Building Quality in Wikipedia: A Theoretical Approach

Guangyuan Zhu*

Dr. Huseyin Yildirim, Faculty Advisor

April 15, 2009

Duke University

Durham, North Carolina

*This honors thesis is submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University. The author can be reached at victor.zhu@duke.edu.

Table of Contents

1 Abstract

Wikipedia is known as a convenient source of user generated information for a wide range of topics, but is it able to compete in quality with a print encyclopedia? In this paper, I formulate a theoretical framework to investigate this topic. I begin with a two-step sequential case for both Wikipedia and print encyclopedia, in which I find that the quality of Wikipedia is generally slightly lower than that of a print encyclopedia. In an extension to an n-person case, I find that the quality of Wikipedia can rise higher than that of print encyclopedia. After taking into account the generally lower costs of contribution for Wikipedia, this small difference will be eliminated.

2 Acknowledgments

I would like to thank my faculty advisor, Professor Huseyin Yildirim, for providing indispensable advice throughout the development of this thesis, as well as Professor David Banks for his help on crucial statistical problems within the formulation of the model. I would also like to thank my classmates for supplying insightful feedback throughout the entire process.

3 Introduction

Many people know Wikipedia to be a convenient resource from which to gather information on topics ranging from the history of Turkey to the latest results of American Idol. What some people may not know is that the information provided in the entries is quite trustworthy. According to a study by Nature Magazine, in a review of 42 articles on both Wikipedia and Encyclopedia Britannica, the former averaged 4 errors and the latter averaged 3 errors, while both had the same number of "serious" errors (Wired 2005 [2]). So in fact, Encyclopedia Britannica has also had its share of publicity regarding erroneous information.

What makes the information in Wikipedia so trustworthy? At first glance, one might think that the reverse would be the case. After all, "most of the articles can be edited by anyone with access to the Internet," although "other editors are always around to advise or correct obvious errors" (Wikipedia 2008 [10]). Wikipedia also acknowledges the fact that "older articles tend to be more comprehensive and balanced, while newer articles more frequently contain significant misinformation, unencyclopedic content, or vandalism" (Wikipedia 2008 [10]). In light of this fact, one may conclude that the quality of Wikipedia articles would generally be lower than that of a traditional print encyclopedia. However, the evidence suggests otherwise. What might explain the existence of a high-quality online encyclopedia written by regular people around the world?

Previous scholarly work on Wikipedia has been focused on the sociological aspects of contribution. The nature of these studies has mainly been anecdotal and statistical. Contributors are generally defined in two groups: novices that write "what they know" on topics of their interest, and experts that have "a concern for the quality of the Wikipedia itself" (Bryant 2005 [4]). Intrinsic motivation, such as "a sense of

relatedness" causes contribution (Zhang 2006 [12]). Wikipedia's success is also related to the dramatic decrease in transaction costs for online collaboration and the creation of an "artificial information economy as a context for collaboration" (Neus 2001 [6]). According to Neus, there is little incentive for vandalism and low-quality work because "it is much 'cheaper' for person B to undo the low-quality change that person A caused, than it is for person A to cause it." It is just a matter of a few clicks of a mouse.

Although this literature can help us understand the rise of Wikipedia, it cannot give us a theoretical explanation for its quality or offer a comparison between the trustworthiness of Wikipedia and a print encyclopedia. In this paper, I present a theoretical framework to explain this phenomenon. I focus on the fact that "Wikipedia is continually updated, with the creation or updating of articles on topical events within seconds, minutes or hours, rather than months or years for printed encyclopedias" (Wikipedia 2008 [10]). I discuss a model in which it is viewed as a public good by its contributors. In my framework, I focus on the "novices," who make up the majority of the people who contribute. Although they are not as devoted as experts, I assume they at least care about the quality of the topic of their interest and have something to bring to the table. To simplify, I ignore acts of vandalism based on the conclusion from Neus 2001 ([6]) that the incentive is lacking. I also investigate theoretically the extent to which the increase in the number of contributors influences average quality.

I begin with the two-person Wikipedia case, in which each person has private knowledge about the topic, has the same cost of contribution, and decides whether or not to contribute sequentially. I then compare the average total quality from the Wikipedia case with that produced by a writer of a regular encyclopedia. Because this writer is employed, I assume he or she must contribute. However, the writer also has the option to do additional research to increase the quality of the work. After analyzing the two-step cases, I extend the model to a generalized infinite-person case

for Wikipedia, and infinite-shot case for print encyclopedia.

4 Literature Review

There exists scant literature on the theoretical nature of contributions to Wikipedia and its relationship to quality. The phenomenon of user-generated content with instant online collaboration is quite recent, such that most research is interested in the issue of motivation. Generally, empirical findings have indicated that high-quality articles on Wikipedia have a larger number of edits than lower-quality articles (Wilkinson and Huberman 2007 [11]).

In one recent empirical paper, data regarding a sample of users and their respective edits is analyzed for frequency and quality. It finds that "the highest quality contributions come from the vast numbers of anonymous 'Good Samaritans' who contribute infrequently" (Anthony et al., 2005 [3]). These anonymous persons who rarely contribute produce high quality posts because they do so only out of interest in the topic. In contrast, highly active registered users tend to contribute high-quality posts because they are "true believers in a collective good." In another paper, Polborn 2007 ([7]) presents Wikipedia as a "club" in which the members obtain utility through the inflow of new members or users. Thus, the contributions exist to inform and influence others. These findings support the formulation of Wikipedia as a public good, from which the contributors benefit as it grows. However, they do not come close to hinting at the theoretical level of quality compared to that of a print encyclopedia.

In regards to general theory on public goods, there exists very little literature on best-shot aggregation of public goods. This is the approach that I will be taking first. Hirshleifer 1983 ([5]) first framed the best-shot rule, where "the socially available amount is the maximum of individual quantities." He formulates the best-shot social

composition function as $X = \max_i(x_i)$ and analyzes the best-shot case with an example involving suppliers, in which the one with the lowest Total Cost produces an efficient output such that his Marginal Cost equals the sum of all Marginal Rates of Substitution. He goes on to explain that actual provision will be much lower, since even the most efficient sole producer of public good will produce only so much that his Marginal Cost equals his individual Marginal Rate of Substitution. In my framework, the quality of Wikipedia can be conceived as the highest of sequential contributions, a best-shot case. However, I will assume that the nature of contribution depends only on a fixed cost for all suppliers, beliefs about how much others will contribute, and the amount (quality) supplied by the previous mover.

I initially construct contribution to Wikipedia as a sequential game where each agent does not know the wealth (quality) of the other agent. Varian 1994 ([9]) discusses such a framework in which each contributor has incomplete information about other contributors and the second contributor makes his optimal choice given the first agent's contribution. The first contributor has a prior distribution on how much the second will contribute. This game results in the first agent contributing less to the public good since he is uncertain about the type of the second agent; he does not want to crowd out some of the public good that the second agent would have contributed. This will also be manifested in my paper with some interesting results. However, I will show that even with such free-riding, the average overall contribution will still be high.

Employing the discussions and results from existing literature, this paper constructs a theoretical model of contributions to Wikipedia. My first hypothesis is that in the two-person case, the quality of Wikipedia will be comparable to that of print encyclopedia. My second hypothesis is that as the number of contributors increases, the quality of Wikipedia will increase.

5 Theoretical Framework

5.1 Wikipedia as a 2-person sequential game

Suppose there are n individuals who are interested in a topic. Each individual has some knowledge $q_i \in [0, 1]$ about this topic. Thus, we assume that q_i's are independent draws from a uniform distribution over $[0, 1]$. Each person i privately observes q_i as in Varian's model. Each person also has a cost of contribution $c \in [0, 1]$ for all individuals. Let Q_k be the aggregate quality after k individuals contribute. In our best-shot formulation, $Q_k = \max\{q_i, ..., q_k\}$.

These individuals go on Wikipedia and see an article of their interest. They derive utility from the overall information quality, the final Q. The rationale for this is that contributors like to learn more about the topic of their interest, and contributors like others to learn as much as possible about the topic of their interest, as discussed in previous literature. These individuals make the decision whether or not to contribute (no research is done as the players produce only their innate quality). If they do decide to contribute, they are made sequentially, as it is in reality. Each person reads and then posts if he or she has better knowledge. We assume that reading is costless. Thus, each person's utility is $U_i = Q_k - c$.

