wikipedia* reduces the costs of participation. A 'wiki' is an online document in which every edit made to it is saved as a unique document. *Wikipedia* is a collection of wiki-pages on specific topics for which the entire edit history of the topic is available. This means that any user can view past edits, add his or her own content, and even restore a previous version of the content. The formal policies of *Wikipedia*, as well as the wiki technology, help to limit (though not prevent) negative contributions, such as nonsense contributions or so-called graffiti attacks. For example, Ciffolilli (2003) argues that because *Wikipedia* is a wiki that saves all past versions of every article, it is very easy for friendly contributors to 'clean up' a damaged article. Research by IBM similarly shows that graffiti and damage to controversial topic pages are repaired quickly at *Wikipedia* (Wattenberg and Viegas 2003).

Beyond the cost factor, a plausible reason for the apparent high quality of *Wikipedia* is that contributors can *benefit* from participating, such as by building a reputation within the community. Reputation systems are powerful mechanisms for overcoming collective action problems (Cheshire and Cook 2004; Kollock 1998; Raub and Weesie 1990). Indeed, reputation systems are the basis for success of other new Internet-based institutions, such as the auction website *eBay* (Kollock 1999). Some researchers argue that reputation systems could be the basis for all secure Internet-based communication and exchange (e.g., Camp et al 2002; Cheshire and Cook 2004). In studies of various open source projects, one of the primary reasons cited for making contributions is the individual incentives of skill-development and building a reputation

(Kollock 1999; Lakhani and von Hippel 2003; Lakhani and Wolf 2005; Lerner and Tirole 2002; von Krogh et al 2003). Reputation mechanisms motivate participation in open source goods because they provide the basis for status in the community (Stewart 2005).

Wikipedia recognizes the power of reputation and encourages contributors to become 'registered users' by outlining the benefits of having an account (http://en.wikipedia.org/wiki/Wikipedia:Why_create_an_account, April 2007). According to *Wikipedia*, there are now over 4 million registered user accounts, "plus an unknown, but quite large, number of unregistered contributors" (http://en.wikipedia.org/wiki/Wikipedia:Wikipedians April 2007).

Though registered-user names are still merely 'cheap' pseudonyms (Friedman and Resnick 1999) that are easily abandoned and not necessarily tied to an individual's real identity, they provide a mechanism for establishing and tracking reputation. For any given subject in *Wikipedia*, users can view the history of contributions. A user can see edits that were contributed by registered *Wikipedians* (see below), while anonymous contributors have no name but merely have an IP address listed. An IP or Internet-Protocol address is a 32-digit number used to identify a computer or device on computer networks connected to the Internet. Clicking on a registered user name takes one to the "user's page," *Wikipedia*-space where registered users create personalized pages about themselves and their contributions to *Wikipedia*, if they choose to do so. *Wikipedia* even lists the top 1,000 contributors with the most edits, some of whom have been identified by name in the popular press (e.g., Terdiman 2005). Contributors with no interest in reputation can remain anonymous. Though anonymous users are listed by IP address only, it is possible to view the history of an IP address similar to a registered user, if more

than one contribution is made. As shown below, however, the majority of anonymous users have only one contribution.

In addition to reputation, some contributors are motivated by an apparent strong commitment to the community of "*Wikipedians*" (Giles 2005; http://en.wikipedia.org/wiki/Wikipedia_community, accessed 4/2/2007). The salience of a group identity can increase contributions to collective goods (Dawes 1980; Kramer and Brewer 1984; Dawes, van de Kragt and Orbell 1990; Turner and Tajfal 1986), including open-source projects (Raymond 1999) and virtual communities (Wellman and Gulia 1999), even though such groups often exist only in virtual 'online' space. Such contributors may even be "zealots", Coleman's (1990) term for true believers in a collective good who contribute for purely *intrinsic* value beyond rational expectations (see, e.g., Lakhani and Wolf 2005; Raymond 1999).

Wikipedia clearly presents itself as a community: "*Wikipedians* are the people who write and edit articles for *Wikipedia*…*Wikipedian* [] suggests someone who is part of a group or community. So in this sense, *Wikipedians* are people who form the *Wikipedia* Community" (http://en.wikipedia.org/wiki/Wikipedia:Wikipedians accessed April 2007). One of the top links on the main webpage is for the "Community Portal" which contains information about many different ways that users can participate in the community of *Wikipedia*.

