The Production and Consumption of Quality Content in Peer Production Communities

A Dissertation
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Morten Warncke-Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY

Loren Terveen and Brent Hecht

December, 2016

Dedicated in memory of

John. T. Riedl, Ph.D

professor, advisor, mentor, squash player, friend

Acknowledgements

This work would not be possible without the help and support of a great number of people. I would first like to thank my family, in particular my parents Kari and Hans who have had to see yet another of their offspring fly off to the United States for a prolonged period of time. They have supported me in pursuing my dreams even when it meant I would be away from them. I also thank my brother Hans and his fiancée Anna for helping me get settled in Minnesota, solving everyday problems, and for the good company whenever I stopped by. Likewise, I thank my brother Carl and his wife Bente whom I have only been able to see ever so often, but it has always been moments I have cherished. Thanks as well to my nieces and nephews who brighten up the day whenever I get to visit, and help me remember that discovering the world is always a lot of fun!

John Riedl was a fantastic advisor and mentor to me, and Loren Terveen and Brent Hecht did an excellent job of helping me refocus and keep going after John's unfortunate passing. Since then, Loren and Brent have been wonderful advisors who have kept me on track with producing solid research and I am grateful for everything I have learned from them. I would also like to thank the other two members of my committee, Ching Ren and John Carlis, who have both showed me how to be a better researcher.

Thanks as well to everyone at GroupLens Research for creating and supporting an environment where science prospers. First of all, thank you to Joe Konstan, John Riedl, and Loren Terveen for seeing my potential when I did not myself and being instrumental in my move to Minnesota to pursue a Ph.D. Secondly, thank you to my senior students who led by excellent example and helped shape my work: Tony Lam, Aaron Halfaker, Katie Panciera, and Michael Ekstrand. Thanks also to the undergraduate students I got to work and publish papers with: Vlad Ayukaev, Vivek Ranjan, and Connor McMahon. I would also like to thank my fellow graduate students Jacob Thebault-Spieker and Isaac Johnson for helping me understand how geographic science works and helping me make sense of our findings. GroupLens has also had some great lab assistants and I thank them for helping out with everyday issues, in particular I would like to thank Angela Brandt for welcoming me to GroupLens and helping me understand how things work in the Midwest. Lastly,

as mentioned GroupLens has a fantastic environment where science prospers, and I would have liked to name every single person for being awesome and helping that happen!

My fellow researchers also deserve many thanks. Dan Cosley helped make our Wikipedia machine learning work a reality and I have greatly benefited from his research experience and perspectives. He also created SuggestBot, which has now been serving recommendations to Wikipedia contributors for about ten years and keeps going strong. Zhenhua Dong was a wonderful collaborator on my first paper and his tireless efforts still inspire me. Last, but very much not the least, thank you to Anuradha Uduwage for defying categorisation by being a great collaborator, graduate student, and friend. I would also like to thank my fellow researchers at the Wikimedia Foundation for the opportunities they have given me as a Research Fellow there. Thanks also goes to students and faculty at the University of Washington for welcoming me to the Pacific Northwest. In particular I would like to thank professors Benjamin Mako Hill and David McDonald for providing me with places to contribute and study, as well as graduate students Amirah Majid and Amanda Menking for making the research lab a place that feels like home.

I am also sending thanks to my friends on both sides of the Atlantic ocean who help keep my eyes open to fresh perspectives, have my back when I struggle, make sure I laugh, and bring a smile to my face whenever I think of them.

I cannot thank my girlfriend Cameo enough for her support over the past three and a half years as I have worked on completing this thesis. She has helped me find alternative perspectives, asked challenging questions, but first and foremost made sure that I do not go through a single day without feeling appreciated nor without at least one good laugh.

Finally, I would like to thank the National Science Foundation for financial support under a variety of grants: IIS 08-08692, IIS 08-45351, IIS 09-68483, and IIS 11-11201.

Abstract

Over the past 25 years, commons-based peer production [Ben02] has become a vital part of the information technology landscape. There are successful projects in different areas such as open source software (e.g. Apache and Firefox), encyclopedias (e.g. Wikipedia), and map data (e.g. OpenStreetMap). A common theme in all these communities is that they are mainly volunteer-driven and that contributors are able to self-select what they want to work on. Studies on contributor motivation in peer production have found "fun" and "appropriate challenges" to be strong factors [Nov07; LW05], both associated with the sensation of vital engagement often referred to as "flow" [NC03]. Peer production contributors also often refer to altruism, the desire to be helpful to others, as a motivating factor [BH13]. To what extent does this bottom-up, interest-driven, volunteer-based content production paradigm *meet the needs of consumers of this content*?

This thesis presents our work on improving our understanding of how peer production communities produce quality content and whether said quality content is produced in areas where there is demand for it. We study this from three perspectives and make contributions as follows: we investigate what textual features describe content quality in Wikipedia and develop a high-performance prediction model solely based on features contributors can easily improve (so called "actionable features"); we apply a coherent framework for describing and evaluating quality improvement projects in order to discover factors associated with the success and failure of these types of projects; we introduce an analytical framework that allows is to identify the misalignment between supply of and demand for quality content in peer production communities and measure the impact this has on a community's audience.