Consider the two-person case in which the first individual I_1 decides to contribute, then I_2. Because this is a sequential game, we will use backwards induction and start with I_2. After reading the post by I_1, I_2 decides to contribute if and only if $q_2 > c$ and $q_2 - c > q_1$. This is since before contributing, his utility is q_1, and after contributing, $Q = q_2$, and his utility is $Q - c$ (Here we assumed that $q_1 \leq 1 - c$ because if otherwise, I_2 will never contribute since $q_2 > 1$, thus violating our initial formulation). Then I_1, knowing I_2's reaction and that $q_2 \in [0, 1]$, decides whether or not to contribute. We at least know that I_1 will not contribute if $q_1 < c$.

Given what we know about I_2's strategy, we will calculate I_1's strategy and $\overline{Q}_1$, the expected quality after I_1 moves. After we know this, we will then apply I_2's strategy and calculate $\overline{Q}_2$, the expected quality after I_2 moves.

5.2 Encyclopedia as a 2-shot sequential game

Suppose a writer is employed to write for a print encyclopedia. In order to write an article, he or she incurs a cost $r_1 \in [0, 1]$ in order to draw a quality $q_w \in [0, 1]$, which can be interpreted as doing initial research to gather information. If the writer finds that q_w is not high enough, he or she may incur another cost $r_2 \in [0, 1]$ to draw another quality $q_x \in [0, 1]$, which can be interpreted as doing additional research, which may or may not gather better resources, to attempt to write a better article. Since the writer is employed, he or she cares about the quality of his or her work. Finally, the writer incurs a cost $c \in [0, 1]$ as the cost of writing the article.

Here, r_1 is trivial, since the writer must incur this cost, whatever it is, to research and write the initial article. This cost can be interpreted as the writer's regular daily effort used to do his or her job. Hence, we assume that this cost is not associated with the writer's motivation to write initially. However, r_2 is important, since the writer has a choice here whether or not to incur this additional cost. To simplify, we will merge r_2 and c into the overall cost of contribution $s \in [0, 1]$ (this assumes that $r_2 + c \leq 1$). We will formulate the writer's utility function as $U_w = Q_k - s$ where Q_k is the quality after all research and writing is done. The writer will only incur additional cost if the expected utility he or she receives from the final article after doing additional research outweighs just writing the initial article.

Given what we know about the writer's preferences, we will calculate his or her strategy and $\overline{Q}_w$, the overall expected quality.

5.3 Wikipedia as a n-person simultaneous game

Ultimately, we are interested in the n person case, since the premise of Wikipedia involves the collaboration of potentially millions of people. Thus, we are interested in the case as $n \to \infty$. If we attempt a sequential model as before, we run into significant hindrances, as with each additional stage the number of calculations explodes. Thus, we formulate the n-person case as a simultaneous game in order to simplify the model.

The setup is similar to the sequential case, where there are n individuals who are interested in a topic, each individual has some private knowledge $q_i \in [0, 1]$ about this topic, and each person has a cost of contribution $c \in [0, 1]$. We introduce a cutoff value $q*$, where individuals only contribute if their personal q_i is larger than $q*$

Let x_i be the draws of quality from a uniform distribution between 0 and 1. Let

$$\hat{x}_i = \begin{cases} x_i \text{ if } x_i \geq q* \\ 0 \text{ if } x_i < q* \end{cases} \tag{1}$$

Let $Y_n = \max\{\hat{x}_1, \hat{x}_2, ..., \hat{x}_n\}$, and let $Q_n = \max\{Y_{n-1}, q_i\}$. The aggregate quality is the maximum of the qualities of all contributions, including my own, if I contributed. Thus, each person's utility is $U_i^1(q_i) = Q_n - c$ if he or she contributes. Each person's utility is $U_i^0 = Y_{n-1}$ if he or she does not contribute. This makes sense, since if the person contributes, his or her utility is the resulting aggregate quality minus the contribution cost, and if the person does not contribute, his or her utility is just the aggregate quality from all other contributions. To find the cut-off $q*$, we set $E[\max\{Y_{n-1}, q*\}] - c = E[Y_{n-1}]$. If the person's quality is equal to the cut-off, then he or she should be indifferent between contributing and not contributing. Finally, we will find $E[Y_n]$, the expected aggregate quality with n persons, using the value found for the cut-off point.

5.4 Encyclopedia as a n-shot simultaneous game

The motivation for an n-shot game for print encyclopedia is not as strong as for Wikipedia, since it is unlikely that writers of encyclopedias do research more than a few times. However, for comparison purposes, we will find the optimal number of times that these writers should do research, given their s, cost of research and writing. We will compare the expected encyclopedia quality that arises from their research with the expected Wikipedia quality that arises from contributions by an infinite population.

Again, the basic assumptions are the same as in the sequential case, in which the writer draws his or her quality from a uniform distribution between 0 and 1 and has the opportunity do further research and writing by incurring a cost s for each revision. Let w_i be draws of quality from a uniform distribution between 0 and 1. Let $Y_n = \max\{w_1, w_2, ..., w_n\}$. The researcher's utility is $U_r = Y_n - ns$.

6 Analysis

6.1 Wikipedia as a 2-person sequential game

6.1.1 Result 1

A: The average quality after the first contributor moves is:

$$\overline{Q}_1 = \begin{cases} \frac{1+q_1^2-c^2}{2} & \text{if } q_1 \leq 1-c \\ q_1 & \text{if } q_1 > 1-c \end{cases} \tag{2}$$

See Appendix for proof of this result.

This make sense, since if $q_1 \leq 1-c$, I_1's decision to contribute is dependent on the cost, but if $q_1 > 1-c$, I_1 knows that I_2 will not contribute if I_1 contributes, so

the expected quality depends only on q_1.

B: The nature of contribution is:

$$\text{For } c < 2 - \sqrt{3} : \begin{cases} \text{contribute if } q_1 > \sqrt{2c} \\ \text{don't if } q_1 \leq \sqrt{2c} \end{cases} \tag{3}$$

$$\text{For } 1 \geq c \geq 2 - \sqrt{3} : \begin{cases} \text{contribute if } q_1 > \frac{1+2c-c^2}{2} \\ \text{don't if } q_1 \leq \frac{1+2c-c^2}{2} \end{cases} \tag{4}$$

See Appendix for proof of this result.

See Figure 1.

As one can see in Figure 1, I_1 is more likely to contribute for areas of low c than in areas of high c. This is very intuitive, since an increase in the cost of posting allows only those with higher q_1's to obtain positive utility from posting. One can also see that the rate of increase of the q_1 level necessary to contribute is high in the beginning. This is reasonable, since at low c, I_1 expects I_2 to also contribute, so I_1 has more incentive to free-ride off the contributions of I_2.

Observe that there is a kink at where $c = 2 - \sqrt{3}$ and $q = -1 + \sqrt{3}$. This is the rightmost point at which for the given c and where I_1 contributes, $q_1 \leq 1 - c$. As $c > 2 - \sqrt{3}$, the rate at which the quality level must increase in order for I_1 to contribute abruptly decreases due to the kink. This can be interpreted in the following sense: once the q_1 must be greater than $1 - c$ for I_1 to contribute, I_1 knows that I_2 will not contribute if I_1 contributes. Also, it will be less likely for I_2 to draw such a high quality if I_1 is able to contribute, so for each given c, I_1 has more incentive to contribute to the public good.