According to this discussion, contributors to *Wikipedia* are motivated by two factors: (1) reputation and/or (2) commitment to the group identity of the *Wikipedia* community. How might these motivations influence the quality of contributions? Certainly motivations influence participation in *Wikipedia*. Contributors interested in

building a reputation will register since this is the only way to establish a reputation, while contributors with no interest in reputation will remain anonymous. Identity with the community, in contrast, has implications for the level of participation. Users who identify as *Wikipedians* participate in the community by making many contributions, while contributors who do not identify with the community will likely have few contributions.

It is straightforward to consider the quality implications for contributors at the intersection of strong interest in reputation and a strong *Wikipedia* identity, i.e., registered users with many contributions. They are the committed-expert contributors and zealots expected by advocates of open-source, and so are expected to have high quality contributions. The ability to identify and track the contributions of registered users, particularly over many contributions, also suggests such users interested in gaining a positive reputation will make many contributions. Moreover, we can also expect that registered users' with the most contributions will have the highest quality, else they would not be able to gain a positive reputation. This discussion suggests the following hypotheses:

Hypothesis 1a: Registered users will have more contributions than non-registered users.

Hypothesis 1b: Quality will be highest among registered users with many contributions, and

Hypothesis 1c: Quality will increase with participation (number of contributions) for registered users.

What are the implications for quality for anonymous contributors? Virtually all theories of social dilemmas would predict low quality from anonymous contributors, especially those with low levels of participation, since they would seem to have little motivation or incentives to contribute at all. Yet the lore of open-source suggests that anonymous one-time contributors are as important as the zealots. Who are these Good Samaritan contributors? They are likely to be of two types. The first type of Good Samaritans may be, like the zealots, experts in a particular field. These experts do not care about their reputation in *Wikipedia* (no registration), nor are they committed to *Wikipedia* as a community (few contributions). Instead they care about their area of expertise and so contribute to that topic only. Taking the time to register would increase the costs of contributing for these Good Samaritans, and since they are not interested in reputation and do not identify with the community itself, they have no reason to incur these costs. Given their expertise in the subject matter, however, their contributions will be of high quality.

The second type of Good Samaritan is simply the user who sees a mistake or a hole and makes a contribution to address it. These contributions are likely to be shorter and less substantive than others and so will be less likely to be edited or changed in the future.

In contrast to registered users whose quality is highest at high levels of participation, in both cases of the Good Samaritans, we expect that it is anonymous users with the fewest contributions that will have the highest quality.

But what are the quality implications for anonymous users with high levels of participation? As noted above, high participation levels suggest that the contributor

strongly identifies with the *Wikipedia* community. Why would a *Wikipedian* who strongly identifies with the community by participating at a high level choose to remain anonymous? One possibility might be that the multiple contributions from a single IP-address are not from the same contributor at all, but rather the result of proxies or dynamic IP-address allocation in some large companies and universities. Though plausible, it is unlikely that dynamic IP-addresses account for many contributors, in part because *Wikipedia* frequently blocks such networks and proxy servers.

Another possibility for why some anonymous users might have many contributions is that such users know their contributions are of low quality and do not want to be identified through a registered user name. Alternatively, many contributions may mean that these users are strongly committed to the *Wikipedia* community, but unlike the registered *Wikipedians* described above, their interest may be *negative* rather than positive. These would-be "hackers" may actively seek to contribute low-quality content to harm the community.

The motivations of anonymous contributors, including both Good Samaritans and high participation-anonymous contributors, leads to the following hypotheses:

Hypothesis 2a: Anonymous users will have shorter contributions than registered users.

Hypothesis 2b: Anonymous users with few contributions will have high quality content, and

Hypothesis 2c: Quality of contributions will decrease with participation for anonymous users.

We now turn to data from *Wikipedia* contributors to analyze these questions.

III. Data and Methods

We selected samples of *Wikipedia* contributors from the populations of both the French and Dutch language sites as of March 1, 2005.[2, 3] By March 2005 *Wikipedia* was well known, but significant debate over the quality of its content had not yet occurred (Encyclopedia Britannica 2006; Giles 2005). As of March 1 2005 there were a total of 53,901 contributors to the French language site and 33,217 contributors to the Dutch language site. The sampling procedure consisted of compiling a list of all contributors within each language group, then drawing two random draws within each language of up to 1,000 contributors for each user-type (registered and anonymous), for a total of n=7,058. (See Table 1.) Since registered users are over-represented in our sample compared to their distribution among all contributors, we weight the analyses based on the population proportions of each user-type within each language group.

