# Misalignment Between Supply and Demand of Quality Content in Peer Production Communities

**Morten Warncke-Wang, Vivek Ranjan, Loren Terveen, and Brent Hecht**
GroupLens Research, University of Minnesota
Minneapolis, MN 55455, USA
morten@cs.umn.edu, ranja008@umn.edu, terveen@cs.umn.edu, bhecht@cs.umn.edu

## Abstract

In peer production communities, individual community members typically decide for themselves where to make contributions, often driven by factors such as "fun" or a belief that "information should be free". However, the extent to which this bottom-up, interest-driven content production paradigm *meets the needs of consumers of this content* is unclear. In this paper, we introduce an analytical framework for studying the relationship between content production and consumption in peer production communities. Applying our framework to four large Wikipedia language editions, we find extensive misalignment between production and consumption in all of them. We also show that this misalignment has an enormous effect on Wikipedias readers. For example, over 1.5 billion monthly pageviews in the English Wikipedia go to articles that would be of much higher quality if editors optimally distributed their work to meet reader demand. Examining misalignment in more detail, we observe that there is an excess of high-quality content about certain specific topics, and that the majority of articles with insufficient quality are in a stable state (i.e. not breaking news). Finally, we discuss technologies and community practises that can help reduce the misalignment between the supply of and demand for high-quality content in peer production communities.

## Introduction

People all over the world turn to sites like Wikipedia for information as casual as the relationship status of favourite celebrities (Spoerri 2007) and as serious as facts about a disease with which they have just been diagnosed (Schwartz et al. 2006). Some of this content is in the form of high-quality articles with neutral treatments of the subject, excellent structure, and appropriate detail. However, other content is lacking in detail, is not written in an encyclopedic tone, and is generally of lower quality (Stvilia et al. 2008; Schneider, Passant, and Decker 2012).

Unlike work allocation processes in traditional content production organisations, peer production communities like Wikipedia generally have no central authority that directs work towards topics that are in high demand by consumers (e.g. Wikipedia readers). Instead, peer production contributors generally do work that they perceive as "fun" (Nov 2007) and work that is simultaneously neither too difficult nor too simple (Lakhani and Wolf 2005). These motivational factors may or may not lead to the production of high-quality content on topics of most interest to consumers.

In this paper, we investigate the relationship between the supply of high-quality peer produced content and the demand for it. Examining four of the most successful Wikipedia language editions (English, French, Russian, and Portuguese), we find a large degree of misalignment between supply and demand. We also find that this misalignment has an enormous impact on Wikipedia readers. For instance, in the English Wikipedia over 1.5 billion monthly views are to articles that would be of much higher quality if work was allocated optimally according to reader demand.

This analysis is based on our *Perfect Alignment Hypothesis* (PAH), a hypothesis that assumes an exact match between the supply of high-quality content and the demand for it. Using pageview data from the Wikimedia Foundation and the Wikipedia community's own article quality assessments, we compare the current state of supply and demand in each Wikipedia against the PAH, providing detailed insight into the amount of misalignment. Our application of the PAH as a tool allows us to uncover both the *extent* and *impact* of misalignment. At the same time, it describes an ideal situation that peer production communities might not be able to reach. In the discussion section we cover this and other aspects of the PAH in more detail, together with other implications of our results.

We also uncover several factors associated with supply/demand misalignment in Wikipedia. For instance, reader demand for some topics (e.g. Lesbian, Gay, Bisexual, and Transgender (LGBT) topics) far exceeds Wikipedia's supply, while other topics have a very large number of high-quality articles relative to the number of people reading them (e.g. military history). Where previous research has shown that Wikipedia is effective at quickly developing content about breaking news subjects (Keegan, Gergle, and Contractor 2013), our findings suggest that the quality of these efforts is not matching the demand. Additionally, we find that a majority of the articles in the highest demand appear to be continuously in high demand, suggesting that peer production communities also need to focus on improving the quality of this type of content.

In summary, the contributions of this paper are as follows:

1. Studying four successful Wikipedia communities, we show that reader demand for and contributor supply of quality content exhibits a great deal of misalignment; there is low demand for many high-quality articles and high demand for articles that need much improvement.
2. We show that this misalignment greatly impacts all four Wikipedia communities; 2 billion monthly pageviews (42.7%) are to articles of much lower quality than they would be if supply and demand were perfectly aligned.
3. Certain topics, such as articles about countries and LGBT issues, are over-represented amongst low-quality/high-demand articles. We also find topics represented on both extremes of the misalignment spectrum (e.g. military history), suggesting that the community can make more optimal choices about where to direct effort.
4. Using time-series analysis, we find that over half of the low-quality/high-demand articles are not related to significant changes in demand, such as breaking news events, suggesting that peer production communities wanting to reduce the misalignment require strategies to tackle both short-term and long-term demand for content.

This paper studies a consequence of task self-selection in peer production communities. We show that while these communities have been greatly successful at producing content, including some of excellent quality, they are less successful at producing quality content *on the topics of most interest to content consumers*. Our results and their implications should help peer production communities make informed decisions about where and how to direct work to meet demand. We next look at related research to provide additional context for this paper and its contributions.

## Related work

Previous research that has studied contributor motivation in peer production communities has found that consumer (reader) demand is generally not a large consideration in how contributors decide to allocate their work. Surveying Wikipedia editors, Nov (2007) found that "fun" was the primary motivator followed by agreement with ideology, for example that "information should be free". Lakhani and Wolf (2005) surveyed contributors to Free/Open Source Software and found that "enjoyment-based intrinsic motivation . . . is the strongest and most pervasive driver."

Despite reader demand not appearing high on the list of motivations expressed by contributors in peer production communities, there is some evidence that it might play a role. Reinoso (2011) studied several different language editions of Wikipedia, and found that views and edits were highly correlated in some languages (e.g. English), but not others (e.g. Japanese). In a study of the effects of redirects, which are special pages that transparently moves the visitor to a different page, Hill and Shaw (2014) also showed that when taking these redirects into account, there is a high correlation between popularity and number of edits to Wikipedia articles. In a working paper, Gorbatai (2014) found a positive relationship between Wikipedia article views and novice edits, but also that these novice edits were associated with a *decrease* in article quality. Contributions by experienced editors were instead associated with an *increase* in quality, but overall there was a very low correlation between popularity and quality. This motivates our work, which aims to paint a clear picture of how supply and demand are distributed, allowing peer production communities to take action and reduce the misalignment.