The research presented in this thesis provides us with a deeper understanding of quality content in peer production communities. These communities have created software, encyclopedic content, and maps that in many ways improve our everyday lives as well as those of millions of others. At the same time we have identified areas where there is a shortage of quality content and discussed future work that can help reduce this problem. This thesis thus lays the foundation upon which we can build improved communities and positively impact a large part of the world's population.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the past 25 years, commons-based peer production [Ben02] has become a vital part of the information technology landscape. Open Source Software projects such as the Linux operating system and the Firefox web browser are well-known examples of computer software developed through peer production. The area of collecting knowledge in the form of encyclopedias has been transformed from printed books edited by select staff into Wikipedia, an online resource that is continuously updated by thousands of volunteers from all over the world and has become one of the world's ten most popular websites[1]. In a way similar to how Wikipedia has been created, OpenStreetMap aims to provide geographic data such as maps for the entire globe. Over the past decade it has gone from an idea to being the basis of geographic applications such as Mapbox[2], and it also supplies maps to Craiglist, the online classified ad website[3]. Lastly, there are "community question and answer sites" such as Stack Overflow, where people can ask questions, typically on a specific topic (e.g. Stack Overflow is for software programming), and a community of volunteers will answer them.

A common theme in all these communities is that they are mainly volunteer-

[1] `http://www.alexa.com/topsites` ranks Wikipedia fifth as of Nov 1, 2016.

[2] `https://www.mapbox.com`

[3] `http://arstechnica.com/tech-policy/2012/08/craigslist-backtracks-drops-exclusive-licensing-on-posts/`

driven and that contributors are able to self-select what they want to work on. While there are various elements of central coordination and oversight in some of them, e.g. Linus Torvalds controls the core parts of Linux [LT02], task selection is typically controlled by the volunteer themselves. Studies on contributor motivation in peer production have found "fun" and "appropriate challenges" to be strong factors [Nov07; LW05], both associated with the sensation of vital engagement often referred to as "flow" [NC03]. Peer production contributors also often refer to altruism, the desire to be helpful to others, as a motivating factor [BH13].

While the decentralised volunteer-driven process in peer production is different from more traditional approaches, it produces high-quality content. Yochai Benkler argues that peer production has major advantages in "acquiring and processing information about human capital" [Ben02]. In the software industry the open source-developed Apache web server was for many years the dominant platform[4], out-competing commercial solutions. Wikipedia's notion of what constitutes high quality is similar to traditional encyclopedias [Stv+08a] and the accuracy of its articles has been found to be comparable to other encyclopedias [Gil05; Che06].

The question is whether this volunteer-driven self-selection process leads to the creation of quality content with a certain bias. There is the potential of bias in where content is produced (e.g. topic areas in Wikipedia), as well as bias in the way it is produced (e.g. content about a specific topic). This thesis concerns itself with the former of those to, where for instance research on Wikipedia discovered that there is a much larger amount of content in articles about movies with a dominantly male audience compared to those with a more female audience [Lam+11]. A study that

[4]Per the Netcraft web server survey: `https://news.netcraft.com/archives/2016/09/19/september-2016-web-server-survey.html`

compared biographies in Wikipedia to those in the Encyclopædia Britannica found that the former had better coverage and longer articles, but was also more likely to be missing articles about women [RR11]. The problem is not limited to Wikipedia, a study of the quality of the map of England in OpenStreetMap found that lower-income areas have considerably lower coverage and content quality [Hak10]. OpenStreetMap also uses keywords known as "tags" to describe points of interest, and a study of these found a much larger variety of tags available to describe places related to prostitution than those related to child care [Ste13].

This thesis focuses on the production and consumption of quality content in peer production communities. Specifically, we examine this topic from the following three perspectives:

1. What is quality content and how can we predict it?
2. How is quality content produced?
3. Is quality content produced where there is demand for it?

We study these issues in two representative peer production communities: Wikipedia and OpenStreetMap. Both are very successful and have a long history, Wikipedia being founded in early 2001 and OpenStreetMap in 2004. They also produce different types of content. Wikipedia is an encyclopedia and is thereby mainly text, whereas OpenStreetMap's content is in the literature referred to as "volunteered geographic information" [Goo07] (VGI), which is information about the location of things such as points of interest as well as description of these (e.g. a building may be labelled as a restaurant). We study all of the three issues in Wikipedia and examine the supply/demand imbalance in OpenStreetMap as well.

In all three areas we advance the state of the art. First, in Chapter 3 we investigate what textual features describe content quality in Wikipedia and develop a high-performance prediction model solely based on features contributors can easily improve (so called "actionable features"). Secondly, in Chapter 4 we consolidate and extend existing research using a coherent framework for describing and evaluating quality improvement projects. We then apply this framework to a diverse set of projects and discover factors related to the success and failure of these projects. Lastly, in Chapter 5 we identify the misalignment between supply of and demand for quality content in peer production communities. We also measure the impact this misalignment has on content consumers and discover characteristics of where the issue occurs.