Variables

We hypothesize that contributor motivations effect the quality of their contributions. That is, we are not measuring the quality of *Wikipedia* content *per se*, but rather the quality of *Wikipedia* contributors. We measure the quality of contributions *quantitatively* as the rate of each contributor's content retained in the current version of the topic article. Retention is only one quantitative dimension of the quality of a contribution, and likely a conservative measure to the extent that contributors and editors are *satisficing* (Simon 1957) rather than maximizing with regard to content, that is,

adding to or editing an entry until it is 'good enough' rather than until it is in some sense "perfect" or "complete."

The dependent variable is the *retention rate, R,* of contributions, measured as the percent of characters retained per contribution by each contributor. More specifically, we measure the number of characters retained, C, in a given article, summed across all edits (contributions), j, for each contributor, i, divided by the sum of the total number of characters, T, in each topic article edited per contributor.

$$R_i = \frac{\sum_{j=1}^{K} C_{ij}}{\sum_{j=1}^{K} T_{ij}}$$

For each contributor, we use the *Wikipedia* differencing algorithm[4] to compare the differences among three documents: (1) *edit*, the content submitted to each topic article by the contributor, (2) *previous*, the version of the article prior to the edit, and (3) *current*, the version of the article as it exists on the day the sample was drawn. *Edits* generally occur in time prior to the time point at which *current* is measured, so *current* does not in general equal *edit*, though it is possible if the contributor contributed all of the current content. We measure the retention of an edit by calculating the number of characters from a contributor's *edit* (comparing *edit* to *previous*) that are *retained* in the *current* version (comparing *edit* to *current*) as a percentage of the total number of characters in the article. For example, compare the following illustrative sentences, *previous*: "Public goods are unlike private goods;" e*dit*: "Public goods, in contrast to private goods, are non-excludable;" and *current*: "In contrast to private goods, public goods are non-excludable and non-rival." Comparing *edit* to *current*, we find that (when considering

longest common subsequences) 62 of the total 75 characters in the current version are retained for a *retention rate* of 83% (note that spaces are counted in the character count).

As illustrated in this example, a contributor's edit may include any of the following: added material, edited or deleted content, as well as content kept from the previous version. That means that our measure of *retention* includes all characters in the version 'submitted' by the contributor, no matter how much or how little of the content was added, deleted or changed by the contributor. The reasoning for this is that a contributor has the opportunity to add, edit or delete whatever she chooses, so preserving content from earlier versions is taken to mean at least tacit acceptance of its quality. It is important to note that *Wikipedia* requires that contributors edit on the granularity of whole entries. For example, the data structure does not permit "journaling" in which a contributor might submit an edit such as: "like before, except change sentence 23 as follows." The number of characters added, retained and total are pooled across all edits made by each contributor. Overall, the mean retention rate at *Wikipedia* is 72%. (See Table 2.)

We recognize that *retention rate* does not take into account all important features of content quality in *Wikipedia*, including, for example, "edit wars", in which two or more contributors continually change the content of a topic-entry, sometimes merely using the *wiki* to return the article to a previous version of the text. Other important factors that we cannot address are the amount of time lapsed between edits, or the status of the content, e.g., whether the topic being edited is "under construction" or in the parlance of *Wikipedia*, a "stub" in which only a very brief entry on the topic exists. These issues are most important when evaluating the quality of content itself, i.e., the

coverage of specific topic areas in which the history of the 'page' is important. In this study, however, we are interested in evaluating the quality of *contributors*, so we analyze their retention rates.

The key independent variables are whether a contributor is registered or anonymous and their number of contributions. Contributor registration status is measured by whether they have a *registered user name* or not. Number of contributions is measured as the number of times a contributor made an *edit*. On average, contributors made over 9 edits, with a range of 1-50 edits. Given the significant positive skew of this measure, we take the natural log in the analyses. Finally, our analyses also control for *language* area (French = 1, Dutch = 0), the total size of each article, measured as the total number of characters (natural log), and the size of the contribution, measured as the number of characters added per edit (natural log). Contribution size controls for the likelihood that the smaller the contribution the more likely it is to be a minor change and thus more likely to be retained. Article size controls for the possibility that registered and anonymous users contribute to fundamentally different types of *Wikipedia* topics. Since *Wikipedia* content is constantly evolving, at any given time there are many "new topics" with relatively small existing entries, as well as many well-established topics with a great deal of existing content. It may be that anonymous users are more likely to contribute only to well-established articles, or conversely only to newer topics with less existing content.