Our work is also motivated by previous research that looked at similarities and differences between reader and editor communities. West et al. (2012) used browser toolbar data to show that contributors are more active users of various Internet services (e.g. news sites and YouTube) compared to readers. For medical topics, Wikipedia has been shown to be a very popular information resource (Heilman et al. 2011), but one that does not necessarily supply information "clinically important to patient safety and care" (Clauson et al. 2008).

The work that specifically motivates the research presented here is that by Lehmann et al. (2014) and Gorbatai (2011). Lehmann et al. found that among biography articles in the English Wikipedia, the most popular articles were not necessarily those of the highest quality, and vice versa. Gorbatai identified a similar mismatch between popularity and quality. The goal of our research is to build on this work by both broadening and deepening our understanding of the relationship between the supply of and demand for quality content in peer production communities. More specifically:

1. We are the first to examine misalignment using a *grounded definition of alignment* (the Perfect Alignment Hypothesis). This allows us to scale our analysis to *entire language editions* (i.e. not just biography articles) while at the same time using structured frameworks for both popularity and quality.
2. We are the first to examine misalignment in *multiple communities*. We look at four separate peer production communities rather than just the English Wikipedia, establishing the robustness of our findings through replication.
3. We are the first to examine the *impact* of supply/demand misalignment, showing that the effects are substantial for millions of Wikipedia readers.
4. We are the first to examine the *character* of supply/demand misalignment, showing that certain topical domains are more associated with misalignment than others and that trending topics are a partial cause of misalignment, but not a dominant one.

## Research questions

The related work suggests that the efforts of contributors in peer production systems do not lead to an information repository whose quality is aligned with reader demand. Our first research question investigates the extent of this:

**RQ1:** How widespread is misalignment in peer production communities?

Misalignment of supply and demand will impact information consumers if topics of high interest (demand) do not

have high-quality content. Therefore, we pose a second research question that seeks to measure this impact:

**RQ2:** What is the impact of this misalignment on content consumers?

If supply and demand of quality content are misaligned, it is useful to understand the nature of this misalignment. Our third research question has two parts, each of which sheds a different light on misalignment:

**RQ3a:** What topics are over-represented amongst the low-quality/high-demand and high-quality/low-demand artifacts?

**RQ3b:** To what extent are low-quality/high-demand artifacts associated with significant surges in attention (i.e. "trending topics")?

We present findings that address these questions in the results section. First, however, we need a precise way to characterise (mis)alignment between the supply of and demand for high-quality content in peer production communities.

## The Perfect Alignment Hypothesis

Related work has indicated that supply and demand of content quality may be misaligned in peer production communities; we want a general way to measure this. We do so with a construct we call the *Perfect Alignment Hypothesis* (PAH). In this section, we define the PAH and in the subsequent sections we use it to study misalignment in Wikipedia.

Ideally, all artifacts in a peer production community would be of the highest possible quality. However, all peer production communities — even the very large English Wikipedia community — have a limited number of contributors and all contributors have a limited amount of available time. Given these limitations, some artifacts necessarily will be of lower quality. The Perfect Alignment Hypothesis imagines a situation in which the limited supply of contributor work is optimally applied such that the quality of artifacts perfectly matches the demand for them. In other words, under the Perfect Alignment Hypothesis, the Spearman's correlation coefficient between quality and consumer demand is exactly 1.0.

For example, in the English Wikipedia the quality scale is (from lowest class to highest): Stub, Start, C, B, Good Article, A, Featured Article. Our dataset from the English Wikipedia contains 4,353 Featured Articles, and under the conditions of the PAH, these would also be the 4,353 most viewed articles. The next-most-viewed 793 articles would be in the A class, then 19,914 Good Articles, and so on, with 2.2 million stubs being the least-viewed articles.

In the following sections, we will use the Perfect Alignment Hypothesis to understand exactly how far away each of our four Wikipedia communities is from "optimal". As we will see, the PAH allows both an overview of the general amount of (mis)alignment, and at the same time insight into how the demand varies across the spectrum of quality.

## Methods and Datasets

To enable the study of (mis)alignment in peer production communities, we needed examples of successful communities. We chose to study four Wikipedia language editions – English, French, Russian, and Portuguese – because they all have large amounts of content, active contributor communities, use a sufficiently fine-grained quality scale, and members of each community have provided quality ratings for a large proportion of their articles. Each of these four language editions have adopted a six- or seven-class assessment scale that editors use to assess the quality of an article.

The highest quality rating, in English called "Featured Article", is only given to articles that provide complete coverage of a specific topic in a "professional, outstanding, and thorough" way[1], such as the English article on Barack Obama. The lowest-quality articles are often called "stubs", which only provide a "very basic description of [a] topic" of only a paragraph or two. While it is community members without guaranteed subject matter expertise that are providing article quality ratings, Wikipedia's notion of article quality has been found to map closely onto pre-Wikipedia notions of encyclopedic quality (Stvilia et al. 2008), and research has also shown that these quality ratings correlate relatively well with reader judgement of article quality (Kittur and Kraut 2008). Choosing Wikipedia editions with a fine-grained quality scale allows us to capture features associated with article quality (e.g. references to sources, usage of illustrative images) across the full quality spectrum, an improvement over using a proxy measure like article length as applied by Lehmann et al. (2014).

It is important to note that although these four language editions are representative examples of successful peer production communities, they are all within the Wikipedia universe. Future research should investigate misalignment in other successful peer production communities (e.g. OpenStreetMap or Stack Overflow) and unsuccessful communities (e.g. many wikis on Wikia (Roth, Taraborelli, and Gilbert 2008; Zhu, Kraut, and Kittur 2014)).