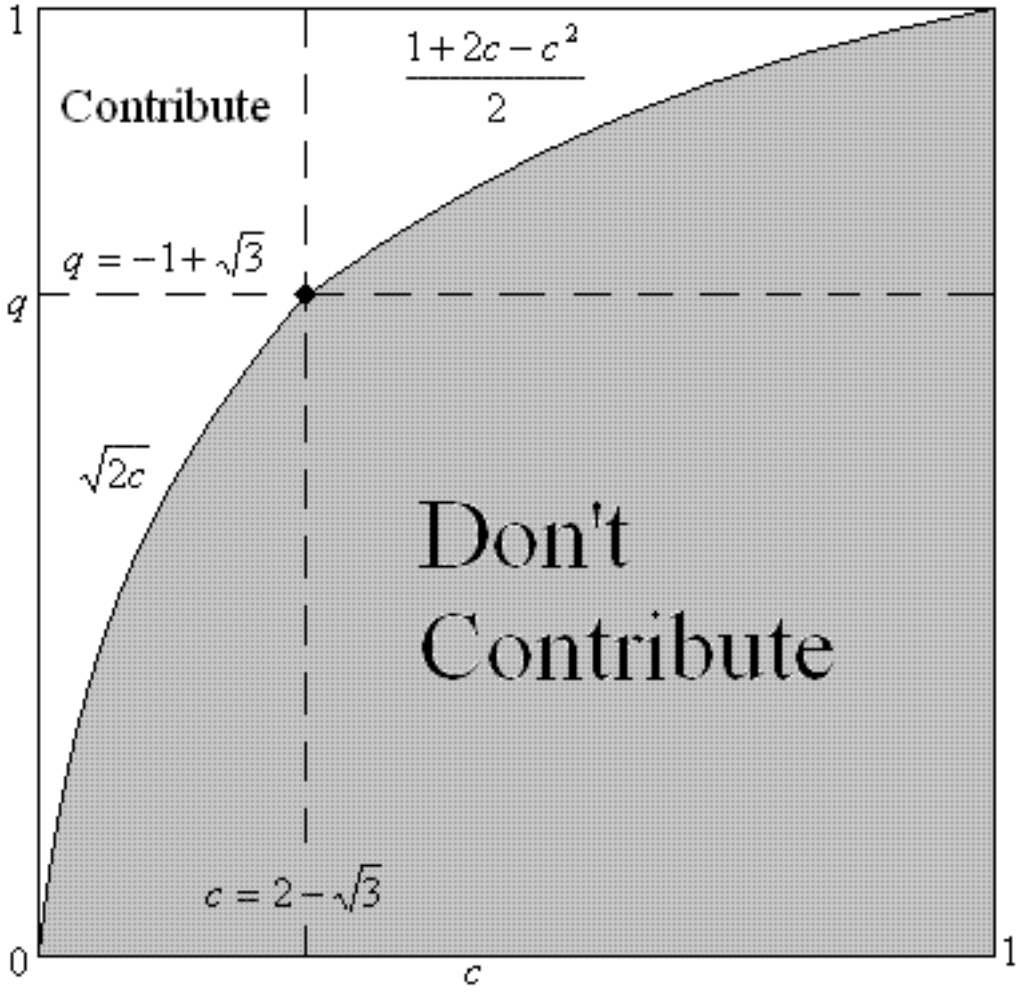
1
Contribute
$\frac{1+2c-c^2}{2}$
$q = -1+\sqrt{3}$
q
$\sqrt{2c}$
Don't
Contribute
$c = 2-\sqrt{3}$
0
c
1

Figure 1:

6.1.2 Result 2

$$\overline{Q}_2 = \begin{cases} \frac{4-2c\sqrt{2c}-3c^2+2c^3}{6} & \text{for } c < 2-\sqrt{3} \\ \frac{(c^2-1)(c^2-2c-1)}{4} - \frac{(c-1)^2(c^2-2c-3)}{8} & \text{for } c \geq 2-\sqrt{3} \end{cases} \tag{5}$$

See Appendix for proof of this result.

See Figure 2.

Figure 2 is strictly decreasing for $c \in (0, 1]$, meaning the slope is 0 at $c = 0$. This implies that at extremely low costs of contribution, both parties are likely to contribute, leading to a slow decrease in expected quality level as c increases. The general shape of the graph is rather reasonable, since as cost increases, contributions are likely to decrease, leading to a lower expected overall quality level.

Also observe the kink at $c = 2 - \sqrt{3}$. In relation to the case discussed in the first result, there is less free-riding as $c > 2 - \sqrt{3}$, since I_1 has more incentive to contribute. Thus, expected overall quality decreases at a slower pace.

6.2 Encyclopedia as a 2-shot sequential game

6.2.1 Result 1

$$\text{For: } \begin{cases} q_w < 1 - \sqrt{2s} & \text{writer researches further} \\ q_w \geq 1 - \sqrt{2s} & \text{writer stops researching} \end{cases} \tag{6}$$

See Appendix for proof of this result.

This result is intuitive, since as s, the cost of more researching, increases, the writer has less incentive to do so.

6.2.2 Result 2

$$\overline{Q}_w = \begin{cases} \frac{2-s\sqrt{2s}}{3} & \text{for } s < \frac{1}{2} \\ \frac{1}{2} & \text{for } s \geq \frac{1}{2} \end{cases} \tag{7}$$

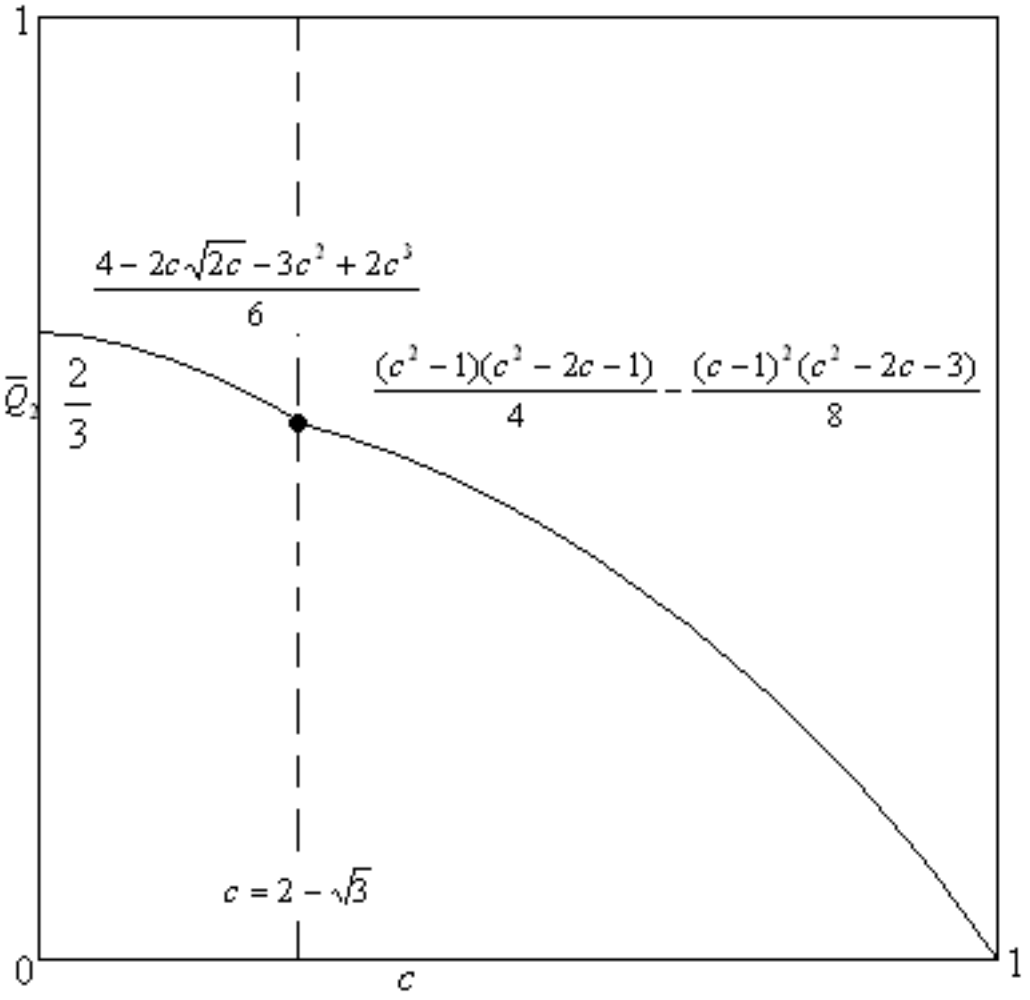
1
$\frac{4-2c\sqrt{2c}-3c^2+2c^3}{6}$
$\bar{Q}_2$
$\frac{2}{3}$
$\frac{(c^2-1)(c^2-2c-1)}{4}-\frac{(c-1)^2(c^2-2c-3)}{8}$
$c=2-\sqrt{3}$
0
c
1

Figure 2:

See Appendix for proof of this result.

See Figure 3.

As s increases, the expected quality decreases, which is intuitive. However, when $s > 1/2$, the function $1 - \sqrt{2s}$ becomes negative. Also at that point, the average quality of having only the first attempt, $\overline{q}_w = 1/2$ becomes greater than the average quality of having both attempts $\overline{q}_x$. Thus, for $s > 1/2$, the expected average quality will just be $1/2$, as the writer will no longer choose to do further research.

