Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of *Wikipedia*

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

November 2005

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

WORD COUNT: 6,159 (main text, notes, references)

TABLES: 7

FIGURES: 2

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

Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of *Wikipedia*

Abstract

One important innovation in information and communication technology developed over the past decade was organizational rather than merely technological. Open source production is remarkable because it converts a private commodity (typically software) into a public good. A number of studies examine the factors motivating contributions to open source production goods, but we argue it is important to understand the causes of *high quality* contributions to such goods. In this paper, we analyze quality in the open source online encyclopedia *Wikipedia*. We find that, for users who create an online persona through a registered user name, the quality of contributions increases as the number of contributions increase, consistent with the idea of experts motivated by reputation and committed to the *Wikipedia* community. Unexpectedly, however, we find the highest quality contributions come from the vast numbers of anonymous "Good Samaritans" who contribute infrequently. Our findings that Good Samaritans as well as committed "Zealots" contribute high quality content to *Wikipedia* suggest that open source production is remarkable as much for its organizational as its technological innovation that enables vast numbers of anonymous one-time contributors to create high quality, essentially public goods.

Word count: 188

Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of *Wikipedia*

I. Introduction

Of the significant advances in information and communication technology over the past decade, some important innovations were organizational rather than technological (Neff and Stark 2003; O'Mahony 2003). One of the most important organizational innovations is the emergence of open source production, which involves 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). Essentially, open source production is remarkable because it converts a private commodity (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 rather than a production process because it appears to give rise to the strong commitment and group identity often found in social movements (Raymond 2001; Stallman 1999; Torvalds and Diamond 2001).

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. The central questions for understanding open source production are who contributes to open source goods, and why?

Despite much hype that a distributed community of anonymous participants create high quality goods via open source production, 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. In open source production, these mechanisms not only motivate participation but are the basis for status in the community (Stewart 2005).

Though important, explanations of the motivations of the majority of open source contributors leave unanswered two additional questions. First, if not all content comes from committed experts, what motivates the one-time contributors? Second, and most important, are contributor motivations related to the *quality* of open source goods? Though public goods are chronically under produced, 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? This paper seeks to answer these questions and makes three contributions. First, we theorize the relation between contributor motivations in open source goods and the quality of contributions. Second, we use data from the online, open-source encyclopedia, *Wikipedia.org* to test hypotheses about contributor motivations and quality. Finally, our findings suggest that open source may be not only an efficient organizational innovation, but also a new site for exploring and identifying new forms of collective action processes.

II. The Case of Wikipedia

Wikipedia, the online, open-source 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, has the most content with over 830,000 articles (as of 11/2005). *Wikipedia* describes itself as "a Web-based, multi-language, free-content encyclopedia written collaboratively by volunteers and sponsored by the non-profit Wikimedia Foundation. It has editions in roughly 200 different languages (about 100 of which are active) and contains entries both on traditional encyclopedic topics and on almanac, gazetteer, and current events topics" (http://en.wikipedia.org/wiki/Wikipedia).

Not only is *Wikipedia* content open access, but the creation (and revision) of the content is also entirely open source 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 for a history of Wikipedia). In contrast, *Wikipedia* as it now exists succeeded by replacing professional contributions and expert peer review with their most democratic extremes: anyone can contribute, or edit any content with no proof of identity or qualifications.

The value of *Wikipedia* is the quality of its content, yet its overall quality is a much debated issue, even within *Wikipedia* (see e.g., http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_commentary/Wikipedia_quality; http://en.wikipedia.org/wiki/Criticism_of_Wikipedia). Popular accounts from both critics (e.g., Orlowski 2005) and fans (e.g., Terdiman 2005; Wagstaff 2004) raise questions about quality based on concerns about contributors, though it is important to point out that none of these accounts are based on systematic analysis of content, despite the vehemence of the arguments. In the only systematic study of the quality of content that we are aware of, *Wikipedia* is found to be comparable in quality to traditional encyclopedias (Lih 2004).