All of the four language editions we studied use templates to organise assessed articles into a well-defined set of article categories reflecting the assessment rating. We identified the appropriate structure of category names for each language edition and gathered datasets of articles for each, removing "non-article" types such as lists and disambiguation pages[2]. Articles without assessment were also discarded because their quality is undefined. This data gathering process resulted in the number of rated articles as listed in Table 1.

We measure demand using Wikipedia article pageviews as made available by the Wikimedia Foundation[3], the best available source for per-article view data. Following Hill and Shaw's (2014) suggested best practice for handling Wikipedia article views, all results in this paper account for pageviews to an article coming in through redirects.

One of our research questions investigates shifts in article demand. We are most interested in understanding short-

[1] https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

[2] Pages listing links from a common term to specific variants, e.g. https://en.wikipedia.org/wiki/John_Smith

[3] http://dumps.wikimedia.org/other/pagecounts-raw/

| Language | Size | Ratings | Classes |
|---|---|---|---|
| English | 4.67M | 3.6M | 7 |
| French | 1.58M | 929k | 6 |
| Russian | 1.18M | 170k | 7 |
| Portuguese | 862k | 444k | 6 |

Table 1: Overview of the four Wikipedia language editions studied. "Size" is in number of articles, "Ratings" is the number of articles given at least one quality rating, and "Classes" shows the number of classes used in the rating scale. Abbreviations: M=million, k=thousand.

| Language | % Aligned | % HQ/HD | % HQ/LD |
|---|---|---|---|
| English | 62.5 | 5.0 | 47.2 |
| French | 67.8 | 4.3 | 62.7 |
| Portuguese | 77.6 | 8.9 | 52.0 |
| Russian | 45.8 | 5.9 | 23.6 |

Table 2: Overall proportion of alignment, proportion of highest-quality articles in the highest demand class (HQ/HD), and proportion of highest-quality articles in the two lowest demand classes (HQ/LD).

term shifts and expect articles that are in continuously high demand to remain stable through a dataset with a shorter time span. Due to Wikipedia having a weekly cycle for both reader views and edits across language editions (ten Thij et al. 2012; Yasseri, Sumi, and Kertész 2012), we define a study period of four weeks to approximate a calendar month. For the English edition, we gathered data from July 27 to August 24, 2014, while for the other three editions our data gathering spans November 2-30, 2014.

## Results

In this section, we present results for each of our research questions. We first study the amount of misalignment in our four Wikipedia editions before measuring the impact of misalignment on content consumers. We then turn our attention to studying (a) where the misalignment occurs, and (b) whether low-quality/high-demand articles are associated with significant shifts in attention.

### RQ1: The Extent of Misalignment

To understand the amount of misalignment in each of the four language editions, we first define a set of "Quality assessment classes", labelled $Q_1$ through $Q_7$, which will correspond to the equivalent Wikipedia assessment class in order from lowest to highest (for editions with six classes, $Q_6$ will be the highest quality class).

We also define a set of hypothesised "Perfect Alignment Hypothesis classes", labelled $PAH_1$ through $PAH_7$. For each Wikipedia edition, the number of articles in a PAH class will be equal to the corresponding quality class (e.g. in English Wikipedia there are 4,353 Featured articles and 2.2 million stubs, thus $Q_7$ and $PAH_7$ both contain 4,353 articles, and $Q_1$ and $PAH_1$ both contain 2.2 million articles). For each edition, we rank the articles by total number of views across the defined four-week window, and assign PAH classes according to rank (e.g. in English Wikipedia the 4,353 most-viewed articles are in the $PAH_7$ class, and the least-viewed 2.2 million are in $PAH_1$). We can then compare the actual quality assessments against the classes the PAH suggests articles should be in if supply and demand were in perfect alignment.

This comparison is shown in the confusion matrices in Tables 3-6. Under the PAH, we would expect that all cells off of the diagonal would have a zero in them, and this is clearly not the case. At a more detailed level, we see that across all languages there is a considerable number of very popular articles that are also not of the highest quality, as found in the non-grey cells in the rightmost column. In the English Wikipedia, 852 articles in the second-lowest quality class ("Start") are so popular that under the PAH they ought to be top quality ("Featured Articles"). Two such articles are "Wedding", a general topic that one would expect to be popular, and "Cisgender", for which Wikipedia has an opportunity to provide important information about a sensitive topic to readers.

Also visible in each of Tables 3-6 is the reverse phenomenon: articles being of significantly higher quality than they would be under the PAH, which are the cells in the bottom left corner of these tables. We see that in French Wikipedia (Table 5), the demand for 796 (57.6%) of the highest quality articles ("articles de qualité") does not justify their quality, landing them in the $PAH_2$ class. Some of these articles are about rather narrow topics (e.g. the themes in Robert Browning's poetry), but as we will later see in our investigation of misaligned topics, many are not.

To get a better understanding of alignment and misalignment, we measure the overall proportion of aligned articles, the proportion of highest-quality articles that are in alignment, and the proportion of highest-quality articles that are in $PAH_1$ and $PAH_2$. The results are shown in Table 2, and as we can see, in three of the four Wikipedia editions the majority of articles are in alignment. This is due to most articles being found in the two lowest quality classes, often named "Stub" and "Start", whereas there are much lower numbers of articles in the other classes. For example, as used in our explanation of the Perfect Alignment Hypothesis, our English Wikipedia dataset has 4,353 Featured Articles (0.1% of the total), but over 2.2 million Stub-class articles (62.4%). In Table 3, 1.7 million Stubs (76.8%) are in alignment, and the results for French and Portuguese are similar. In contrast, the Russian Wikipedia appears to have a significantly lower proportion of aligned articles. This might be due to our dataset of assessed articles in Russian covering a lower proportion of the total number of articles (5.9%) compared to the other Wikipedias.