IV. Results

Table 3 shows the bivariate results for each variable by user type. Anonymous and registered users differ in important ways. Overall, registered users contribute more content across a greater number of edits compared to anonymous users, consistent with Hypotheses 1a and 2a. Surprisingly however, anonymous users contribute higher quality content overall compared to registered users. Given the expected motivations of reputation and identity among registered users, i.e., the zealots and committed experts, this is remarkable.

Table 4 shows the retention rates for contributors by the intersection of the two contributor motivations, reputation and commitment. Both committed experts and Good Samaritans have high quality contributions, supporting hypotheses 1b and 2b. Committed experts' (cell 1) contributions are of significantly higher quality compared to registered users with fewer contributions (cell 2). They are also significantly higher than anonymous users with similar numbers of contributions.

TABLES 3 AND 4 ABOUT HERE

Good Samaritans (cell 4 in Table 4) make the highest quality contributions overall. Good Samaritans contribute higher quality content than either registered users with similar levels of participation (cell 2), other anonymous users with more contributions (cell 3), and even registered users with many contributions (cell 1) though the latter is significant at the $p<.10$ level.

The bivariate results shown in Table 4 also suggest support for hypotheses 1c and 2c about the relationship between quality and contributions for different types of contributors. Figure 1 displays the estimated regression lines for the quality of contributions (retention rate) regressed on commitment (log number of contributions) for

both registered and anonymous users. Figure 1 shows that indeed quality changes with the extent of participation but in exactly the opposite direction for registered versus anonymous users. Anonymous users' quality is highest at low levels of commitment, and decreases as participation increases, while the opposite is true for registered users for whom quality increases with participation.

TABLES 5 AND 6 ABOUT HERE

We now turn to the multivariate analysis. Table 5 shows the results of multivariate regressions of the quality of contributions on levels of participation, controlling for article size, size of contribution and language, for registered and anonymous users. Hypotheses 1b and 2b are both supported in Table 5. Whereas *log edits* is positive for registered users, indicating increasing quality with increasing participation, it is negative for anonymous users.

It is important to note that the control variables are also significant in explaining the quality of contributions. The shorter a contribution is the higher its quality, for both registered and anonymous users. Quality is also higher when the topic article being edited is larger, regardless of the type of contributor. It may be that the larger a topic articles is, the more complete the information already included, so only those certain of their knowledge (i.e., experts, whether registered or anonymous) contribute to such articles. In addition, French contributors in general are less likely to have their contributions retained compared to Dutch contributors. We do not speculate as to why this may be the case.

FIGURE 2 ABOUT HERE

Another way to look at the relationship between quality and quantity for different types of contributors is to examine the effects among those with few contributions

compared to those with many. Table 6 shows the results of quality regressed on the type of user, controlling for the amount contributed, article size and language among those with fewer than five edits, and those with five or more edits. Consistent with the findings presented above, among those with fewer than five edits, registered users, compared to anonymous users (the omitted category), have significantly lower quality, but for those with five or more edits, registered users have higher quality.

V. Discussion and Conclusion

Why should we care about understanding the quality of *Wikipedia* contributions? One reason is that *Wikipedia* is becoming a "source of record" increasingly cited by mainstream print and news media (Lih 2004). For example, a search for *Wikipedia* in the top world newspapers in Lexis/Nexis for the period January 1-May 25, 2007 yielded 300 articles. (See also http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_media.) In part because of its exposure in mass media, readers of the *Wikipedia.org* website also are increasing dramatically. According to a website that tracks the traffic (number of visitors) to websites, www.alexa.com, the *Wikipedia* website ranks 10th in the Global Top 500 websites (accessed May 2007). As of October 2005 *Wikipedia* ranked as the top reference site (www.alexa.com).

While contributors to *Wikipedia* vary in their interests in reputation and their commitment to the group, *readers'* main interest is simply the *quality* of the content, *i.e.*, whether the material is accurate and reliable. *Wikipedia* readers, however, are highly uncertain about the quality of its content because they cannot rely on editors or publishers to screen for quality as they can when using a brand name encyclopedia. Readers'

uncertainty may lead them to look at types of contributors for different signals of quality, such as registration or high levels of participation. A registered user name provides access to the history of contributions for that contributor (*i.e.,* reputation), and as such, readers may look to a contributor's history, or even take registration itself, as a signal of quality. Alternatively, readers may consider that a strong identity in *Wikipedia* is necessary for quality content, and so expect that only those with many contributions (*i.e., Wikipedians*, whether registered or not) will contribute high quality content. To the extent that readers look for the intersection of registration and high participation, our analysis suggests they will indeed find high quality content from the committed expert contributors. Either signal alone, however, suggests they will not find high quality material. Further, attention to these signals alone may hinder readers from recognizing the high quality contributions of Good Samaritans who contribute one-time only and anonymously.