Each perspective corresponds to a chapter in this thesis. We cover related work and background literature in Chapter 2. In Chapter 6 we discuss the broader impact of this work as well as how it has opened up venues for future research. As we will see, the work done in this thesis lays down the foundation for future studies that will further enhance our understanding the underlying issues of where quality content is created in peer production communities, and development of sociotechnical solutions that can assist communities with creating or improving quality in areas where it is going to have a greater impact on the community's audience.

Chapter 2

Background

Peer production communities, or more specifically "commons-based peer production" [Ben02], are typically communities made up of volunteers who make contributions to some kind of repository, where said repository is not owned by any specific individual, hence the term "commons". These contributions can come in the form of knowledge (e.g. how to make certain decisions) or as production of content that is part of the repository (e.g. software code). Because this repository is not owned by a specific individual or group, anyone is free to adopt it, adapt it, or extend it, thus making either new or derivative works. The licenses of the repository often specifically define that derivative works also be openly available, for instance the Creative Commons License[1] defines a "share alike" clause that may be used and this require any derivative work to also be openly licensed.

There are many types of peer production communities around. Some of the more successful ones are Open-Source Software projects such as the Linux operating system, the Apache web server, or the PostgreSQL database server. This thesis focuses on two other communities: Wikipedia, which aims to create a freely available encyclopedia[2], and OpenStreetMap (OSM), where the goal is to "create and provide free geographic data such as street maps to anyone who wants them."[3] [HW08]. Both of

[1] `https://creativecommons.org`
[2] `https://en.wikipedia.org/wiki/Wikipedia:Five_pillars`
[3] `https://blog.openstreetmap.org/faq/`

these communities have been immensely successful towards reaching their goals with thousands of people making contributions, and these communities also share much data freely, thereby providing us with a rich resource on which to do research.

2.1 Peer-produced Content Quality

This thesis concerns itself with the production and consumption of *quality content* in peer production communities. What does "quality content" mean in this context? We will describe this in more detail and refer to relevant research, first for Wikipedia, then for OpenStreetMap.

2.1.1 Content Quality in Wikipedia

The goal of Wikipedia is, as described in the five pillars referred to earlier, to create a freely available encyclopedia. Encyclopedias have been around for several hundred years, the French Encyclopédie was started in the mid-eighteenth century. There is therefore an existing notion of quality of encyclopedias, and studies of content quality in the English Wikipedia edition have compared the two, finding that what Wikipedia considers "quality content" is similar to that of traditional encyclopedias [Stv+05b].

What the English Wikipedia regards as quality has not all been static since that language edition was started in January 2001. When the category for the highest quality articles, named "Featured Articles" was created, the only criteria for being included was "brilliant prose" [Stv+08a]. Now articles go through a peer review process before reaching that status, and are required to be both well-written as well as meeting other criteria such as all claims referencing relevant external sources, proper media usage such as illustrative images and/or video [VWM07].

There are over 290 Wikipedia editions[4], and their notion of quality differs somewhat. All of them share the "five pillars of Wikipedia" mentioned previously, meaning that they all have the same goal of creating an encyclopedia, they have to cite reliable sources, and so forth. At the same time, there are nuances in for instance what criteria they use to define articles as the highest quality, allowing them to be promoted to "Featured Article" status. In a study, Stvilia et al. [SAY09], found that the English and Arabic editions had the exact same requirements for promotion, while the Korean edition differed slightly. It is therefore important to keep in mind that culture might affect what these communities perceive as quality. The Wikipedia editions are also very different in size, meaning that a smaller edition might focus on expansion, while one that has most subjects covered might focus on improving the quality of existing content, a topic we return to in Chapter 4 when we discuss quality improvement projects.

2.1.2 Content Quality in OpenStreetMap

Where Wikipedia aims to build an encyclopedia, OSM aims to create freely accessible maps. This type of content is commonly referred to as "Volunteered Geographic Information" [Goo07] in the research literature. When it comes to the quality of geographic information, it has been standardised by the International Organization for Standardization (ISO) in standard number 19157 [13]. That standard defines five parameters of quality: completeness, logical consistency, positional accuracy, temporal accuracy, and thematic accuracy.

Similarly as for Wikipedia, the OpenStreetMap community is concerned about

[4] `https://meta.wikimedia.org/wiki/List_of_Wikipedias`

information quality. On the OSM wiki we can find pages about quality assurance[5], and they also have pages describing issues related to accuracy[6] and completeness[7]. In addition to these, there is also a vast amount of information regarding the usage of "tags", a set of key-value pairs of text that can be used to describe features of the map[8] such as a building being a restaurant (where the key would be "amenity" and the value "restaurant"). When it comes to these tags, information quality is sustained through community norms on what is regarded as appropriate usage, which is then documented on the wiki and implemented in some of the popular map editing tools built for OSM.