Table 2 also reveals how these communities have been producing content of the highest quality in areas with a rather narrow audience. Only 4-9% of the highest-quality articles are in high enough demand to warrant their top quality rating under the conditions of the PAH. At the same time, we see that approximately 50-60% of these highest-quality articles are in comparatively low demand as they would be in one of the two lowest quality classes. Understanding

| | $PAH_1$ | $PAH_2$ | $PAH_3$ | $PAH_4$ | $PAH_5$ | $PAH_6$ | $PAH_7$ |
|---|---|---|---|---|---|---|---|
| $Q_1$ | 1,710,819 | 477,687 | 30,701 | 6,647 | 657 | 16 | 64 |
| $Q_2$ | 454,270 | 477,547 | 92,585 | 37,148 | 6,130 | 190 | 852 |
| $Q_3$ | 43,255 | 71,012 | 26,749 | 19,056 | 6,259 | 232 | 1,344 |
| $Q_4$ | 14,408 | 30,669 | 13,707 | 12,102 | 5,447 | 262 | 1,351 |
| $Q_5$ | 3,649 | 9,416 | 3,192 | 2,136 | 953 | 62 | 506 |
| $Q_6$ | 132 | 398 | 128 | 92 | 31 | 0 | 12 |
| $Q_7$ | 59 | 1,994 | 846 | 766 | 438 | 32 | 218 |

Table 3: Confusion matrix for supply (rows) and demand (columns) in English Wikipedia. Under perfect alignment all articles would be in the diagonal (grey) cells.

| | $PAH_1$ | $PAH_2$ | $PAH_3$ | $PAH_4$ | $PAH_5$ | $PAH_6$ | $PAH_7$ |
|---|---|---|---|---|---|---|---|
| $Q_1$ | 49,363 | 28,060 | 4,646 | 759 | 203 | 214 | 47 |
| $Q_2$ | 28,969 | 25,330 | 6,513 | 1,618 | 491 | 593 | 167 |
| $Q_3$ | 3,827 | 6,906 | 2,814 | 909 | 342 | 613 | 234 |
| $Q_4$ | 666 | 1,550 | 787 | 323 | 110 | 297 | 175 |
| $Q_5$ | 425 | 705 | 72 | 30 | 7 | 8 | 6 |
| $Q_6$ | 42 | 953 | 525 | 158 | 64 | 139 | 76 |
| $Q_7$ | 0 | 177 | 288 | 111 | 36 | 93 | 44 |

Table 4: Confusion matrix for supply (rows) and demand (columns) in Russian Wikipedia. Under perfect alignment all articles would be in the diagonal (grey) cells.

| | $PAH_1$ | $PAH_2$ | $PAH_3$ | $PAH_4$ | $PAH_5$ | $PAH_6$ |
|---|---|---|---|---|---|---|
| $Q_1$ | 548,038 | 125,803 | 6,796 | 138 | 243 | 94 |
| $Q_2$ | 124,025 | 78,243 | 13,712 | 574 | 972 | 578 |
| $Q_3$ | 8,392 | 11,402 | 4,476 | 345 | 797 | 518 |
| $Q_4$ | 273 | 490 | 217 | 25 | 67 | 69 |
| $Q_5$ | 314 | 1,370 | 370 | 27 | 66 | 63 |
| $Q_6$ | 70 | 796 | 359 | 32 | 65 | 60 |

Table 5: Confusion matrix for supply (rows) and demand (columns) in French Wikipedia. Under perfect alignment all articles would be in the diagonal (grey) cells.

| | $PAH_1$ | $PAH_2$ | $PAH_3$ | $PAH_4$ | $PAH_5$ | $PAH_6$ |
|---|---|---|---|---|---|---|
| $Q_1$ | 323,012 | 39,270 | 3,520 | 174 | 71 | 48 |
| $Q_2$ | 38,937 | 18,483 | 5,004 | 455 | 226 | 151 |
| $Q_3$ | 3,346 | 4,453 | 3,068 | 602 | 361 | 323 |
| $Q_4$ | 343 | 495 | 307 | 69 | 55 | 91 |
| $Q_5$ | 369 | 287 | 71 | 15 | 11 | 11 |
| $Q_6$ | 88 | 268 | 183 | 45 | 40 | 61 |

Table 6: Confusion matrix for supply (rows) and demand (columns) in Portuguese Wikipedia. Under perfect alignment all articles would be in the diagonal (grey) cells.

characteristics of these strongly misaligned articles can help peer production communities decide how and where to allocate resources, and this will be the focus of our third research question. At the same time, these results indicate that misalignment has potentially a large impact on content consumers, which is the topic we turn to next.

## RQ2: The Impact of Misalignment

Priedhorsky et al. (2007) used the notion of a *damaged view* to understand the impact of vandalism in Wikipedia. Similarly, we define the notion of a *misaligned view*, a view of an article that supplies a quality level not in alignment with its demand. Based on the confusion matrices shown in Tables 3-6, we define two types of misalignment: *excess quality (ExQ)*, where quality is higher than demand suggests; and *insufficient quality (InQ)*, where high demand is not met with high quality.

There are different degrees of misalignment, as shown in our confusion matrices, varying from no misalignment to articles of maximum quality being minimally popular. One approach to measure the impact is to use the distance between an article's $Q$ and $PAH$ class (e.g. a $Q_2$ article in $PAH_6$ has distance $d = 6 - 2 = 4$). A drawback with this approach is that the range of the distance varies depending on the assessment class; in English Wikipedia the $Q_7$ range is $[-6, 0]$, and $Q_2$ has range $[-1, 5]$. Since the number of articles varies greatly between classes, the results will most likely be strongly skewed. To avoid this problem, we collapse larger degrees of misalignment into a single category. If the distance between an article's assessment class and PAH class is greater than or equal to two, we define it as *strong* misalignment. For example, a $Q_2$ class article is in strong misalignment if its PAH class is $PAH_4$ or higher. As there are six or seven classes in total, two classes of misalignment will typically mean a significant increase or decrease in quality. Similarly, we define *moderate* misalignment as one-class misalignment.