A more important reason to care about the quality of *Wikipedia* is because it serves as an apparently successful example of a new form of production: open-source production (Kogut and Meitiu 2001; von Hippel 2002). Open source production essentially involves creating a public good, and therefore entails the same social dilemma that confronts the production and maintenance of other public goods. The intersection of two well-known mechanisms for overcoming social dilemmas, reputation and group identity, account for some of the variation in the quality of contributions to the open source encyclopedia, *Wikipedia*. Consistent with the expectations of the open source community and with previous studies of open source goods, we find that zealots and highly committed experts contribute high quality content. Yet, these mechanisms fail to

account for the very high quality content provided by anonymous Good Samaritans who do not care about reputation, and contribute only a few times.

The findings of lower quality for anonymous contributors with high participation indicate a strong but negative interest in the collective good which, if left unchecked, could destroy the open source good much as other commons can be destroyed by similar collective action problems. To deal with the negative impact of this group of contributors *Wikipedia* has instituted a policy that requires contributors to register after some number of anonymous contributions. Of course a policy of required registration is somewhat contrary to the ideal of open source and open access, and could potentially inhibit Good Samaritans. Since the majority of anonymous contributors make only one, high quality edit, such a policy may not be very problematic.

Our finding that anonymous Good Samaritans contribute high quality content to open source goods is both novel and unexpected. One reason the role of Good Samaritans may have been overlooked in other studies of collective goods is because we rarely have data for all contributions, large and small, over the entire production history of public goods. For example, studies of participation in social movements focus on the role of individual incentives, social networks and collective resources (e.g., McAdam1982, 1988; Opp et al 1995) that facilitate the contributions of highly committed participants. Alternatively, laboratory studies of collective goods necessarily create highly structured contexts that do not allow participation from actors outside of the study, such as potential Good Samaritan contributors who happen to pass by. However, it also may be because of the scope of open source production, which enables vast numbers of

contributors to participate, that Good Samaritan contributors can play such an important role in producing collective goods.

Sociologists have argued that social actors vary in both resources and levels of motivation to contribute to collective goods, so a critical mass of heterogeneous contributors is necessary to produce them (Marwell and Oliver 1993; Heckathorn 1992). While recognizing that production functions vary across types of collective goods (Marwell and Oliver 1993; Heckathorn 1992, 1996), open source production reduces the costs of contributing and expands the population of potential contributors so much that a critical mass is more likely to be reached early in the production process, and to be maintained throughout the ongoing production of open-source goods. In other words, open source production alters the *quantity* of producers, which in turn affects the *quality* of the production process itself. Our findings that one-time, anonymous Good Samaritans, as well as committed experts, contribute high quality content to *Wikipedia* suggest that open source production enables the exploitation of untapped productive resources that overcome barriers to efficient production of collective goods.

Notes

1. Clean air, bridges and ocean habitats are all examples of public goods. Economists define public goods, in contrast to private goods, as a type of good that is non-excludable and non-rival, and often also requires joint production. Non-excludable means that once the good is produced it is available to all, though 'all' may be restricted by geography (e.g., you have to be in the White Mountains to breathe the clean air) or other characteristics, such as citizenship. A non-rival good is one in which consumption of the good does not reduce its availability. Finally, many public goods must be collectively (jointly) produced either because the vastness of the resources required prevent one individual from producing it, or because the good itself requires the contributions of many actors (e.g., a group discussion). Public goods often are under-produced because individual and collective interests do not align9, 20 - 21 and because they lack a critical mass of potential contributors17.

2. Data are available on request from the authors, on the condition that it not be shared subsequently or used for commercial purposes (please send requests via email to: wikidatarequest@dartmouth.edu).

3. The nature of the sampling procedure inhibited us from extracting data from the significantly larger English-language Wikipedia. It is possible that our findings apply only to the French and Dutch language content, because of cultural differences or other unknown reasons. Future research on other language areas is necessary to verify the findings we report here.

4. Wikipedia uses a PHP port of Perl's Algorithm::Diff module 1.06, which uses the Longest Common Subsequence approach to computing string differences. PHP is an

open-source programming language used for developing applications, dynamic web content, and software.