Given that the creation of its content is completely open, quality depends entirely on the types of contributors to *Wikipedia*. Yet, as noted by critics, why would any actor, let alone an expert, contribute? That is, from a completely rational perspective, *Wikipedia* is a collective good entailing at least some costs when contributing (e.g., time) and no expected individual gain, so no rational actor would be expected to contribute. Concerns about who contributes, and the possiblity that they are indeed *not* experts, is the key reason for critics' claims that *Wikipedia* content must be of low quality, for why, they ask, would any real expert participate?

In the lore of open-source communities there are two types of contributors, the strongly committed expert and the passerby contributor. Strongly committed experts are contributors who not only have the expertise to contribute high quality content, but also care a great deal about the collective good itself and are willing to contribute consistently to make sure it is high quality. Open Source production is fueled by these experts who contribute much of the content (Ghosh and Prakash 2000; Mockus et al 2005; Lakhani and von Hippel 2002) for a variety of reasons. In studies of various open source projects, one of the primary reasons for experts to contribute 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). Another important factor motivating contributors to open source is commitment to the open source community (Lakhani and von Hippel 2003; Lakhani and Wolf 2005; Lerner and Tirole 2002; Raymond 1999; von Krogh et al 2003; Wellman and Gulia 1999). Finally, still other factors motivating contributions to open source goods include the very low costs of contributing (Lerner and Tirole 2002), and possibly because contributors are "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).

The passerby contributor, in contrast, is a user who 'wanders in' to the website and, like a Good Samaritan, contributes to a topic, typically one time only. Participation by Good Samaritan contributors is enabled by the 'wiki' technology, which both expands the potential population of contributors, and reduces the costs of participation. In a 'wiki' every edit made is saved as a unique document. This means that any contributor 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, limit 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).

Thus, we have somewhat conflicting expectations about the types of contributors who provide content: committed experts with high levels of involvement in the community, and anonymous one-time contributors. What are the implications of these two types of contributors for understanding quality?

High quality contributions are expected from the skilled and committed experts known to contribute to other open source goods because they can benefit from building a reputation (e.g., Lakhani and von Hippel 2003; Lerner and Tirole 2002). Reputation systems are powerful mechanisms for overcoming collective action problems (Cheshire and Cook 2004; Kollock 1998; Raub and Weesie 1990), and are considered the basis for success of other new Internet-based institutions, such as the auction website eBay (Kollock 1999). Because reputation systems can facilitate trust as well as contributions to collective goods some researchers advocate such systems as the basis for all secure Internet-based communication and exchange (e.g., Camp et al 2002; Cheshire and Cook 2004).

Wikipedia recognizes the power of reputation by allowing interested contributors to become 'registered users.' Users register by creating a user name and providing an email address. *Wikipedia* actively encourages users to register in order to establish a reputation: "If you create an account, you can pick a username. Edits you make while logged in will be assigned to that name. That means you get full credit for your contributions in the page history…. While we welcome anonymous contributions, logging in lets you build trust and respect through a history of good edits" (http://en.wikipedia.org/wiki/Wikipedia:Why_create_an_account%3F). According to *Wikipedia*, there are "over 250,000 user accounts, along with an unknown (but quite large) number of unregistered contributors," (http://en.wikipedia.org/wiki/Wikipedia:Wikipedians July 2005). *Wikipedia* does not require proof of identity or qualifications to participate, but contributors can make contributions in two different ways – either anonymously or as a registered user.

While user names are still merely 'cheap' pseudonyms (Friedman and Resnick 1999), meaning they are easily abandoned and not necessarily tied to an individual's real name and identity, they provide a way to track a contributor's history. 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, and a recent article in the popular press highlighted some of these individuals by name (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 edit is contributed from a particular address. As shown below, however, the majority of anonymous users have only one edit.

A different type of incentive for contributors is the desire to be part of the *Wikipedia* community. *Wikipedia* clearly presents itself as a community. According to *Wikipedia*: "*Wikipedians* are the people who write and edit articles for *Wikipedia*…. The ending of *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). 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*. In this way *Wikipedia* may be similar to other open-source projects (Raymond 1999), and virtual communities (Wellman and Gulia 1999) in which many participants strongly identify as a member of the group, even though the group exists only in virtual 'online' space. Experimental research demonstrates that the salience of a group identity can motivate actors to contribute to collective goods, such as open source goods (Dawes 1980; Kramer and Brewer 1984; Dawes, van de Kragt and Orbell 1990; Turner and Tajfal 1986).