Combining the notion of *strong* and *moderate* misalignment with *excess quality* and *insufficient quality*, we get a Likert-type scale with five categories: Strong ExQ, Moderate ExQ, Alignment, Moderate InQ, Strong InQ. We first use this scale to collapse our confusion matrix rows and columns, then combine them with *misaligned views* to aggregate article views over our four-week time span. The result is an estimate of the monthly impact of misalignment on Wikipedia's readers, and Table 7 provides an overview.

Whereas we previously found large proportions of overall alignment, the results shown in Table 7 make it clear that the misalignment that does exist has an *enormous impact* on content consumers. Across these four Wikipedias, two billion monthly pageviews are to articles that are in the Strong InQ category. In other words, *articles that are more than two quality classes lower than they would be if the supply of quality was in alignment with demand receive 2 billion pageviews a month*. We can also see that the proportion of views going to articles of insufficient quality varies across the language editions, but is substantial throughout. In all language editions, well over half of the pageviews go to articles that are either of moderate insufficient quality or strong insufficient quality. In the English Wikipedia, articles of strong insufficient quality alone receive close to half of the pageviews, and in the Russian Wikipedia, they receive more than half. Overall, these results suggest that the average Wikipedia reader frequently encounters articles that would be of much higher quality if the community distributed quality optimally according to reader demand.

## RQ3: Characterising Misalignment

**RQ3a: Misaligned Topics** For this research question, we investigated whether the supply and demand of quality content is especially misaligned for certain topical domains (e.g. biographies). In order to answer this question, we first had to identify a mechanism for categorising articles. Many Wikipedia editions have a robust dataset of user-generated category memberships, but these can be difficult to leverage to assign articles to a set of higher-level categories (Nastase and Strube 2008; Hecht 2013). Fortunately, the English Wikipedia has WikiProjects, which are groups of Wikipedia contributors interested in the same topic. As article quality assessments are done by WikiProject members, every article in our English dataset is associated with at least one project. For example, "WikiProject LGBT studies" covers articles about LGBT supporters and activists (e.g. Harvey Milk) as well as articles such as "Gay". Whereas the projects in the English Wikipedia have been studied extensively (Kittur, Pendleton, and Kraut 2009; Chen, Ren, and Riedl 2010; Choi et al. 2010; Forte et al. 2012; Zhu, Kraut, and Kittur 2012; Morgan et al. 2014), much less is known about the project infrastructure of the other language editions, leading us to focus this work on the English Wikipedia.

From our investigation into the extent of misalignment, we find two categories of misaligned articles that are strong candidates for further analysis. First are the most popular articles that are not also of the highest quality. Given their popularity and the huge impact of misalignment as we saw previously, these should be the articles the community is most interested in improving under the PAH. As we are studying the English Wikipedia, these articles are found in the rightmost column of Table 3, with the exception of articles already in $Q_7$. There are 4,135 articles in this class, which we will refer to as the "Needs Improvement" (NI) dataset.

The second group of articles are those that have reached the highest quality, but have relatively low popularity. Studying these should inform us about where the community exerts excess effort (under the PAH). These articles are found in the bottom row of Table 3. We focused on the leftmost two cells ($PAH_1$ and $PAH_2$) as they are in particularly strong misalignment and account for almost half (47.2%) of all top-quality articles. We will refer to these articles as the "Spent Effort" (SE) dataset.

The number of articles within the scope of each WikiProject differs, for example biographies are about five times more common than articles about the United States, and we have to account for these differences in underlying probability. To do so, we use *Relative Risk* (RR) to measure the extent to which a topic is over-represented, as that tells us "how much risk is increased or decreased from an initial level" (Davies, Crombie, and Tavakoli 1998). In our case,

| Language | | Strong ExQ | Moderate ExQ | Aligned | Moderate InQ | Strong InQ | Total |
|---|---|---|---|---|---|---|---|
| *English* | **N** | 89,902,800 | 202,851,495 | 858,479,337 | 1,072,060,036 | 1,696,921,186 | 3,920,214,854 |
| | **%** | 2.3 | 5.2 | 21.9 | 27.3 | 43.3 | 100.0 |
| *Russian* | **N** | 7,039,690 | 10,443,592 | 29,893,771 | 41,770,575 | 112,343,208 | 201,490,836 |
| | **%** | 3.5 | 5.2 | 14.8 | 20.7 | 55.8 | 100.0 |
| *French* | **N** | 6,724,070 | 18,807,331 | 105,348,978 | 132,932,332 | 120,913,575 | 384,726,286 |
| | **%** | 1.8 | 4.9 | 27.4 | 34.6 | 31.4 | 100.0 |
| *Portuguese* | **N** | 2,494,655 | 7,909,048 | 41,430,473 | 45,546,214 | 60,576,005 | 157,956,395 |
| | **%** | 1.6 | 5.0 | 26.2 | 28.8 | 38.3 | 100.0 |

Table 7: Number (N) and proportion (%) of article views per (mis)alignment category for each of the four Wikipedia editions. Proportions are relative to each language edition's total number of views in 28 days, as listed in the rightmost column.

| Rank | Topic | N | Rel. Risk |
|---|---|---|---|
| 1 | Countries | 144 | 506.9 |
| 2 | Pop music | 97 | 38.9 |
| 3 | Internet | 84 | 37.6 |
| 4 | Comedy | 134 | 21.9 |
| 5 | Technology | 58 | 15.8 |
| 6 | Religion | 121 | 15.8 |
| 7 | Science Fiction | 70 | 15.5 |
| 8 | Rock music | 84 | 11.4 |
| 9 | Psychology | 60 | 11.1 |
| 10 | LGBT studies | 136 | 9.1 |

Table 8: Topics most strongly over-represented in the Needs Improvement (NI) dataset, limited to topics w/at least 50 NI articles. "N" columns lists number of NI articles.

| Rank | Topic | N | Rel. Risk |
|---|---|---|---|
| 1 | Cricket | 65 | 159.0 |
| 2 | Tropical cyclones | 112 | 99.3 |
| 3 | Middle Ages | 87 | 13.4 |
| 4 | Politics | 147 | 12.0 |
| 5 | Fungi | 53 | 9.1 |
| 6 | Birds | 78 | 8.2 |
| 7 | Military history | 404 | 5.3 |
| 8 | Ships | 88 | 5.0 |
| 9 | England | 72 | 4.9 |
| 10 | Australia | 258 | 4.3 |

Table 9: Topics most strongly over-represented in the Spent Effort (SE) dataset, limited to topics w/at least 50 SE articles. "N" column lists number of SE articles.

the relative risk is the probability of encountering a topic in the NI/SE dataset divided by the probability of encountering a topic in the entire English Wikipedia dataset.