Research on information quality in OSM has mainly studied two of the five parameters in the ISO standard: positional accuracy and completeness. The common approach in these studies have been to compare OSM against a *gold standard* dataset, for example studies of OSM in England has compared it to datasets from the Ordnance Survey [Hak10; Hak+10], the government organisation that has been mapping the United Kingdom since the 1700s. Other papers have been taken the same approach in other areas such as Germany [ZZ10; LVK11] and Ireland [Cie+10]. One study [Ars+13] focuses on the town of Heidelberg, Germany, and assesses three of the five parameters: positional accuracy, completeness, and semantic accuracy. They find that OSM to a large degree covered the area with as much data and accuracy as the golden standard dataset they compared it with.

[5] `http://wiki.openstreetmap.org/wiki/Quality_assurance`
[6] `http://wiki.openstreetmap.org/wiki/Accuracy`
[7] `http://wiki.openstreetmap.org/wiki/Completeness`
[8] `http://wiki.openstreetmap.org/wiki/Map_Features`

2.2 How is Content Quality Assessed?

2.2.1 Quality Assessment in Wikipedia

Wikipedia's contributors assess quality of articles using an ordinal scale with seven categories[9]. In ascending order they are: Stub, Start, C, B, Good Article, A, and Featured Article[10]. Each category comes with a set of criteria which an article needs to meet, for instance a C-class article is "substantial, but is still missing important content or contains much irrelevant material." Wikipedia's volunteers then individually assess the articles and apply a category to them, with the exception of "Good Article" and "Featured Article" status, which requires a peer review process [VWM07].

The readers of Wikipedia might have other criteria for assessing quality. For instance it has been shown that readers mainly assess quality through the trustworthiness of an article, which they determine by looking at whether the article uses images and properly references external sources [LS10b]. Other research has utilised Amazon's Mechanical Turk to recruit non-contributors and compare their assessment with that of Wikipedia's contributors. An early study by Kittur and Kraut [KK08] found significant overlap ($r_s = 0.54$) between comparable set of categories. A more recent study [KR16] examined medicine-related articles and found instead that non-contributors were unable to provide fine-grained quality assessment, they were instead only able to group articles into a binary high/low categorisation scheme.

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

[10] There are also categories for other types of content (e.g. lists), and some seldom used categories (e.g. "B+"). We will not concern ourselves with these additional categories and instead focus on the most commonly used seven categories in this thesis.

2.2.2 Quality Assessment in OpenStreetMap

OSM does not have quality assessment categories like Wikipedia does, meaning that OSM contributors do not define specific areas of the map as having a certain amount of quality. Instead, research on OSM suggests that completeness can be determined by looking at map edit patterns over time. As an area moves towards completeness, in other words higher quality, the level of edit activity diminishes [GBR14]. When it comes to how users of the map data OSM provides assess quality, the equivalent to Wikipedia's readers. there appears to be little research available.

2.3 Can We Automate Quality Assessment?

When it comes to OpenStreetMap, there appears to be a scarcity of work studying whether it is possible to automate quality assessment. This may partly be due to the usage of external datasets to determine positional accuracy, as described earlier, where these other datasets are not freely available. Some of the mapping tools that have been built to support OSM contributors do feature some forms of automated error detection[11]. These appear to rely heavily on OSMs tagging system, for instance to detect a road crossing a body of water without the road being tagged "bridge". In summary, there appears to be potential for a contribution to the literature by looking at sociotechnical solutions to quality assessment in OSM.

For Wikipedia, the research literature started out by seeking to understand "how", meaning they looked at how the content was produced and whether it resulted in encyclopedic quality. Having established that it does work, focus moved on to whether technology such as machine learning could be utilised to support the community by

[11] See `http://wiki.openstreetmap.org/wiki/Quality_assurance` for a list of tools

easing the effort required to assess quality, an otherwise completely manual process. We will categorise the Wikipedia research into three main categories, although some of the work spans multiple categories.

2.3.1 Contributor-based Assessment

Some studies of assessment in Wikipedia have focused on understanding how different contributors affect quality, or what we could call a "contributor-based perspective". One approach is to look at *editor reputation*, where an editor's reputation is commonly based on whether they contribute content that lasts [AD07; Hal+09]. It is then possible to predict the quality of an article by looking at the reputation of its contributors and how much content they contributed to it [SY12].

While editor reputation has been found to play a role, many other aspects of quality has also been studied. One study investigated the correlation between number of edits/contributors and quality, finding that higher quality articles have more contributors and more edits [WH07]. At the same time, who these contributors are and what types of work they do affects quality [LR11]. For instance, it has been found that the experience of the creator of an article can determine if it reaches higher levels of quality [SH07]. An alternate approach to experience used a definition of social capital, and similarly found article's created by contributors with higher social capital would end up with higher quality [NGL11].

Contributors in Wikipedia do not act alone, and research has built networks of collaborators to study how properties of these networks relate to quality. Two studies added properties to the edges in a network graph and found that this could be used to predict article quality [Bra+09; WI11]. Wu et al. [WHC12] built a network between contributors and articles and mined edge patterns, which they referred to as *motifs*,

from this network. Using these patterns they then predicted article quality with good success. Recent work studied whether affinity networks provide information that is not already captured by other approaches [KR16], finding that adding data about these networks can significantly improve quality predictions.