According to this discussion, contributors to *Wikipedia* are motivated by two different factors: (1) reputation and/or (2) commitment to the group identity of the *Wikipedia* community. Any contributor who has a strong interest in 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. That is, contributors who identify strongly with the community will participate a lot (many contributions) while contributors who do not identify with the community are likely to have low participation levels (few contributions).

TABLE 1 ABOUT HERE

What are the implications for quality when considering the intersection of these two sources of motivation? It is straightforward to consider contributors at the intersection of strong interest in reputation and a strong *Wikipedia* identity, i.e., registered users with many contributions (see cell one in Table 1). They are the committed-expert contributors and zealots expected by advocates of open-source online communities. The ability to identify and track the contributions of registered users, particularly over many contributions, makes them both interested in contributing a lot and suggests that their quality improves over time, else they would not be able to gain a positive reputation. This discussion suggests the following hypotheses:

Hypothesis 1a: Registered users will have higher participation levels than non-registered users.

Hypothesis 1b: The quality of contributions will increase with participation for registered users.

The expectation of increasing quality with participation for registered users implies that registered users with few contributions (cell 2 in Table 1) will be of lower quality. Why might this be the case? Three plausible explanations exist. First, a user may register prior to contributing, so be unsure about whether her contribution is high quality or not. Upon learning that it is low quality, she may abandon that registered name. Second, and related to the first, if there is a learning curve in making quality contributions such that registered users get better over time, a snapshot sample of registered users with few contributions may be early in their 'career' as contributors so be of lower quality, while registered users with many contributions will have improved over time. Finally, a user may be aware that his contribution is not high quality, but use a registered user name to try to fool others with a signal associated with high quality. For example, studies of Amazon.com reviews suggest that authors may engage in many questionable practices and often promote specific agendas while attempting to build their identities as experts (David and Pinch 2005).

Hypothesis 1c: Registered users with low participation levels will contribute lower quality content than strongly committed registered users (i.e., those with high participation).

What are the implications for quality for anonymous contributors? Virtually all theories of social dilemmas would predict low quality contributions from anonymous contributors, especially those with low levels of participation, since they would seem to have little motivation or incentives to contribute. Yet the lore of open-source suggests that anonymous one-time contributors (cell 4 in Table 1) 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 actually increase the costs of contributing for these Good Samaritan contributors, 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. Alternatively, the second type of Good Samaritan contributors are likely to be merely passers-by who see a mistake or a hole and make a contribution to address it. These contributions are likely to be much shorter than others, and therefore less likely to be edited or changed in the future, making them appear high quality. Thus, in contrast to registered users whose quality is highest at high levels of participation, anonymous users with the fewest contributions will be the highest quality.

But what are the quality implications for anonymous users as participation increases (cell 3 in Table 1)? As noted above, contributors with high participation levels strongly identify 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 is 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. Another possibility, however, is that such users know their contributions are of low quality and do not want to be identified through a registered user name. As their high levels of contribution suggest, these users are strongly committed to the *Wikipedia* community, but unlike the *Wikipedians* described above, their interest is *negative* rather than positive. These would-be "hackers" may actively seek to contribute low-quality content to harm the community. This discussion of the motivations for anonymous contributors, including both Good Samaritans and high participation-anonymous contributors, leads to the following hypotheses:

Hypothesis 2a: Anonymous users with few contributions will have high quality content.

Hypothesis 2b: The quality of contributions will decrease with participation for anonymous users.

Hypothesis 2c: Anonymous users with many contributions will contribute low quality content.

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

III. Data and Methods

We selected a sample of *Wikipedia* contributors from the population of both the French and Dutch language sites as of March 1, 2005.[2] At that time 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 2 for a breakdown of the sample by user type and language.) 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. Since registered users are over-represented in our sample compared to their distribution among all contributors, we weight the analyses below based on the representation of each user-type within each language group.