Table 8 describes the topics that are most over-represented amongst articles in the NI dataset. In order to filter out extremely specific topics (e.g. "Human Computer Interaction": 3 articles), which are affected by very low sample sizes, and balance specificity and generality, we restrict Table 8 to topics with more than 50 articles in the NI dataset. We see that countries is by far the most disproportionately represented topic. Most articles within the scope of this topic are general knowledge articles about a specific country, and as we see most of these are in high demand and have limited quality. There are also some pop culture topics such as "pop music", "comedy", and "rock music". Lastly, we find two important topics, psychology and LGBT, making the top 10, indicating that Wikipedia is an major resource for knowledge about these topics, but needs to deliver more high-quality content to be in alignment with demand.

Table 9 shows the topics most over-represented in the SE dataset, again limited to topics with at least 50 articles. Cricket is the highest ranked topic, perhaps exemplified by the existence of ten Featured Articles about players on the Australian cricket team in England in 1948. Articles about cricket were also found in the NI dataset, for instance the article about the game itself has enough demand to be $PAH_7$ but is now a middle-quality article[4]. This might be because it is easier or more exciting (Lakhani and Wolf 2005; Nov 2007) to write biographies about cricket players than it is to perfect an article about a more general topic.

Examining Tables 8 and 9 together, we find that articles about countries need improvement, whereas articles related to two specific countries, England and Australia, are over-represented on the opposite side of the misalignment spectrum. Previous research has identified "self-focus bias" in the English Wikipedia (Hecht and Gergle 2009). These results further substantiates that this is an issue the community should be aware of and seek to mitigate, as there is a disconnect between audience demand and contributor effort.

Our results also contain examples of how volunteer groups in peer production communities are not making "efficient" choices about where to supply quality improvements. One of the topics listed in Table 9, military history, is a very successful WikiProject with over a thousand active members (Forte et al. 2012). Its members have created several hundred articles that have reached the highest quality, but as we can see a large number of these articles are not in particularly high demand (e.g. several articles about battleships). At the same time, this project is also associated with 179 articles in the NI dataset (relative risk $= 1.16$), such as the articles about NATO, and the Vietnam War. We have shown that misalignment has a big impact on content consumers, and these results that point to "inefficient" effort focus motivate socio-technical solutions that we will return to in our discussion section.

[4]The article was demoted from Featured Article status in 2008.

**RQ3b: Demand Stability in Misalignment** The previously described misaligned topics included ones such as film and music, where the latest news about a celebrity or event could mean dramatic changes in the demand for specific articles. In this section we again study our Needs Improvement (NI) dataset, the most popular articles that are not also of the highest quality. As before, these are arguably the articles the community would be most interested in improving, but if they seek to do so, to what extent are they chasing a moving target (i.e. improving an article that will soon be unpopular once the subject is no longer in the news)? Here we analyse the extent of stability in demand for these articles.

We require a robust way to model temporal patterns in our article view data. Moving window detection is one approach that has been used on Wikipedia data to detect bursts (Emanuelsen and Holaker 2013), which would allow us to identify spikes in demand such as the death of the famous actor Robin Williams. Increases in demand can also be less dramatic but sustained over a longer period of time. In order to make it possible to detect both types of changes, we used the popular and well-studied ARIMA models for time-series data (Brockwell and Davis 2002; Hyndman and Khandakar 2008).

We first verify that this approach can successfully identify these types of demand changes in our English Wikipedia pageview data. From the NI dataset, we picked a random sample of 100 articles, manually inspected their article views during our four-week window, and labelled them as either having a significant surge in demand or not during this period. In this testset, 50 articles had a surge, and 50 did not. For each article, we downloaded pageview data for the eight weeks prior to our study window, and used that data to train an ARIMA model. Then, for each day in the four-week window, we forecasted a 99.7% confidence interval, labelled an article as having a surge if its view rate was larger than the confidence interval, and updated the model with that day's views. On this testset, all articles with a surge were labelled correctly, while two non-surging articles were incorrectly labelled positive due to random fluctuations in demand, for a total accuracy of 98%.

This approach was then applied to all the 4,135 articles in the Needs Improvement dataset. Of these, 1,918 (46.4%) were predicted as having a significant uptake in demand during our four-week window. Since the NI articles are the ones the community should be most interested in improving, discovering that the majority of these are in a stable state of high demand is an important finding, it suggests there are fundamental shortcomings in how peer production communities prioritise effort. At the same time, this result shows that it is important for these communities to also pay attention to fluctuations in demand, as those are also frequent amongst these examples of low-quality/high-demand content. This duality will be brought up again in the next section, where we discuss the implications of our findings.

## Discussion

Through our modelling of the *Perfect Alignment Hypothesis*, we investigated misalignment between quality content supply and demand in peer production communities, both on a general level, as well as in more detail. We found extensive alignment for low-quality/low-demand content, but strong misalignment for high-quality/high-demand content: a large proportion of high-quality content was in low demand, and the vast majority of high-demand articles were not of the highest quality. This misalignment has a big impact on content consumers: in Wikipedia a huge percentage of views are to articles that would be of considerably higher quality if quality supply and demand were more in alignment. In our coverage of previous research, we pointed to studies showing that contributors are primarily motivated by "fun" (Lakhani and Wolf 2005; Nov 2007). Previous work has also identified several types of biases in peer produced content (Hecht and Gergle 2009; Haklay 2010; Hecht and Gergle 2010; Lam et al. 2011; Reagle and Rhue 2011; Stephens 2013). Our results fit into this greater line of work, suggesting that the misalignment between supply and demand of quality content is another important issue that peer production communities need to put continued efforts into solving.