2.3.2 Article-based Assessment

Research has also focused on the articles themselves, seeking to understand what properties describe quality and allows software tools to distinguish between varying levels of quality. High-quality articles have a vastly different edit pattern over time than low-quality articles [WP09]. When it comes to the words used in the article, the variety of words, or the "writing style" also has an effect [LS10a; XL11]. However, it also turns out that just looking at the amount of content in the article to a large degree defines its quality, higher level articles have more content [Blu08; Hu+07; WP09].

Some of the research has also used a combination of features. A paper by Stvilia et al. [Stv+05a] mapped features from the information quality literature on to Wikipedia's articles and built a model that combined both properties of the content as well as who authored it. This model showed good performance at distinguishing Featured Articles from a random set of other articles. Dalip. et al. [Has+09] tested a large number of features, aiming to find out which ones were best suited for distinguishing between different classes of quality. They found that, again, the amount of content and the writing style were again the highest performers.

2.3.3 Article Flaw Detection

Where other research typically aims to determine whether an article is of high or low quality, some research has also aimed to understand if we can detect the flaws an article might have. Software tools to help contributors identify and remove flaws could be created, and a flawless article could the be labelled "high quality". The main body of research in this area used one-class classifiers and found that it was able to identify four out of the ten types of flaws they investigated [ASL12]. An international competition on quality flaw detection was held in 2012 [AS12], where two of the entries showed promising performance.

2.4 The Production of Quality Content

We now turn our attention to how quality content is produced in peer production communities. First, we will examine some of the research that has studied how to increase the amount of content produced both in Wikipedia and VGI communities. Then we will examine Wikipedia content production in more detail before we round off with a similar examination of OpenStreetMap.

2.4.1 Increasing the Amount of Contributions

Research has looked at many different ways of increasing the amount of contributions to a peer production community. Soliciting contributions from consumers of the content is one way. Halfaker et al. [HKT13] nudged Wikipedia readers into submitting feedback on article quality, finding that this adds value as long as the system design makes it easy to weed out low-quality contributions. A study of the Cyclopath bike-

mapping community found that naturally occurring feedback on user behaviour could be used to improve suggested bike routes or add annotations to the map [MT14].

Recommender systems can also be used as a tool to match contributors with tasks related to their interest. This is called "Intelligent Task Routing" (ITR) and was first used to request contributions from users in a movie recommendation system [Cos+06]. They found that some personalised strategies were successful, but that one of them was outperformed by the random baseline strategy. ITR has also been studied in the context of Wikipedia in the form of an article recommender called "SuggestBot" [Cos+07]. In that case, three personalised strategies were about four times more successful than a random baseline at eliciting contributions to articles.

In the social psychology literature, we find two examples of interventions aiming to increase contributions. Ling et al. [Lin+05] found that appealing to users' unique capabilities and giving them specific and challenging goals resulted in more contributions. Rashid et al. [Ras+06] did a followup study where they found that displaying the estimated value of a contribution had a positive impact. Further, identifying with the member group and the way a person viewed the member group also had positive effects.

Returning to the Cyclopath VGI community, a study where contributors were requested to do work [PMT10] found that they would do more work than explicitly requested. Secondly, user familiarity with a given area would strongly affect the type of contributions they would make.

The amount of existing content and structure can also affect how contributions are made to a peer production community. Solomon and Walsh [SW12] studied whether seeding content in a wiki would alter the contributions made, finding that those who started from a blank slate would contribute more content. However, the study also

found that participants who worked on an initial structure would follow it, meaning that it is possible to steer user behaviour towards certain contributions.

2.4.2 Producing Quality Content in Wikipedia

Content production in Wikipedia has largely focused on the *amount* of content in the form of the number of articles. So much so that at the opening plenary for Wikimania 2006, the conference for volunteers working on Wikipedia and other Wikimedia Foundation-related projects, Jimmy Wales encouraged the projects that already had a large number of articles to turn their focus towards ensuring content quality[12].

Wikipedia is a collaborative environment, so one area of research looks at how Wikipedia's contributors coordinate their efforts when writing article content. As we discussed earlier, high-quality articles have a large number of contributors and contributions, with intense cooperative behaviour [WH07]. This was further studied by Kittur and Kraut [KK08], finding that quality would increase faster when a concentrated group of contributors worked together. Lastly, Arazy and Nov [AN10] further increased our understanding of this phenomenon by discovering that contributor concentration does not have a direct effect on article quality, it is instead mediated through how they coordinate their efforts. They also reported that article quality was directly effected by a diverse set of contributors provided that at least some of them have a lot of Wikipedia experience, similar to how we previously discussed experienced contributors' effect on quality [SH07; NGL11].

When collaborating it is also worth asking if simultaneous work is more or less effective than asynchronous sequential work. André et al. [AKK14] studied this and

[12] `https://wikimania2006.wikimedia.org/wiki/Opening_Plenary_(transcript)#Quality_initiative_.2833:20.29`

found simultaneous work to be less effective than sequential, but that the effect was mitigated by assigning specific roles to participants. There are very few formal roles in Wikipedia, contributors instead assume different roles depending on the situation [Ara+15; Yan+16]. At the same time, Wikipedia is highly efficient at producing content during breaking news events [KGC13], a situation where a large number of contributions typically occour over a short period of time.