References

Camp, Jean, Helen Nissenbaum, and Cathleen McGrath. 2002. "Trust: A collision of paradigms." *Lecture Notes in Computer Science* 2339:91-105.

Cheshire, Coye, and Karen S. Cook. 2004. "The Emergence of trust networks under uncertainty – Implications for Internet interactions." *Analyse & Kritik* 26: 220-240.

Ciffolilli, Andrea. 2003. "Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of Wikipedia." *First Monday*, 8 (12). Available from: http://firstmonday.org/issues/issue8_12/ciffolilli/.

Coleman, James. 1990. *Foundations of Social Theory*. Cambridge, MA: Belknap Press.

David, Shay and Pinch, Trevor John, 2005. "Six Degrees of Reputation: The use and abuse of online review and recommendation systems." Presented at the Economic Sociology and Technology Conference, September 23-24, 2005, Ithaca, NY. http://ssrn.com/abstract=857505

Dawes, Robyn. 1980. "Social Dilemmas." *Annual Review of Psychology* 31:169-193.

_____, Alphons J. C. van de Kragt, and John M. Orbell. 1990. "Cooperation for the Benefit of Us-Not Me, or My Conscience." Pp. 97-110 in *Beyond Self-Interest*. Edited by Jane J. Mansbridge. Chicago: University of Chicago Press.

Friedman, Eric, and Paul Resnick. 2001. "The Social Costs of Cheap Pseudonyms." *Journal of Economics and Management Strategy* 10(2):173-xx.

Ghosh, Rishab, and V. Ved Prakash, 2000. "The Orbiten free software survey. *First Monday* 5 (7). Available from: http://firstmonday.org/issues/issue5_7/ghosh/.

Hardin, Garret. 1968. "The Tragedy of the Commons." *Science* 162:243-48.

Heckathorn, Douglas D. 1992. " ". *Advances in Group Processes* 9:41-xx.

_____. 1996. "The Dynamics and Dilemmas of Collective Action." *American Sociological Review* 61:250-277.

Kogut, Bruce, and Anca Metiu. 2001. "Open-source software development and distributed innovation." *Oxford Review of Economic Policy* 17(2):248-64.

Kollock, Peter. 1998. "Social Dilemmas: The anatomy of cooperation." *Annual Review of Sociology* 24:183-214.

_____. 1999. "The Production of trust in online markets." *Advances in Group Processes* 16:99-123.

Lakhani, Karim, and Eric von Hippel, "How open source software works: "free" user-to-user assistance. *Research Policy* 32:923-43.

_____, and Robert G. Wolf. 2005. "Why Hackers do what they do: Understanding motivation and effort in free/open source software projects" Pp.3-21 in *Perspectives on free and open source software*, J. Feller, B. Fitzgerald, S. Hissam, K.R. Lakhani, Eds. Cambridge, MA: MIT Press.

Lerner, Josh, and Jean Tirole. 2002. "Some simple economics of open source." *Journal of Industrial Economics* L(2):197-234.

Lih, Andrew. 2004. "Wikipedia as participatory journalism: Reliable sources?" Paper presented at *5th International Symposium on Online Journalism,* University of Texas, Austin, April 16-17, 2004.

Marwell, Gerald, and Pamela Oliver. 1993. *The Critical Mass in Collective Action*. Cambridge, England: Cambridge University Press.

McAdam, Doug. 1982. *Political process and the development of black insurgency, 1930-1970*, Chicago: University of Chicago Press.

_____. 1986. "Recruitment to high-risk activism: The case of Freedom Summer." *American Journal of Sociology* 92(1):64-90.

Neff, Gina and David Stark. 2003. "Permanently Beta." Pp. 173-188 in *Society Online*, edited by Philip Howard and Steve Jones. Thousand Oaks, CA: Sage Publications.

Neumann, Peter. 2005. "Attaining robust open source software." Pp.123-6 in *Perspectives on free and open source software*, J. Feller, B. Fitzgerald, S. Hissam, K.R. Lakhani, Eds. Cambridge, MA: MIT Press.

O'Mahony, Siobhan. 2003. "Guarding the commons: how community managed software projects protect their work." *Research Policy* 32:1179-98.

Open Source Initiative. http://www.opensource.org/docs/definition_plain.php (9/2005).

Opp, Karl-Dieter, Peter Voss, and Christiane Gern. 1995. *The Origins of a Spontaneous Revolution: East Germany, 1989*. Ann Arbor: University of Michigan Press.