6.3 Comparison between sequential games

See Figure 4.

When compared to the quality of a print encyclopedia, the two-person Wikipedia case produces a strictly lower quality at every point except where c or s equals 0. We place both on the same axes because they have the same interpretation: a cost associated with contribution.

Although there is a quality difference, at least for the two-person case, this difference is quite low for low enough $c < 1/2$. Here, we are comparing quality where c equals s. However, it is much easier to contribute to Wikipedia than to a print encyclopedia. Anyone can just go online and, in a few minutes, type out a submission using their existing knowledge. With a print encyclopedia, there exists a research and review process that may take months or even years. The review process for Wikipedia is just more easy submissions correcting mistakes or adding information. Thus, there is reason to believe that c is generally less than s by a significant amount. If this were the case, then the quality difference may be a lot smaller, or even reversed. This is especially likely since we assumed earlier that $s = r_2 + c$, so s and c can only be equal if $r_2 = 0$, or there is no cost of researching. Looking at quality equal to $1/2$, where for $c \leq 1/2$ or $s \leq 1/2$ the difference between Wikipedia and print encyclopedia is largest, the

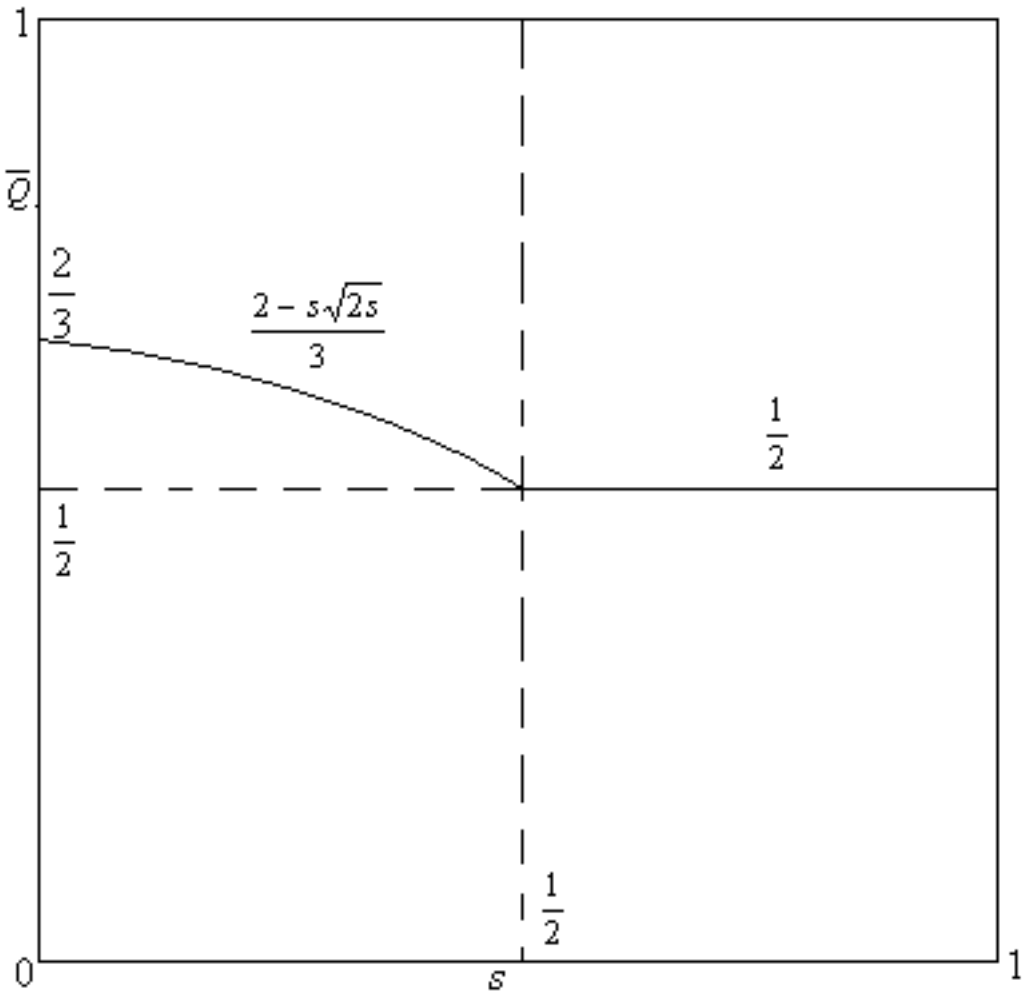

Figure 3:

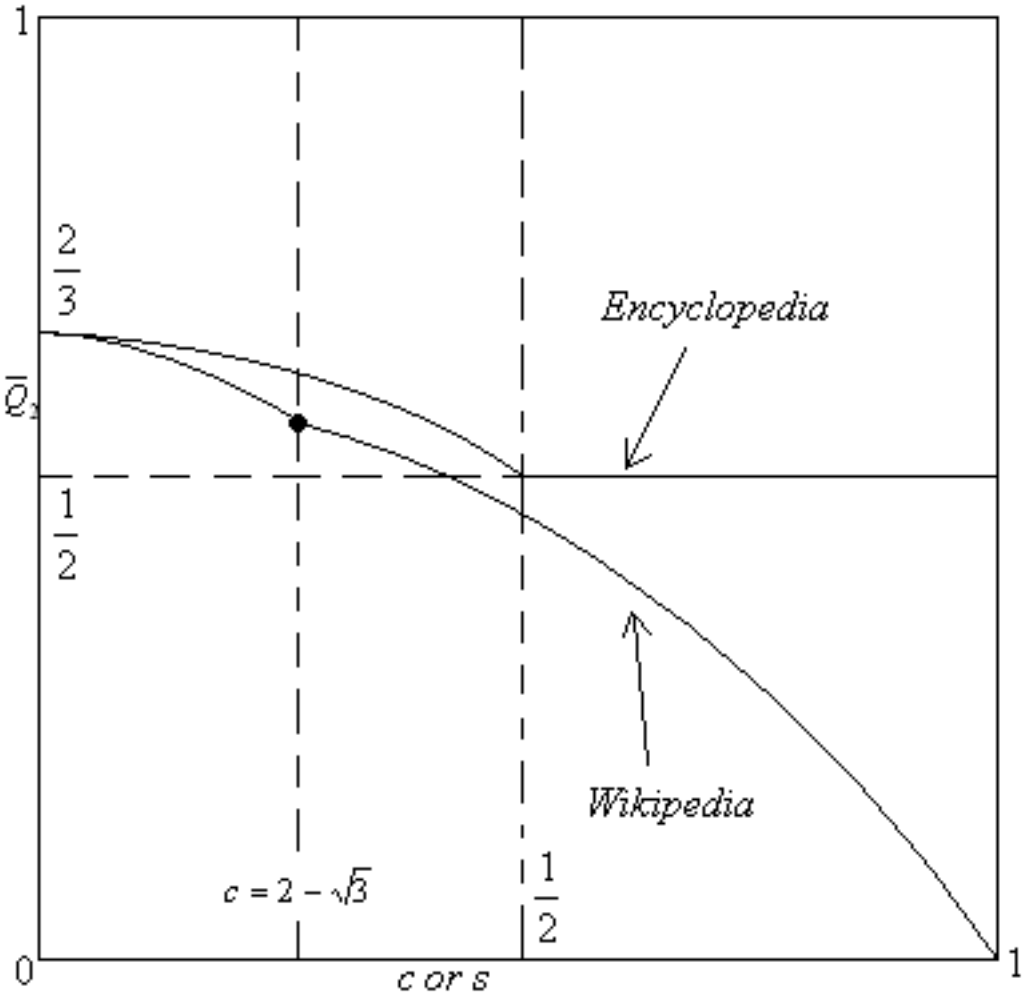

Figure 4:

difference there between c and s is only about 0.086. Thus, if c is generally less than s by 0.086, then Wikipedia quality would at least match that of print encyclopedia for $s \leq 1/2$. In reality, the effort put into each revision of a print encyclopedia article is likely many times greater than the effort put in by any single writer editing a Wikipedia article. The cost of researching for a print encyclopedia article is also likely to be as or even more time-consuming than writing it.

In addition, there is reason to believe that with more contributors, the quality of Wikipedia will increase. We will investigate this later in the paper.