The production of quality content is of course also affected by what specific work contributors do. Wattenberg, Viégas, and Hollenbach [WVH07] found that Wikipedia's contributors choose to do particular tasks some of the time. For instance will they create alphabetical lists of articles that need specific improvements and go through them in sequence, which shows up as distinct patterns when visualised.

Quality Improvement Projects in Wikipedia

Research has also looked specifically at some of the different quality improvement projects that take place in Wikipedia, a topic that we study more closely in Chapter 4. One such type of project is "Collaboration of the Week" (CotW), a focused effort typically lasting one to two weeks. These efforts are organised by WikiProjects, which are volunteer groups of Wikipedia contributors that are interested in a specific topic or task [CRR10; For+12; KK08; KPK09; Mor+13; Mor+14]. Zhu, Kittur, and Kraut [ZKK12] studied CotW and found that their article improvement goals strongly motivated project members to increase the nubmer of contributions they made, and that the effect also spilled over to other article within a WikiProject's topic area.

The Wikipedia Education Program[13] (WEP) is a project where educators and students across the globe work on improving Wikipedia articles as class assignments.

[13] `https://outreach.wikimedia.org/wiki/Education`

It started in 2010 as the Public Policy Initiative[14] (PPI). Lampe et al. [Lam+12] surveyed PPI participants, asking whether the project motivated them to continue contributing to Wikipedia after course completion. Students that reported actively participating and who were aware of Wikipedia's global reach were also more likely to say they would continue contributing.

A project related to the WEP is the Association for Psychological Science's (APS) Wikipedia Initiative[15]. This project was studied by Farzan and Kraut [FK13], comparing participants against a cohort of similar Wikipedia contributors. Farzan and Kraut found that participants added considerably more content, and that the content survived on par with that contributed by subject matter experts with PhDs.

The PPI and WEP has also been studied by the Wikimedia Foundation, although the results have not been published in peer-reviewed venues. Their findings for the PPI reported that the average article improved to an intermediate amount of quality [Rot], while their study of the WEP [WMF] found a smaller increase in quality.

2.4.3 Producing Quality Content in OpenStreetMap

The research literature has revealed less information on how content is produced in OpenStreetMap, at least compared to the large number of Wikipedia studies. Two specific aspects of production in OpenStreetMap has received attention: mapping parties and the Humanitarian OpenStreetMap Team (HOT).

Mapping parties are local workshops that are designed to introduce new users and contributors to the community by gathering data and updating the map [HW08]. Studying mapping parties in London, Hristova et al. [Hri+13] found that participants

[14] `https://outreach.wikimedia.org/wiki/Public_Policy_Initiative`
[15] `http://www.psychologicalscience.org/index.php/members/aps-wikipedia-initiative`

were very active contributors, of whom many joined OSM during its initial phase. They also tend to be long-time contributors to OSM. Lastly, they also concluded that these mapping parties were unsuccessful in retaining newcomers to the community.

The Humanitarian OpenStreetMap Team[16] is a group of volunteers who come together to produce VGI content in areas where humanitarian aid is needed. Two key instances where HOT participated are the Haitian earthquake in 2010, and the 2015 earthquake in Nepal. Oliver et al. [Oli+14] surveyed 252 disaster response volunteers, of which 118 were OSM volunteers, in order to learn more about the motivation of these volunteers. Their results indicated that the four main motivational factors were: personal satisfaction, altruism, increased understanding about a given disaster, and gaining and improvising geospatial knowledge.

2.5 Consumption of Quality Content

Wikipedia is one of the worlds most popular websites[17], but what types of content is consumed there? Research into reader habits have found that it spans a diverse range of topics. It might be something casual, such as the relationship status of their favourite celebrity [Spo07]. Frequently, Wikipedia is also used to look up more serious information such as facts about a disease with which someone has recently been diagnosed [Sch+06].

When it comes to OSM, less is known about how it is used. Whereas Wikipedia is largely a centralised source of information, people go to Wikipedia's website to read the information, OSM is to a much larger extent decentralised. OSM content is reused by other websites and applications, many of them very popular ones such

[16] `https://hotosm.org`

[17] `http://www.alexa.com/topsites` lists it as number 5 as of October 13, 2016.

as MapBox[18], Apple Maps, and Craiglist[19]. We can therefore draw some conclusions about *how* OSM content is consumed, but know little about the specifics.

2.6 Misalignment Between Production and Consumption

Unlike work allocation processes in traditional content production organisations, peer production communities generally have no central authority that directs work towards topics that are in high demand by consumers (e.g. Wikipedia readers). Contributors to peer production communities generally do work that they perceive as "fun" [Nov07], work that is simultaneously neither too difficult nor too simple [LW05], or that aligns with their altruistic interests [Ard08; BH13]. These motivational factors may or may not lead to the production of high-quality content on topics of most interest to consumers.