Variables

Our dependent variable is a measure of the quality of contributors' contributions, *not* the quality of *Wikipedia* content per se. That is, we are not measuring the quality of *Wikipedia* articles but of *Wikipedia* contributors. We measure the quality of contributions *quantitatively* as the percentage of a contributor's edit retained in the current version of the article. This measure of quality measures the *survivability* of a contribution and is only one dimension of a contributor's quality. It is likely a conservative measure to the extent that contributors are *satisficing* rather than maximizing (Simon 1957), that is, adding to and editing an entry until it is 'good enough' rather than until it is in some sense 'complete.'

For each contributor, we use the *Wikipedia* differencing algorithm[3] to compare the differences between three documents: (1) *edit*, the edit submitted by the contributor, (2) *previous*, the version of the article prior to the edit, and (3) *current*, the current version of the article as it exists on the day the sample was drawn. *Edits* generally occur in time prior to the current time point at which *current* is measured, so *current* does not in general equal *edit*, though it is possible. We measure the quality of an edit by calculating the number of characters from a contributor's *edit* that are *retained* in the *current* version, measured as the percentage retained of the total number of characters in the entry (retained in current/total in current). For example, compare the following 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 *percentage retained* 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/changed or deleted content, as well as content kept from the previous version. That means that our measure of *percentage retained* 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 measure of quality 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 quality (percentage retained) of contributors to Wikipedia is 72%. (See Table 3 for means of all variables.)

We recognize that this measure of quality does not take into account important features of *Wikipedia*, such as edit wars, articles under construction, etc. These issues are most important when evaluating the quality of content, 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 histories and retention rates.

The key independent variables are whether a contributor is registered or anonymous and the number of contributions. Contributor registration status is measured by whether they have a *registered user name* or not. Level of contribution 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), the total size of each topic article, measured as the total number of characters in the article (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 4 shows the bivariate results for each variable by user type. Anonymous and registered users differ in important ways. Overall, registered users contribute more content more often compared to anonymous users, consistent with Hypothesis 1a. Anonymous users, however, contribute higher quality content, a surprising finding given the expected motivations of reputation and identity, particularly for the zealots and committed experts.

Table 5 reproduces the intersection of contributor motivations shown in Table 1 to first examine the simple relationships identified in hypotheses 1c, 2a and 2c. We find support for all hypotheses. Both committed experts and Good Samaritans have high quality contributions. Supporting hypothesis 1c, committed experts' (cell 1) contributions are of significantly higher quality compared to registered users with fewer contributions (cell 2).

TABLES 4 AND 5 ABOUT HERE

Good Samaritans (cell 4 in Table 5) make the highest quality contributions overall. Good Samaritans contribute higher quality content than either registered users with similar levels of participation (cell 2) or other anonymous users who have higher levels of participation (cell 3), supporting hypotheses 2a and 2c.

The bivariate results shown in Table 5 also suggest support for hypotheses 1b and 2b about the divergent relationship in quality for different types of contributors across levels of participation. Figure 1 displays the estimated regression lines for the quality of contributions (% retained) regressed on participation (log edits) for both registered and anonymous users and shows that indeed quality changes with the amount of participation but in exactly the opposite directions for registered versus anonymous users. Anonymous users' quality is very high at low levels of participation, but decreases as participation increases, while the opposite is true for registered users for whom quality increases with participation. Also note that anonymous users with low participation (*i.e.*, Good Samaritans) have the highest quality overall.

TABLES 6 AND 7 ABOUT HERE

We now turn to the multivariate analysis. Table 6 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 6. 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 7 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. Figure 2 illustrates the mean quality, adjusted for article size, language and contribution size, for different types of users with different levels of contribution. Anonymous and registered users are significantly different from one another within each contribution category ($p<.01$), and quality in the 5+ edits category is significantly higher than in the 1-4 edits category ($p<.05$).

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-July 30, 2005 yielded 29 articles. See also http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_the_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 has had a 50% increase in visitors over the past 3 months. 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 feelings of identity, the main interest of readers is simply the *quality* of the contributions. *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 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 a successful example of an apparently 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.

Our finding that anonymous Good Samaritans contribute high quality content to open source goods is both novel and unexpected by social science theory. 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 that it is only via open source production that Good Samaritan contributors can play such an important role in producing collective goods.

Is there something different about open-source production that motivates these one-time anonymous contributors? 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; *cf.* 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. 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).