Furthermore, this paper does not touch on reader's strategies. Intuitively, the reader may choose to sacrifice quality for ease of access. Since Wikipedia is free and online, the cost of using Wikipedia is much lower than that of a print encyclopedia, which one must buy or borrow. This implies that if Wikipedia has only a small disadvantage in quality, if any, readers should choose to use Wikipedia rather than the print encyclopedia.

6.4 Wikipedia as a n-person simultaneous game

6.4.1 Result 1

$$q* = c^{\frac{1}{n}} \tag{8}$$

See Appendix for proof of this result.

See Figure 5.

Thus, if $q_i > c^{\frac{1}{n}}$, then contribute; otherwise, don't contribute. This makes intuitive sense, since as cost increases, the threshold increases. People need to have higher quality to overcome the cost and contribute. As n increases, the cut-off increases, indicating that the free-rider problem still persists in this case. With more people possibly contributing, each individual person has less incentive to contribute.

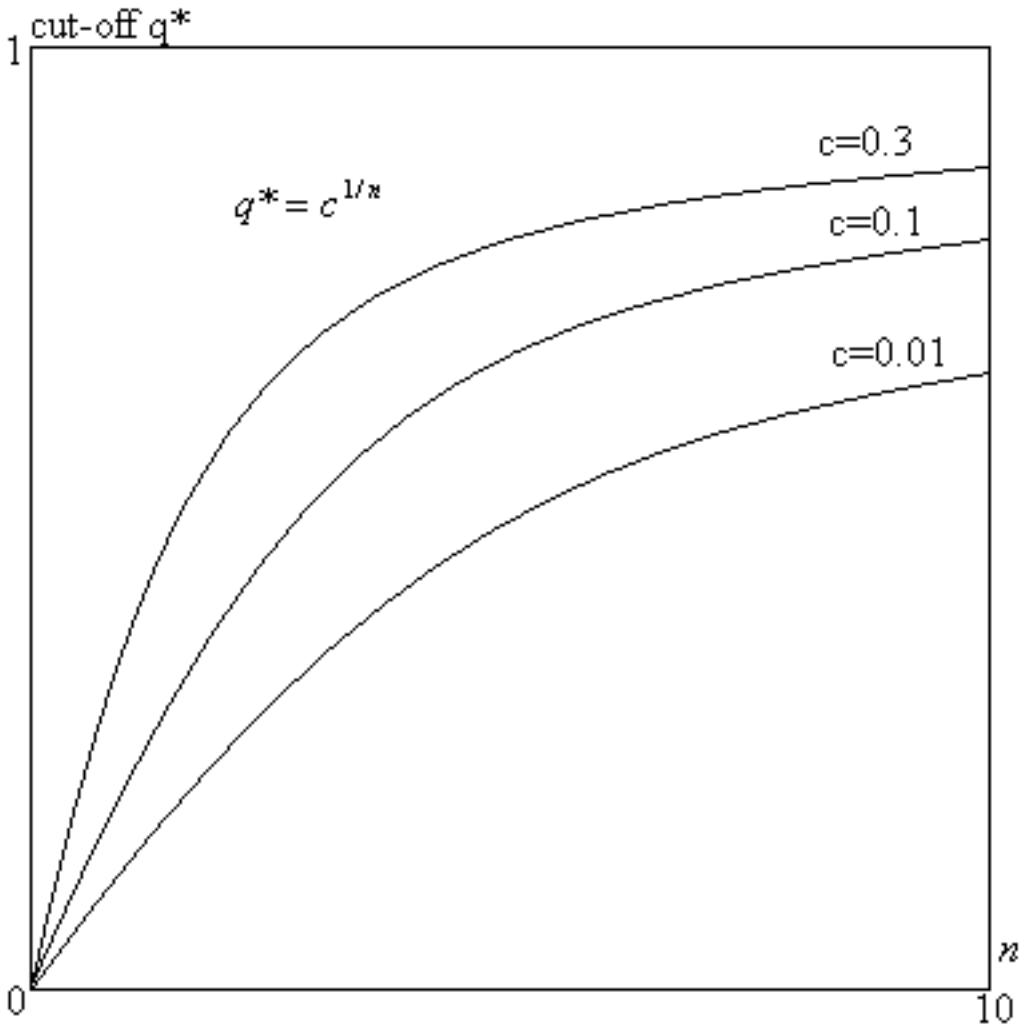

Figure 5:

6.4.2 Result 2

$$E[Y_n] = \left(\frac{n}{n+1}\right)(1 - q*^{n+1}) = \left(\frac{n}{n+1}\right)(1 - c^{\frac{n+1}{n}}) \tag{9}$$

See Appendix for proof of this result.

See Figure 6.

As $n \to \infty$, $E[Y_n]$ goes to $1 - c$. The expected quality increases quickly at first, and then more slowly, indicating that as n increases, the free-rider problem becomes worse. Still this is surprising, since even with the free-rider problem, the expected quality converges to a positive number. For low costs of contribution, which is likely with Wikipedia, the expected quality in the end can actually be very high. To contrast, if $c^{\frac{1}{n}}$ is not substituted for $q*$, in other words if there existed a beneficent ruler that set the threshold value $q*$, then $E[Y_n]$ converges to 1, since the threshold $q*$ should be less than 1. This means that when each individual plays strategically, the expected quality converges to a level lower than optimal. Thus, the free-rider problem is still evident in strategic play. Those with the highest levels of quality are not likely to contribute since they already possess the best information possible.

6.4.3 Result 3

Expected number of contributors is:

$$-\ln c \text{ as } n \to \infty \tag{10}$$

See Appendix for proof of this result.

Since Wikipedia has millions of users, setting n to be infinity is not unreasonable. If $c = 0.1$, then the expected number of contributors would be about 2.3, and the expected quality would be about 0.67. If $c = 0.01$, then the expected number of contributors would be 4.6, and the expected quality would be 0.77. As shown, even

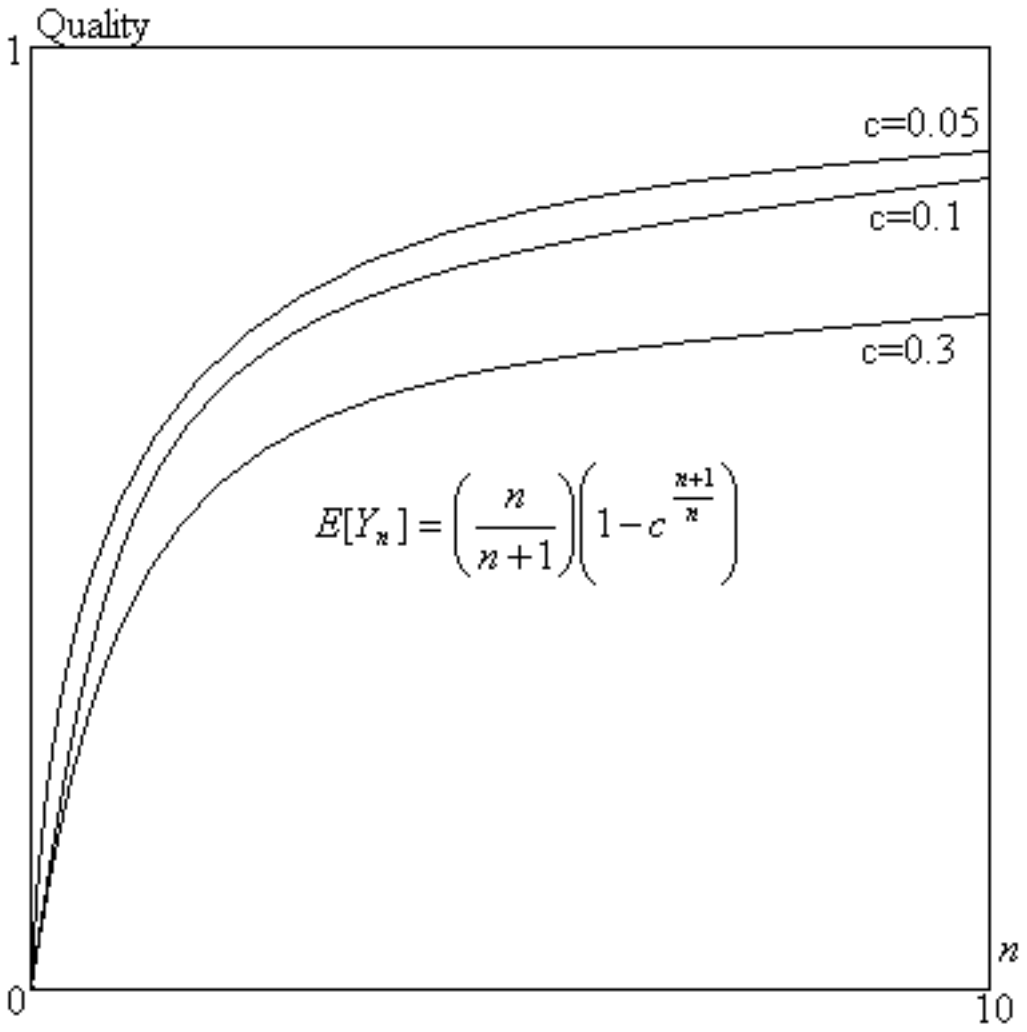
Quality
1
c=0.05
c=0.1
c=0.3
$E[Y_n] = \left(\frac{n}{n+1}\right)\left(1 - c^{\frac{n+1}{n}}\right)$
n
0
10

Figure 6:

with a low number of contributors, in the single digits, expected quality can be quite high. Again, we point out that it is reasonable to assume that the cost of contribution for Wikipedia is quite low, implying a high expected quality.