We used the *Perfect Alignment Hypothesis* as a tool to enable us to characterise the mismatch between supply and demand. Is the ideal situation described by the PAH desirable or even attainable? We realise that the success of peer production communities like Wikipedia has been driven largely by very prolific editors (Kittur et al. 2007; Priedhorsky et al. 2007) who maintain high level of activity throughout their lifetime in the system (Panciera, Halfaker, and Terveen 2009). Simplistic attempts to "force" volunteer contributors to work on high-demand topics rather than topics they find interesting and valuable may just cause them to leave or reduce their participation. Creating nuanced work suggestion mechanisms that balance contributor self-interest and audience demand is an important research challenge raised by our findings.

What should the locus of such mechanisms be? Self-organised groups of volunteers who express interest in improving content on a particular topic, the "WikiProjects" we used in our topic analysis, are one intriguing possibility. Previous work has shown that WikiProjects have goal-setting mechanisms that can motivate contributors towards group efforts (Zhu, Kraut, and Kittur 2012), but recall that we found that even within the scope of a project, misalignment occurs; we referred to the military history project, which contains both high-demand/low-quality and low-demand/high-quality articles. However, some of these groups have created hundreds of top-quality articles (Forte et al. 2012), meaning their members have acquired domain-specific knowledge that can benefit other groups as well. To address this, a tool like SuggestBot (Cosley et al. 2007) could be modified to suggest good candidate (i.e. low-quality/high-demand) articles for improvement, and also identify and suggest candidate sets of editors with the required range of topic and Wikipedia expertise, thus enabling efficient production of content in alignment with demand.

We also found a duality in whether highly popular articles are connected to a surge in demand. In the case of Wikipedia, it has been shown that it handles extreme cases of high demand well (Keegan, Gergle, and Contractor 2013),

but there are also examples of less extreme trends. One way to address this problem could be to organise groups of contributors who are willing and able to work on any kind of article, an editorial "rapid response team". The development, deployment, and study of tools to support such groups – for example, to identify trending topics early – are interesting venues for future research.

## Future Work and Limitations

We studied misalignment in the context of Wikipedia. Do our findings generalise to other peer production communities? Or is Wikipedia too idiosyncratic, say because of its scale, social norms, quality standards, or encyclopedic mission? To help assess generality, we have begun to study OpenStreetMap, a significantly different community.

We used Wikipedia's own assessment scale to measure quality. Assessments are done manually by Wikipedians, which has two key implications for our results. First, while Wikipedia quality assessments correspond well to existing notions of encyclopedic quality (Stvilia et al. 2008), they may contain noise as contributors differ in opinion about article quality or make inconsistent assessments. Second, there may be a delay between significant changes to an article and its subsequent (re)assessment, which in our case would translate to the article belonging to a lower assessment class than it should. We measured quality at the end of the study period to reduce this effect; we also note that reassessment delays themselves are a form of misalignment, which deserves further study.

Finally, to measure demand we used article pageview data and counted all pageviews equally. This is a proxy for content demand, as a human reading a Wikipedia article might not be interested in more than the lead section, or they might not read the article at all. The Wikipedia view data we used does not account for visits to Wikipedia's mobile site[5], and it is not known whether mobile views are uniformly distributed across Wikipedia. While we have no reason to suspect they are not, this is a source of uncertainty for our estimates of reader demand.

## Conclusion

We studied alignment between supply and demand of quality content in peer production communities in the context of four large Wikipedia editions and reached the following conclusions:

1. Reader demand and contributor supply for high-quality content exhibits a great deal of misalignment, with low demand for many high-quality articles, and vice versa.
2. This misalignment has a major impact. Across our four Wikipedia editions, 2 billion monthly article pageviews (42.7% of the total) are to articles of much lower quality than they would be if supply and demand were aligned.
3. Certain topic areas, e.g. countries and sensitive topics, are over-represented amongst articles in high demand but of low quality.
4. The misalignment of the articles in the highest demand is not solely due to breaking news or trending topics, as over half of them appear to be in a stable state of high demand.

We identified a number of areas for future design and research with the potential to address the misalignment problem in Wikipedia and generalised our findings to other peer production communities.

[5] https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2014-10-15/Traffic_report

## Acknowledgements

We thank our GroupLens colleagues for their support, our reviewers for their helpful suggestions, the Wikimedia Foundation and the Wiki ViewStats project for facilitating access to Wikipedia data, and all Wikipedia contributors for creating a great encyclopedia for us to study. This work has been funded in part by the National Science Foundation grants IIS-0808692, IIS-0968483, and IIS-1111201.

## References

Brockwell, P. J., and Davis, R. A. 2002. *Introduction to Time Series and Forecasting*, volume 1. Springer, 2 edition.

Chen, J.; Ren, Y.; and Riedl, J. 2010. The Effects of Diversity on Group Productivity and Member Withdrawal in Online Volunteer Groups. In *Proc. of CHI*.

Choi, B.; Alexander, K.; Kraut, R. E.; and Levine, J. M. 2010. Socialization tactics in wikipedia and their effects. In *Proc. of CSCW*, 107–116.

Clauson, K. A.; Polen, H. H.; Boulos, M. N. K.; and Dzenowagis, J. H. 2008. Scope, Completeness, and Accuracy of Drug Information in Wikipedia. *Annals of Pharmacotherapy* 42(12):1814–1821.

Cosley, D.; Frankowski, D.; Terveen, L.; and Riedl, J. 2007. SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. In *Proc. IUI*, 32–41.

Davies, H. T. O.; Crombie, I. K.; and Tavakoli, M. 1998. When can odds ratios mislead? *BMJ* 316(7136):989–991.

Emanuelsen, E., and Holaker, M. R. 2013. Event Detection using Wikipedia. Master's thesis, Norwegian University of Science and Technology.