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

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. Contributor motivations, user type, level of participation and quality of contribution

Level of Identity	Interest in Reputation: Strong	Interest in Reputation: Weak
Strong	1 Registered Users Many contributions ~Zealots & Committed Experts~ High Quality	3 Anonymous Users Many contributions Low Quality?
Weak	2 Registered Users Few contributions Low Quality?	4 Anonymous Users Few contributions ~Good Samaritans~ High Quality?

Table 2. 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 3. Means for *Wikipedia* Contributor Characteristics (unweighted)

	Total	French	Dutch
Number of Cases	7,058	3,566	3,492
Quality (% retained)	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 4. *Wikipedia* Contribution Characteristics by Type of User (unweighted)

	Registered User	Anonymous User
Quality	70.3 (28.4)	74.0** (29.5)
Log Edits	1.9** (1.4)	0.60 (.83)
Log Contribution size	5.0** (1.5)	3.9 (1.7)
Log Article Size	7.8 (1.1)	7.8 (1.3)
French language	.49 (.50)	.50 (.50)

** significantly higher mean ($p < .01$) *Note:* Standard deviations in parentheses.

Table 5. Quality of Contribution by Contributor Motivations

Level of Identity	Interest in Reputation: Strong: Registered Users (RU)	Interest in Reputation: Weak: Anonymous Users (AU)
Strong: High level of participation 5+ contributions	1 ~Zealots & Committed Experts~ 73% (.23) [1,2] (n=1941)	3 69% (.26) (n=469)
Weak: Low level of participation 1-4 contributions	2 67% (.36) (n=1641)	4 ~Good Samaritans~ 75% (.30) [3,4] (n=3007)

[1] = RU with 5+ edits significantly greater than RU with 1-4 edits ($F=47.8$, $p<.001$) cell 1 > cell 2
[2] = RU with 5+ edits significantly greater than AU with 5+ edits ($F=11.3$, $p<.001$) cell 1 > cell 3
[3] = AU with 1-4 edits significantly greater than AU with 5+ edits ($F=14.4$, $p<.001$) cell 4 > cell 3
[4] = AU with 1-4 edits significantly greater than RU with 1-4 edits ($F=70.1$, $p<.001$) cell 4 > cell 2