6.5 Encyclopedia as a n-shot simultaneous game

6.5.1 Result 1

Encyclopedia's expected quality is $\frac{n}{n+1}$
See Appendix for proof of this result.

This approaches 1 as $n \to \infty$. It is evident that there is no free-rider problem here, since there is only one person. With infinite revisions, the encyclopedia article will be perfect, which is expected.

6.5.2 Result 2

Print encyclopedia's expected quality as a function of s is $\frac{1/\sqrt{s}-1}{1/\sqrt{s}}$
See Appendix for proof of this result.

Unless s is 0, the writer will never reach perfect quality, since he or she will stop researching when further effort decreases his or her utility. Using the above result, we can find that, with $c = 0.1$, the writer's optimal strategy is to do research 2.16 times for a quality of 0.6838 (Since the number of research processes is an integer, we would practically take the floor of this number).

6.6 Comparison between simultaneous games

Since the expected number of contributors for Wikipedia as $n \to \infty$ is $-\ln c$, we can substitute $-\ln c$ for n to obtain the expected quality of Wikipedia as a function of c:

$$E[Y_n] = \left(\frac{-\ln c}{-\ln c + 1}\right)(1 - c^{\frac{-\ln c+1}{-\ln c}}) \tag{11}$$

See Appendix for proof of this result.

And we know from before that the optimal expected quality of encyclopedia as a function of s is:

$$\frac{1/\sqrt{s} - 1}{1/\sqrt{s}} \tag{12}$$

See Figure 7.

When plotted against each other, we can see that Wikipedia's expected quality overtakes that of print encyclopedia at $c = 0.1313$. This is quite surprising, since when c or s is very low, Wikipedia actually lags behind print encyclopedia due to its free-rider problem. Those with the highest quality are unlikely to contribute. However, as c or s increases, the number of expected revisions for encyclopedia drops significantly, whereas the number of expected contributors for Wikipedia does not drop as quickly. For example, at c or $s = 0.2$, the expected number of iterations for print encyclopedia is about 1.23, and the expected number of contributors for Wikipedia is about 1.61. Wikipedia overwhelms print encyclopedia with its sheer number of possible contributors. As indicated in the sequential games, s includes both cost of research and contribution for encyclopedia, so s is likely to be significantly higher than c. Given the already small difference, this allows for a shift where Wikipedia overtakes print encyclopedia for all costs of contribution.

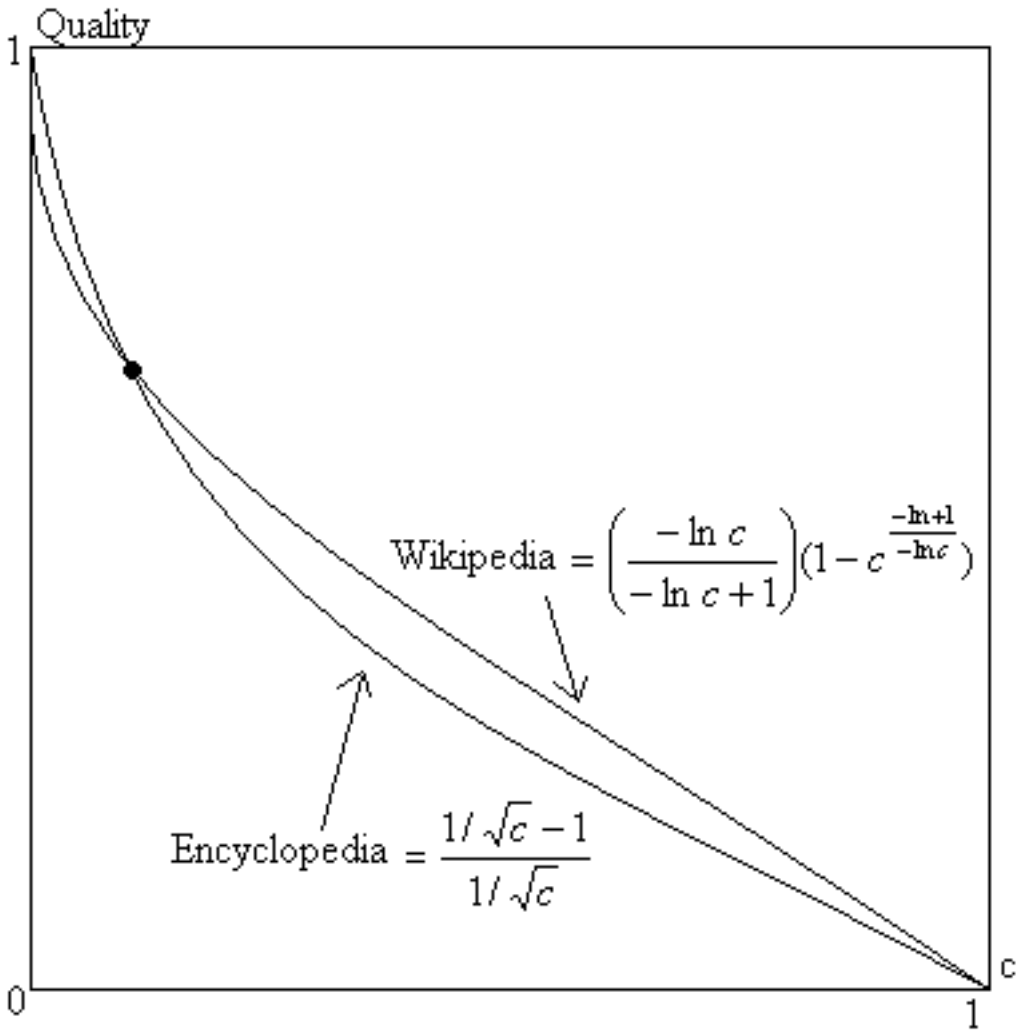
Quality
1
0
1
c
Wikipedia = $\left(\frac{-\ln c}{-\ln c + 1}\right)(1 - c^{\frac{-\ln+1}{-\ln c}})$
Encyclopedia = $\frac{1/\sqrt{c} - 1}{1/\sqrt{c}}$

Figure 7:

7 Conclusion

In this paper, we have formulated a theoretical model to compare the difference between the average expected quality from a case with two-person sequential contributions to Wikipedia and a case with one-writer sequential contributions to print encyclopedia. We have also formulated a theoretical model to compare the difference between the average expected quality from a case with n-person simultaneous contributions to Wikipedia and a case with one-writer simultaneous contributions to print encyclopedia. For the sequential games, we have found that the quality level is strictly higher for print encyclopedia for all costs of contribution c or s, except 0. However, the difference in quality is quite small for $c \leq 1/2$ or $s \leq 1/2$. If we take into account the likelihood that generally, $c < s$, the quality of Wikipedia may turn out to be higher than that of a regular encyclopedia as we shift the curve to accommodate for the difference in cost of contribution. For the simultaneous games, we have found that the quality of Wikipedia converges to a positive number as n reaches infinity. We have also found that the quality level is higher for print encyclopedia at very low costs of contribution, and the quality level is higher for Wikipedia at higher costs of contribution. Again, the difference is very small even at the largest gap, so as we take into account that $c < s$, the quality of Wikipedia may be strictly higher than that of a encyclopedia.