Orlowski, Andrew. 2005. "Wikipedia founder admits to serious quality problems." *The Register* October 18, 2005. Available at: http://www.theregister.co.uk/2005/10/18/wikipedia_quality_problem/

Ostrom, Elinor. 1990. *Governing the Commons*. Cambridge, UK: Cambridge University Press.

Raub, Werner, and J. Weesie. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96:626-654.

Raymond, Eric S. 2001. *The cathedral and the bazaar: Musings on Linux and open source by an accidental revolutionary*. Sebastopol, CA:O'Reilly.

Simon, Herbert. 1957. *Models of man.* New York:Wiley.

Stallman, Richard. 1999. Pp.53-70 in *Open sources: Voices from the open source revolution*, C. DiBona, S. Ockman, M. Stone, Eds. Sebastopol, CA:O'Reilly.

Stewart, Daniel. 2005. "Social status in an open-source community." *American Sociological Review* 70:823-42.

Terdiman, Daniel. 2005. "Wiki becomes a way of life." Wired News. March 8, 2005. Available at: http://www.wired.com/news/culture/o,1284,66814,00.html

Torvalds, Linus, and David Diamond. 2001. *Just for fun: The story of an accidental revolutionary*. New York:Harper Collins.

von Hippel, Eric. 2001. "Innovation in user communities: Learning in open source software." *Sloan Management Rev* 42:82-86.

von Krogh, Georg, Sebastian Spaeth, and Karim Lakhani. 2003. "Community, joining and specialization in open source software innovation: a case study." *Research Policy* 32:1217-41.

Wagstaff, Jeremy. 2004. "Wikipedia: It's Wicked." *Far Eastern Economic Review*, 167(7):38-39.

Wattenberg, Martin, and Fernanda Viegas. 2003. "History flow: results." Available at: http://researchweb.watson.ibm.com/history/results.html

Weber, Steven. 2004. *The success of open source*. Cambridge, MA: Harvard University Press.

Wellman, Barry, and Milena Gulia. 1999. Pp. 167-94 in *Communities in cyberspace*, M.A. Smith, P. Kollock, Eds. New York:Routledge.

Table 1. Population and Sample of Wikipedia Contributors by User Type and Language

Language	User Type		Total
	Registered	Anonymous	
French			
Population	5,690	48,211	53,901
Sample	1,763	1,729	3,492
Dutch			
Population	2,895	30,322	33,217
Sample	1,819	1,747	3,566
Total			
Population	8,585	78,533	87,118
Sample	3,582	3,476	7,058

Table 2. Means for *Wikipedia* Contributor Characteristics (unweighted)

	Total	French	Dutch
Number of Cases	7,058	3,566	3,492
Retention Rate	72.1 (29.0)	70.4 (29.6)	73.7 (28.4)
Number of Edits	9.4 (15.0)	9.0 (14.5)	9.7 (15.5)
Log Edits	1.3 (1.3)	1.2 (1.3)	1.2 (1.4)
Article Size	4,412 (5,886)	5,054 (6,869)	3,784 (4,647)
Log Article Size	7.8 (1.2)	7.9 (1.2)	7.7 (1.2)
Contribution Size	358 (1,545)	358 (1,089)	358 (1,889)
Log Contribution	4.8 (1.6)	5.7 (2.5)	5.7 (2.5)
Registered User	51%	51%	51%

Note: Standard deviations in parentheses.

Table 3. *Wikipedia* Contribution Characteristics by Type of User (unweighted)

	Registered User	Anonymous User	
Quality	70.3 (28.4)	74.0** (29.5)	F = 29.7** df = 1, 7,056
Log Edits	1.9** (1.4)	0.60 (.83)	F = 2,058.0** df = 1, 7,056
Log Contribution size	6.9** (2.3)	4.5 (2.1)	F = 1,955** df = 1, 7,056
Log Article Size	7.8 (1.1)	7.8 (1.3)	F = 0.89 df = 1, 7,056
French language	.49 (.50)	.50 (.50)	F = 0.19 df = 1, 7,056

Note: Standard deviations in parentheses.

** $p < .01$

Table 4. Content Retention Rates by Contributor Motivations

Level of Commitment	Interest in Reputation Strong: Registered Users	 Weak: Anonymous Users
Strong: 5+ contributions	1 73% (.23) [1,2] n=1,941	3 69% (.26) n=469
Weak: 1-4 contributions	2 67% (.36) n=1,641	4 75% (.30) [3,4,5] n=3,007

Note: standard deviations in parentheses.