Forte, A.; Kittur, N.; Larco, V.; Zhu, H.; Bruckman, A.; and Kraut, R. E. 2012. Coordination and Beyond: Social Functions of Groups in Open Content Production. In *Proc. of CSCW*, 417–426.

Gorbatai, A. D. 2011. Exploring Underproduction in Wikipedia. In *Proc. of WikiSym*, 205–206.

Gorbatai, A. D. 2014. The Paradox of Novice Contributions to Collective Production: Evidence from Wikipedia. *Available at SSRN 1949327*.

Haklay, M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design* 37(4):682.

Hecht, B., and Gergle, D. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proc. C&T*, 11–20.

Hecht, B., and Gergle, D. 2010. The tower of Babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proc. CHI*, 291–300.

Hecht, B. 2013. *The Mining and Application of Diverse Cultural Perspectives in User-Generated Content*. Ph.D. Dissertation, Northwestern University.

Heilman, J. M.; Kemmann, E.; Bonert, M.; Chatterjee, A.; Ragar, B.; Beards, G. M.; Iberri, D. J.; Harvey, M.; Thomas, B.; Stomp, W.; et al. 2011. Wikipedia: A Key Tool for Global Public Health Promotion. *Journal of Medical Internet Research* 13(1).

Hill, B. M., and Shaw, A. 2014. Consider the Redirect: A Missing Dimension of Wikipedia Research. In *Proc. OpenSym/WikiSym*, F5:1–F5:4.

Hyndman, R. J., and Khandakar, Y. 2008. Automatic time series for forecasting: The forecast package for R. *Journal of Statistical Software* 27(3):1–22.

Keegan, B.; Gergle, D.; and Contractor, N. 2013. Hot Off the Wiki: Structures and Dynamics of Wikipedias Coverage of Breaking News Events. *American Behavioral Scientist* 57(5):595–622.

Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In *Proc. CSCW*.

Kittur, A.; Chi, E. H.; Pendleton, B. A.; Suh, B.; and Mytkowicz, T. 2007. Power of the Few vs. Wisdom of the Wrowd: Wikipedia and the Rise of the Bourgeoisie.

Kittur, A.; Pendleton, B.; and Kraut, R. E. 2009. Herding the Cats: The Influence of Groups in Coordinating Peer Production. In *Proc. of WikiSym*, 7:1–7:9.

Lakhani, K. R., and Wolf, R. G. 2005. Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects. In J. Feller, B. Fitzgerald, S. H., and Lakhani, K. R., eds., *Perspectives on Free and Open Source Software*. MIT Press.

Lam, S. T. K.; Uduwage, A.; Dong, Z.; Sen, S.; Musicant, D. R.; Terveen, L.; and Riedl, J. 2011. WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. In *Proc. of WikiSym*, 1–10.

Lehmann, J.; Müller-Birn, C.; Laniado, D.; Lalmas, M.; and Kaltenbrunner, A. 2014. Reader Preferences and Behavior on Wikipedia. In *Proc. HT*, 88–97. ACM.

Morgan, J. T.; Gilbert, M.; McDonald, D. W.; and Zachry, M. 2014. Editing Beyond Articles: Diversity & Dynamics of Teamwork in Open Collaborations. In *Proc. of CSCW*, 550–563.

Nastase, V., and Strube, M. 2008. Decoding Wikipedia Categories for Knowledge Acquisition. In *AAAI*.

Nov, O. 2007. What Motivates Wikipedians? *Commun. ACM* 50(11):60–64.

Panciera, K.; Halfaker, A.; and Terveen, L. 2009. Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *Proc. GROUP*, 51–60.

Priedhorsky, R.; Chen, J.; Lam, S. T. K.; Panciera, K.; Terveen, L.; and Riedl, J. 2007. Creating, Destroying, and Restoring value in Wikipedia. In *Proc. of GROUP*, 259–268.

Reagle, J., and Rhue, L. 2011. Gender Bias in Wikipedia and Britannica. *International Journal of Communication* 5(0).

Reinoso, A. J. 2011. *Temporal and behavioral patterns in the use of Wikipedia*. Ph.D. Dissertation, Universidad Rey Juan Carlos.

Roth, C.; Taraborelli, D.; and Gilbert, N. 2008. Measuring wiki viability: an empirical assessment of the social dynamics of a large sample of wikis. In *Proc. WikiSym*.

Schneider, J.; Passant, A.; and Decker, S. 2012. Deletion Discussions in Wikipedia: Decision Factors and Outcomes. In *Proc. WikiSym/OpenSym*. ACM.

Schwartz, K. L.; Roe, T.; Northrup, J.; Meza, J.; Seifeldin, R.; and Neale, A. V. 2006. Family Medicine Patients Use of the Internet for Health Information: a MetroNet Study. *The Journal of the American Board of Family Medicine* 19(1):39–45.

Spoerri, A. 2007. What is popular on Wikipedia and why? *First Monday* 12(4).

Stephens, M. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal* 1–16. 00001.

Stvilia, B.; Twidale, M. B.; Smith, L. C.; and Gasser, L. 2008. Information quality work organization in Wikipedia. *J. Am. Soc. Inf. Sci. Technol.* 59:983–1001.

ten Thij, M.; Volkovich, Y.; Laniado, D.; and Kaltenbrunner, A. 2012. Modeling and predicting page-view dynamics on wikipedia. *CoRR* abs/1212.5943.

West, R.; Weber, I.; and Castillo, C. 2012. Drawing a Data-driven Portrait of Wikipedia Editors. In *Proc. of OpenSym/WikiSym*, 3:1–3:10.

Yasseri, T.; Sumi, R.; and Kertész, J. 2012. Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis. *PLoS ONE* 7(1):e30091.

Zhu, H.; Kraut, R.; and Kittur, A. 2012. Organizing without formal organization: Group identification, goal setting and social modeling in directing online production. In *Proc. of CSCW*, 935–944.

Zhu, H.; Kraut, R. E.; and Kittur, A. 2014. The Impact of Membership Overlap on the Survival of Online Communities. In *Proc. of CHI*, 281–290.