There are several important areas that this paper does not address. One possibility is developing a model that takes aggregate quality as additive instead of, or in addition to, it as maximal. In reality, many posts in Wikipedia develop through additions as well as revisions. One can also formally incorporate reader strategies to explain the wide adoption of Wikipedia. Finally, it may be worthwhile to further investigate the assumptions behind and implications of the results of the n person extension. It is a bit counter-intuitive to see the quality of print encyclopedia higher than that of Wikipedia

for low costs of contribution. After all, isn't Wikipedia all about low contribution costs? Although a reasonable explanation exists in the free-rider problem and a credible resolution exists in assuming higher costs of contribution across the board for print encyclopedia, it may be interesting to investigate how Wikipedia may not have such a low perceived cost of contribution at all, avoiding this discussion altogether. Future models may incorporate the dichotomy of dedicated experts versus anonymous novices who contribute to Wikipedia. Experts are likely to incur research costs similar to the writer of the print encyclopedia, while producing higher quality edits. This may explain how Wikipedia could actually demand higher costs of contribution

Nevertheless, these findings have significant implications for the publishers of encyclopedias. Wikipedia has been shown in this model, as well as in reality, to be a relatively reliable source of information, in many ways comparable to or better than print encyclopedia. It has more users and potential contributors, updates much faster than print encyclopedia, and is accessible instantaneously by anyone with internet. In addition to the *Nature* study, German computing magazine *c't* found Wikipedia to be even better than Encarta, and as of August 11, 2008, there are 490 academic papers in the ScienceDirect database that cites Wikipedia (Wikipedia 2008 [8]). These are testaments to what our model shows: Wikipedia has the potential to surpass the print encyclopedia. It is common knowledge that generally, all encyclopedias are not authoritative references. If they lose their edge on quality, the importance of print encyclopedias for research as well as casual browsing may diminish even further as people turn to another source, both more convenient and reliable.

8 Appendix

8.1 Wikipedia as a 2-person sequential game

8.1.1 Proof of result 1A:

If $q_1 \leq 1 - c$:

$$\overline{Q}_1 = \int_0^{q_1+c} q_1 \, dq_2 + \int_{q_1+c}^{1} q_2 \, dq_2 = \frac{1 + q_1^2 - c^2}{2} \tag{13}$$

If $q_1 > 1 - c$:

$$\overline{Q}_1 = \int_0^{1} q_1 \, dq_2 = q_1 \tag{14}$$

Combining:

$$\overline{Q}_1 = \begin{cases} \frac{1+q_1^2-c^2}{2} & \text{for } q_1 \leq 1 - c \\ q_1 & \text{for } q_1 > 1 - c \end{cases} \tag{15}$$

8.1.2 Proof of result 1B:

Given q and c, if $q_1 > 1 - c$:

$$U_{q_1} = \begin{cases} q_1 - c & \text{if contribute} \\ \frac{1-c^2}{2} & \text{if don't} \end{cases} \tag{16}$$

Contribute if $q_1 - c > \frac{1-c^2}{2}$ or $q_1 > \frac{1+2c-c^2}{2}$

If $q_1 \leq 1 - c$:

$$U_{q_1} = \begin{cases} \frac{1+q_1^2-c^2-2c}{2} & \text{if contribute} \\ \frac{1-c^2}{2} & \text{if don't} \end{cases} \tag{17}$$

Contribute if $\frac{1+q_1^2-c^2-2c}{2} > \frac{1-c^2}{2}$ or $1 - c > q_1 > \sqrt{2c}$.

Suppose $q_1 > 1 - c$, then if

$$\begin{cases} q_1 > \frac{1+2c-c^2}{2} & \text{contribute} \\ q_1 \leq \frac{1+2c-c^2}{2} & \text{don't} \end{cases} \tag{18}$$

$$\begin{cases} 1 - c > \frac{1+2c-c^2}{2} & \text{for } c < 2 - \sqrt{3} \\ 1 - c \leq \frac{1+2c-c^2}{2} & \text{for } c \geq 2 - \sqrt{3} \end{cases} \tag{19}$$

Thus, for $c < 2 - \sqrt{3}$, $q_1 > 1 - c > \frac{1+2c-c^2}{2}$ so I_1 always contributes.

For $1 \geq c \geq 2 - \sqrt{3}$, if

$$\begin{cases} q_1 > \frac{1+2c-c^2}{2} & \text{contribute} \\ 1 - c \leq q_1 \leq \frac{1+2c-c^2}{2} & \text{don't} \end{cases} \tag{20}$$

Suppose $q_1 \leq 1 - c$, then if

$$\begin{cases} q_1 > \sqrt{2c} & \text{contribute} \\ q_1 \leq \sqrt{2c} & \text{don't} \end{cases} \tag{21}$$

$$\begin{cases} 1 - c > \sqrt{2c} & \text{for } c < 2 - \sqrt{3} \\ 1 - c \leq \sqrt{2c} & \text{for } c \geq 2 - \sqrt{3} \end{cases} \tag{22}$$

Thus, for $c < 2 - \sqrt{3}$, if:

$$\begin{cases} \sqrt{2c} < q_1 \leq 1 - c & \text{contribute} \\ q_1 < \sqrt{2c} & \text{don't} \end{cases} \tag{23}$$

For $1 \geq c \geq 2 - \sqrt{3}$, $q_1 \leq 1 - c \leq \sqrt{2c}$ so I_1 never contributes

Combining:

For $c < 2 - \sqrt{3}$, if:

$$\begin{cases} q_1 > \sqrt{2c} & \text{contribute} \\ q_1 \leq \sqrt{2c} & \text{don't} \end{cases} \tag{24}$$

For $1 \geq c \geq 2 - \sqrt{3}$, if:

$$\begin{cases} q_1 > \frac{1+2c-c^2}{2} & \text{contribute} \\ q_1 \leq \frac{1+2c-c^2}{2} & \text{don't} \end{cases} \tag{25}$$

8.1.3 Proof of result 2:

For $c < 2 - \sqrt{3}$ and $q_1 < 1 - c$

$$\overline{Q}_2 = \int_0^{\sqrt{2c}} \int_c^1 q_2 \, dq_2 \, dq_1 + \int_{\sqrt{2c}}^{1-c} \left[\int_0^{q_1+c} q_1 \, dq_2 + \int_{q_1+c}^1 q_2 \, dq_2 \right] dq_1 = \frac{2 - 3c - c\sqrt{2c} + c^3}{3} \tag{26}$$

For $c < 2 - \sqrt{3}$ and $q_1 \geq 1 - c$

$$\overline{Q}_2 = \int_{1-c}^1 \int_0^1 q_1 \, dq_2 \, dq_1 = \frac{2c - c^2}{2} \tag{27}$$

For $c \geq 2 - \sqrt{3}$ and $q_1 < 1 - c$

$$\overline{Q}_2 = \int_0^{1-c} \int_c^1 q_2 \, dq_2 \, dq_1 = \frac{(c^2 - 1)(c - 1)}{2} \tag{28}$$

For $c \geq 2 - \sqrt{3}$ and $q_1 \geq 1 - c$

$$\overline{Q}_2 = \int_{1-c}^{\frac{1+2c-c^2}{2}} \int_c^1 q_2 \, dq_2 \, dq_1 + \int_{\frac{1+2c-c^2}{2}}^1 \int_0^1 q_1 \, dq_2 \, dq_1 = \frac{(c^2 - 1)(c^2 - 4c + 1)}{4} - \frac{(c - 1)^2(c^2 - 2c - 3)}{8} \tag{29}$$

After combining the first two expressions and the last two expressions:

$$\overline{Q}_2 = \begin{cases} \frac{4-2c\sqrt{2c}-3c^2+2c^3}{6} & \text{for } c < 2-\sqrt{3} \\ \frac{(c^2-1)(c^2-2c-1)}{4} - \frac{(c-1)^2(c^2-2c-3)}{8} & \text{for } c \geq 2-\sqrt{3} \end{cases} \tag{30}$$

8.2 Encyclopedia as a 2-shot sequential game

8.2.1 Proof of result 1:

$$E[\max\{q_1, q_2\}] - s \geq q_1 \Rightarrow \int_0^{q_1} q_1 \, dq_2 + \int_{q_1}^{1} q_2 \, dq_2 - s \geq q_1 \Rightarrow q_1 \leq 1 - \sqrt{2s} \tag{31}$$