[1] cell 1 > cell 2 ANOVA F = 47.8, p<.001

[2] cell 1 > cell 3 ANOVA F = 11.3, p<.001

[3] cell 4 > cell 3 ANOVA F = 14.4, p<.001

[4] cell 4 > cell 2 ANOVA F = 70.1, p<.001

[5] cell 4 > cell 1 ANOVA F = 3.59, p<.10

Table 5. OLS Unstandardized Coefficients of Quality of Contributions for Registered versus Anonymous Users (weighted)

	Registered Users	Anonymous Users
Constant	.39** (.04)	.54** (.04)
Log Article Size	.06** (.005)	.05** (.004)
Log Contribution Size	-.03** (.003)	-.03** (.003)
French Language	-.03** (.01)	-.05** (.01)
Log Edits	.02** (.003)	-.01+ (.006)
Adjusted R^2	.07	.08
Unweighted N	3,582	3,476

+ $p < .10$ * $p \leq .05$ ** $p \leq .01$

Note: Standard Error terms in parentheses.

Figure 1. Quality of *Wikipedia* Contributions by Number of Contributions for Registered and Anonymous Users.

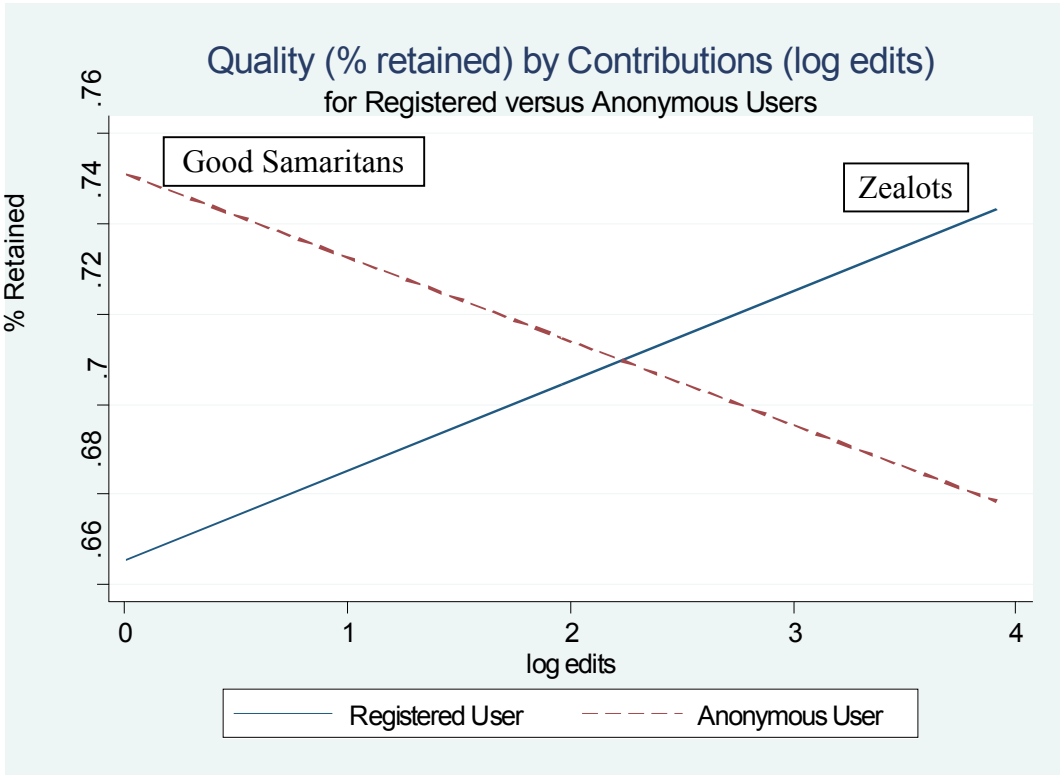

Table 6. OLS Unstandardized Coefficients of Quality of Contributions by Level of Contribution (weighted)

	Few Contributions < 5 edits	Many Contributions ≥ 5 edits
Constant	.54** (.03)	.39** (.05)
Log Article Size	.05** (.003)	.06** (.01)
Log Contribution Size	-.03** (.002)	-.03** (.004)
French Language	-.06** (.01)	-.014 (.01)
Registered User	-.05** (.02)	.05** (.01)
Adjusted R^2	.08	.06
Unweighted N	4,647	2,410

** $p \leq .01$ *Note*: Standard Error terms in parentheses.