There have been several studies of the similarities and differences between contributors and consumers in peer production communities. West et al. [WWC12] used browser toolbar data to show that contributors are more active users of various Internet services (e.g. news sites and YouTube) compared to consumers. For medical topics, Wikipedia has been shown to be a very popular information resource [Hei+11], but one that does not necessarily supply information "clinically important to patient safety and care" [Cla+08].

In addition to these differences in the interests of contributors and consumers, several studies have also looked at how the production of content is affected by bias in the contributor population. Contributors to Wikipedia and OSM are largely male.

[18] `https://www.mapbox.com`

[19] `http://arstechnica.com/business/2012/08/craigslist-is-on-board-openstreetmap-continues-soaring-to-new-heights/`

This has been shown to for instance affect the amount of content about movies in Wikipedia [Lam+11], where those with a mainly male audience will have disproportionally more content. Research on biographies comparing Wikipedia to Encyclopædia Britannica found that the former had better coverage and longer articles but was also more likely to be missing articles about women [RR11], and Wikipedia's related problems with categorisation of novelists attracted media attention[20]. Bias in OSM has also been studied, for instance it was found that there was a much larger variety of tags available to describes places related to prostitution than those related to child care [Ste13]. Socioeconomic status also plays a role in OSM, where lower-income areas in England were found to have considerably lower coverage and content quality [Hak10].

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 [Rei11] 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 [HS14b] 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 [Gor14] 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.

[20] `http://www.nytimes.com/2013/04/28/opinion/sunday/wikipedias-sexism-toward-female-novelists.html`

Work by Keegan et al. [KG10] found that higher quality articles were more likely to attract contributions. Recent work by Kane and Ransbotham [KR16] further refined the connection between consumption, contributions, and quality, finding that consumption leads to contributions, but that as an article gains quality it will attract fewer contributions, there is less work to do.

This thesis is specifically interested in the relationship between production and consumption of quality content, looking at whether high-quality content is produced where there is a demand for it. Lehmann et al. [Leh+14] 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 [Gor11] identified a similar mismatch between popularity and quality. It is this mismatch we study in more detail in Chapter 5.

Chapter 3

Predicting Content Quality Using Actionable Features

In this chapter we extend the state of the art in using machine learning to predict Wikipedia article quality. As we saw in our "background" chapter, previous research has found that Wikipedia's notion of article quality correlates well with traditional notions of encyclopedic quality [Stv+08a]. It has also been found to correlate reasonably well with non-Wikipedians idea of what quality articles should be like [KK08]. Lastly, we referred to the extensive body of research on using machine learning for automatic assessment of article quality.

We saw an opportunity to contribute to the research literature by combining quality *prediction* and quality *improvement* by focusing on *actionable features*, those features that contributors can easily work with. By focusing on those features, the ultimate goal is to build tools that are able to not only accurately assess the quality of the content produced, but that can also provide contributors with specific suggestions on how to further improve it. For instance, it is reasonably straightforward to inform a contributor that an article needs more references to sources and for them to act on it. If we instead seek a certain distribution of contributor experience, or use a measure of the longevity of the contributors' contributions to other articles, it might be non-trivial to determine how that can be meaningfully changed, and most likely difficult for a single contributor to go about changing it.

In addition to being able to suggest specific work types to contributors, our re-

search also aimed to be able to do fine-grained classification in an efficient way. Many of the larger Wikipedia editions use a six or seven level scale, in the English edition there are seven levels[1] and from lowest to highest degree of quality they are: Stub, Start, C, B, Good Article, A, Featured Article. Previous literature in this area has often reduced the problem of predicting Wikipedia article quality to a two-class problem (e.g. Featured Articles versus "everything else"). Our research was motivated in being able to implement it as a part of SuggestBot[2] [Cos+07], a software tool that recommends articles to edit to Wikipedia contributors in seven language editions. We therefore started out with a two-class problem before exploring the feasibility of predicting all seven of the English Wikipedia's quality classes. As we will see, the results of our first study was promising, and in the second study we further improve performance to a level that enables deeper analysis of quality improvement. Lastly, we also noted that many of the features used in previous research are costly to calculate (e.g. editor reputation). Aiming to enable large-scale evaluation of quality, our research preferred using more efficient measurements by focusing on an article revision's text instead of metadata (e.g. "number of editors" or "number of reverts").

The first part of this chapter is based on our paper presented at WikiSym in 2013 [WCR13], where we described the problem and designed the first iteration of our "actionable quality model". In the second part of this chapter, published as an appendix in our 2015 CSCW paper [War+15b], we describe how we made substantial improvements in the accuracy of this model. In addition to these publications, our research has had a significant impact on the community, which we will describe in

[1] `https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment` There are also two quality labels for lists, and some groups might use non-standard quality labels such as "B+", neither of which are in the scope of our research.

[2] `https://en.wikipedia.org/wiki/User:SuggestBot`

more detail at the end of this chapter in section 3.3.