Thus, for

$$\begin{cases} q_1 < 1-\sqrt{2s} & \text{writer researches further} \\ q_1 \geq 1-\sqrt{2s} & \text{writer stops researching} \end{cases} \tag{32}$$

8.2.2 Proof of result 1:

$$\overline{Q}_w = \int_0^{1-\sqrt{2s}} \left[\int_0^{q_1} q_1 \, dq_2 + \int_{q_1}^{1} q_2 \, dq_2 \right] dq_1 + \int_{1-\sqrt{2s}}^{1} q_1 \, dq_1 = \frac{2 - s\sqrt{2s}}{3} \text{ for } s < \frac{1}{2} \tag{33}$$

$$\overline{Q}_w = \int_0^{1} q_1 \, dq_1 = \frac{1}{2} \text{ for } s \geq \frac{1}{2} \tag{34}$$

$$\overline{Q}_w = \begin{cases} \frac{2-s\sqrt{2s}}{3} & \text{for } s < \frac{1}{2} \\ \frac{1}{2} & \text{for } s \geq \frac{1}{2} \end{cases} \tag{35}$$

8.3 Wikipedia as a n-person simultaneous game

8.3.1 Proof of result 1:

CDF of $Y_n = P(\hat{x}_1 \leq y) \times ... \times P(\hat{x}_n \leq y)$

Let

$$P(Y \leq y) = \begin{cases} \alpha & \text{if } y \leq q* \\ \beta & \text{if } y > q* \end{cases} \tag{36}$$

$$\alpha = P(x_1 \leq y) \times ... \times P(x_n \leq y) = q*^n \tag{37}$$

$$\beta = \alpha + \sum_{i=1}^{k} P[(Y \leq y) \text{ and } k \text{ of the uniforms} > q*] \text{ for } k \in [1, n] \tag{38}$$

$$= \alpha + \sum_{i=1}^{k} P(Y \leq y \,|k \text{ uniforms} > q*) P(k \text{ uniforms} > q*) \tag{39}$$

$$= q*^n + \sum_{i=1}^{n} \left(\frac{y - q*}{1 - q*}\right)^k \binom{n}{k} (1 - q*)^k (q*)^{n-k} \tag{40}$$

$$= q*^n + \sum_{i=1}^{n} (y - q*)^k \binom{n}{k} (q*)^{n-k} \tag{41}$$

Since $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-1}$, let $\begin{cases} a = y - q* \\ b = q* \end{cases}$

$$\beta = q*^n + (y - q* + q*)^n - q*^n = y^n \tag{42}$$

Thus,

$$Y_n = \begin{cases} q*^n & \text{for } y < q* \\ y^n & \text{for } y \geq q* \end{cases} \tag{43}$$

PDF of Y_n:

$$g(y) = \begin{cases} 0 & \text{for } y < q* \\ ny^{n-1} & \text{for } y \geq q* \end{cases} \tag{44}$$

$$E[Y_n] = \int_0^{q*} y \times 0 \, dy + \int_{q*}^{1} ny^n \, dy = \left(\frac{n}{n+1}\right)(1 - q*^{n+1}) \tag{45}$$

Let $\max\{Y_{n-1}, q*\} = \hat{Y}$

$$\hat{Y} = \begin{cases} 0 & \text{for } y < q* \\ q*^{n-1} & \text{for } y = q* \\ y^{n-1} & \text{for } y > q* \end{cases} \tag{46}$$

PDF of $\hat{Y}$:

$$\hat{g}(y) = \begin{cases} 0 & \text{for } y \leq q* \\ (n-1)y^{n-2} & \text{for } y > q* \end{cases} \tag{47}$$

$$E[\max\{Y_{n-1}, q*\}] = q*^{n-1} \times q* + \int_{q*}^{1} (n-1)y^{n-1} \, dy \tag{48}$$

$$= q*^n + \left(\frac{n-1}{n}\right)(1 - q*^n) \tag{49}$$

$$= \frac{n-1}{n} + \frac{1}{n} q*^n \tag{50}$$

$E[Y_{n-1}] = \left(\frac{n-1}{n}\right)(1 - q*^n)$

$$\frac{n-1}{n} + \frac{1}{n} q*^n - c = \left(\frac{n-1}{n}\right)(1 - q*^n) \tag{51}$$

$$q*^n = c \tag{52}$$

$$q* = c^{1/n} \tag{53}$$

8.3.2 Proof of result 2:

$$E[Y_n] = \left(\frac{n}{n+1}\right)(1 - q*^{n+1}) \tag{54}$$

Plugging in $q* = c^{1/n}$:

$$E[Y_n] = \left(\frac{n}{n+1}\right)(1 - c^{\frac{n+1}{n}}) \tag{55}$$

8.3.3 Proof of result 3:

$$E[\text{number of contributors}] = n(1 - c^{1/n}) \tag{56}$$

$$= \frac{1 - c^{1/n}}{1/n} \tag{57}$$

$$= -\ln c \text{ as } n \to \infty \tag{58}$$

8.4 Encyclopedia as a n-shot simultaneous game

8.4.1 Proof of result 1:

$Y_n = \max\{w_1, ..., w_n\}$
$P(Y_n \leq y) = P(w_1 \leq y) \times, ..., \times P(w_n \leq y) = q^n$

PDF of Y_n:

$$nq^{n-1} \tag{59}$$

$$E[Y_n] = \int_0^1 qnq^{n-1}\, dq = n \int_0^1 q^n\, dq = \frac{n}{n+1} \tag{60}$$

8.4.2 Proof of result 2:

$$\frac{d}{dn}\left(\frac{n}{n+1} - ns\right) = \frac{1}{(n+1)^2} - s = 0 \tag{61}$$

$$n = \frac{1}{\sqrt{s}} - 1 \tag{62}$$

Plugging into $E[Y_n] = \frac{n}{n+1}$:

$$E[Y_n] = \frac{1/\sqrt{s} - 1}{1/\sqrt{s}} \tag{63}$$

8.5 Comparison between simultaneous games

Plug $n = -\ln c$ into $E[Y_n] = \left(\frac{n}{n+1}\right)(1 - c^{\frac{n+1}{n}})$:

$$E[Y_n] = \left(\frac{-\ln c}{-\ln c + 1}\right)(1 - c^{\frac{-\ln c+1}{-\ln c}}) \tag{64}$$

Bibliography

[1] (2005). Schoolboy spots errors in Encyclopedia Britannica. Retrieved April 27, 2008, from *Guardian* Web site: http://education.guardian.co.uk/schools/story/0,5500,1399038,00.html.

[2] (2008). Wikipedia, Britannica: A Toss-Up. Retrieved April 27, 2008, from *Wired* Web site: http://www.wired.com/culture/lifestyle/news/2005/12/69844.

[3] Anthony, D., Smith, D., & Williamson, T. (2005). Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia. *Department of Computer Science, Dartmouth College.*

[4] Bryant, S., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. *Georgia Institute of Technology.*

[5] Hirshleifer, J. (1983). From weakest-link to best-shot: The voluntary provision of public goods. *Public Choice*, 41(3), 371-386.

[6] Neus, A. (2001). Managing Information Quality in Virtual Communities of Practice. *IQ 2001: The 6th International Conference on Information Quality at MIT.*

[7] Polborn, M. (2007). Competing for Recognition Through Public Good Provision. *University of Illinois, Department of Economics – CESifo (Center for Economic Studies and Ifo Institute for Economic Research).*

[8] Reliability of Wikipedia. (2008). In *Wikipedia: The Free Encyclopedia* [Web]. Wikimedia. Retrieved December 10, 2008, from http://en.wikipedia.org/wiki/Reliability_of_Wikipedia.

[9] Varian, H. R. (1992). Sequential contributions to public goods. *Journal of Public Economics*, 53(2), 165-186.

[10] Wikipedia:About. (2008). In *Wikipedia: The Free Encyclopedia* [Web]. Wikimedia. Retrieved April 27, 2008, from http://en.wikipedia.org/wiki/Wikipedia:About.

[11] Wilkinson, D., & Huberman, B. (2007). Cooperation and quality in Wikipedia. *Information Dynamics Laboratory, Hewlett-Packard Labs.*

[12] Zhang, X., & Zhu, F. (2006). Intrinsic Motivation of Open Content Contributors: the Case of Wikipedia. *HKUST and MIT Center for Digital Business*, Harvard Business School.