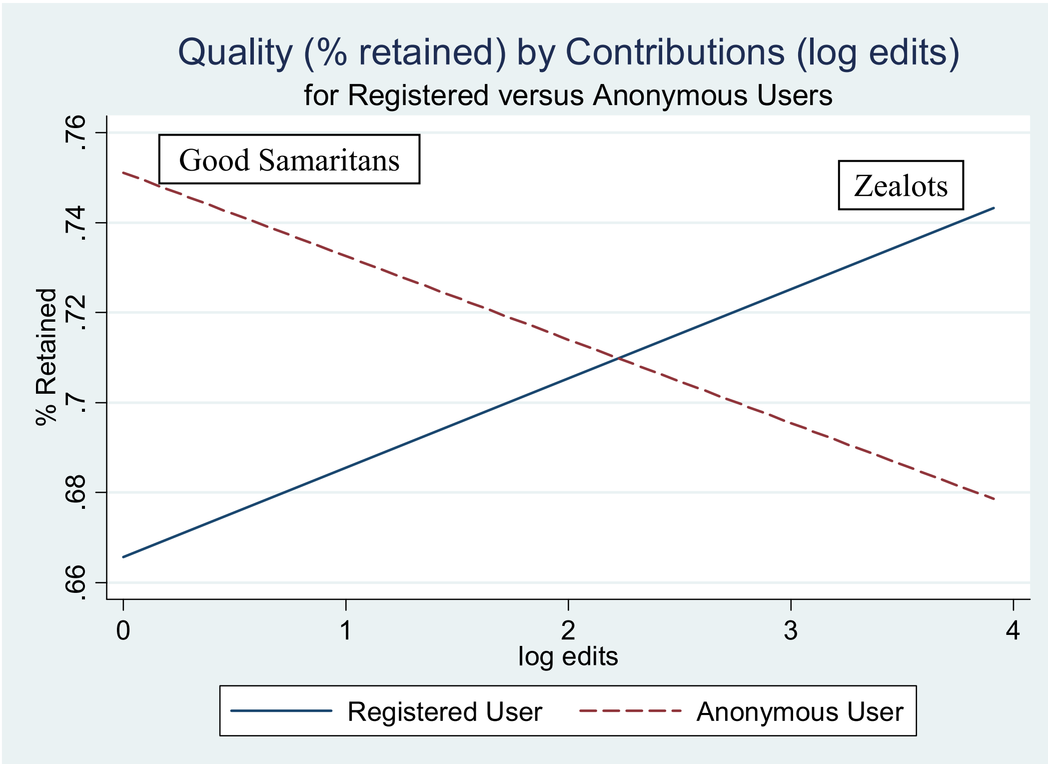

Figure 1. Quality of *Wikipedia* Contributions by Number of Contributions for Registered and Anonymous Users.

Table 6. OLS Unstandardized Coefficients of Quality of Contributions for Registered versus Anonymous Users (weighted)

	Registered Users	Anonymous Users
Constant	.39** (.04)	.54** (.03)
Log Article Size	.06** (.004)	.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

* $p \leq .05$ ** $p \leq .01$ *Note*: Standard Error terms in parentheses.

Table 7. 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 R2	.08	.06
Unweighted N	4,647	2,410

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

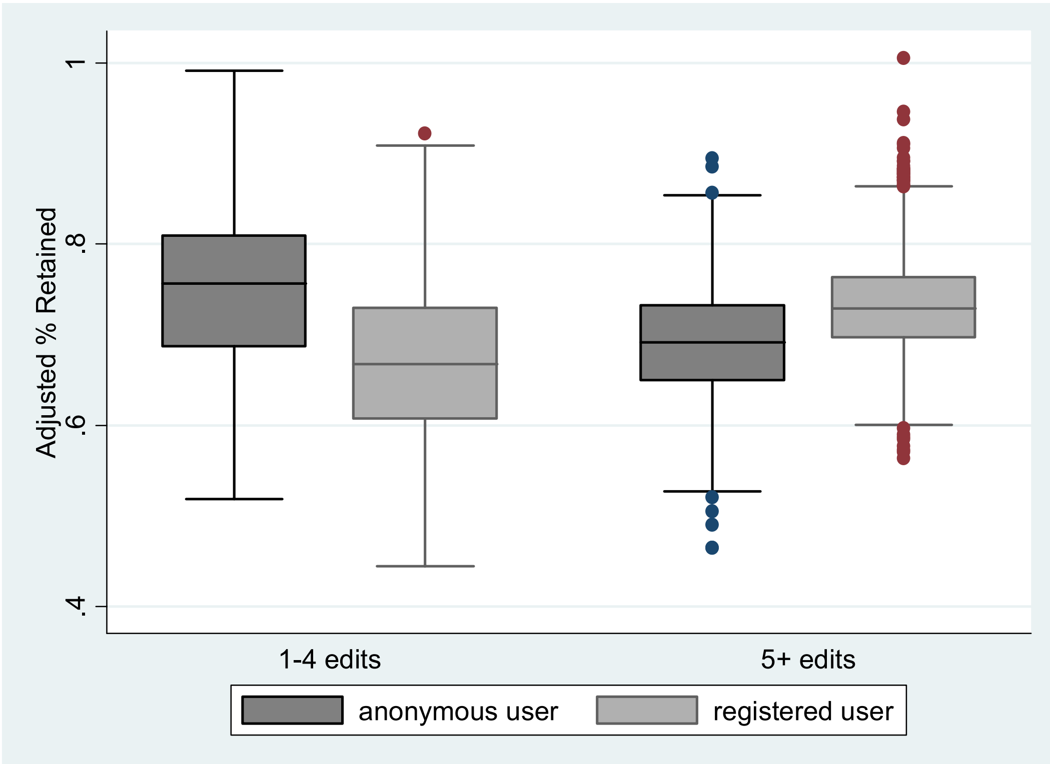

Figure 2. Quality by Number of Contributions for Anonymous and Registered Users.

Notes: Mean contribution quality, adjusted for article size, language and contribution size. Anonymous and registered users are significantly different from one another within each contribution category ($p<.01$), and quality in the 5+ edits category is significantly higher than in the 1-4 edits category ($p<.05$).