3.1 An Actionable Quality Model

In this section we will first discuss how we chose a machine learner that would enable us to learn more about which features were useful for predicting quality. We then discuss a simpler version of the prediction problem that divides the English Wikipedia's seven assessment classes into a "good enough" set and one that "needs work", followed by how we created our training and test data sets using that definition. Using the chosen machine learner and data sets we then evaluate several different feature sets on our two-class problem, where we end up with our five-feature actionable quality model. We then test several different machine learners to understand if others significantly outperform our initial choice, before finally generalising our classification problem to all seven assessment classes.

3.1.1 Technology selection and data collection

Selecting an appropriate machine learner

Our aim is to gain an understanding of the predictive power of different features when classifying article quality, with particular focus on those that are actionable. We therefore prefer algorithms which allow us to inspect the underlying model directly. Blumenstock used a logistic regression where the regression coefficients are exposed [Blu08]. Stvilia et al. used a decision tree classifier, where the tree can be inspected to learn how specific features are used, to build a fairly complex model with a combination of actionable and non-actionable features [Stv+05a]. We chose to use

a decision tree classifier because of the combination of an exposed model and known good performance from Stvilia et al.

Assessment class selection

A common approach to quality modelling is to classify Featured Articles (FAs) versus other articles (e.g., [Stv+05a; WP09]). However, we are interested in distinguishing broadly between articles that need a lot more attention and articles that are already "pretty good". Instead of predicting FAs versus others, we choose a split that reflects whether the articles are in need of more attention from contributors.

From the description of the assessment classes[3] we learn that both FAs and A-class articles are "complete". It is also clear that Good Articles (GA) have received a lot of attention, due to the peer review process involved in reaching GA status. Thus, we choose to split the article space into two classes: one class of articles not in need of more attention, which we label *GoodEnough*, containing FA, GA and A-class articles, and one class of articles needing more attention, which we label *NeedsWork*, containing B-, C-, Start-, and Stub-class articles. Because we include all classes of articles, and set our split not at the best (FA) or worst (Stub and Start) article classes, but somewhere in the middle, we expect this to be a challenging task.

Data collection

Having chosen a decision tree classifier as our technology and defined our classes as $GoodEnough = FA \cup GA \cup A$ and $NeedsWork = B \cup C \cup Start \cup Stub$, we turn our attention to gathering articles for training and testing the classifier. Decision tree

[3] `https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment#Grades`

classifiers prefer training sets where there are roughly the same number of items in each class, so we set out to build such a dataset.

We gathered our data in the period of 27-29 May 2011. First we found the class with the fewest number of articles, which was A-class articles with 827[4]. Our plan was to select the same number of articles from FA, GA, and A, leading to a total of 2,481 articles in the *GoodEnough* class, then sample another 2,481 evenly distributed across the B, C, Start, and Stub classes to create a balanced dataset of roughly 5,000 articles. However, when we crawled A-class articles using the category "Category:A-class articles", we found only 672 actual articles[5].

In the end, we gathered the 672 A-class articles and 800 each from FAs and GAs for a total of 2,272 *GoodEnough* articles. We then chose 568 articles from each of the remaining four classes, for a total of 2,272 *NeedsWork* articles. These were then split 50/50 into a training set and a test set. Note that these assessments are best guesses; a limitation of this data set is that the quality assessment assigned to articles may not reflect their true assessment class, or the underlying distributions in Wikipedia, because articles change in quality and some articles are not assessed.

3.1.2 Establishing a baseline

We start our exploration of article quality assessment with Stvilia et al.'s early but well-known model as a baseline. This model was chosen because it has known good performance and contains a combination of actionable and less actionable features. There are a total of 18 features in all, some of which are added together to make it a seven-dimensional model as presented below.

[4]This was according to "WP 1.0 Bot", which counts the number of articles in each quality class.

[5]The difference appears to come from "WP 1.0 Bot" using WikiProject listings of article assessment instead of counting articles tagged with this category.

1. **Authority/Reputation** = 0.2∗NumUniqueEditors + 0.2∗NumEdits + 0.1∗Connectivity + 0.3∗NumReverts + 0.2∗NumExternalLinks + 0.1∗NumRegUserEdits + 0.2∗NumAnonEdits
2. **Completeness** = 0.4∗NumBrokenWikilinks + 0.4∗NumWikilinks + 0.2∗ArticleLength
3. **Complexity** = Flesch-Kincaid Readability Score
4. **Informativeness** = 0.6∗InfoNoise - 0.6∗Diversity + 0.3∗NumImages
5. **Consistency** = 0.6∗AdminEditShare + 0.5∗Age
6. **Currency** = Current article age in days
7. **Volatility** = Median revert time in minutes

Connectivity is the number of articles reachable through the editors of a given article. *InfoNoise* is the proportion of text content remaining after removing MediaWiki code and stopwords and stemming all words. *Diversity* is *NumUniqueEditors/NumEdits*. Other definitions can be found in the original paper [Stv+05a].

Note that some features, such as the current article age or revert volatility, are practically impossible to directly change; others, involving the mix of anonymous-to-registered or admin-to-regular edits, are in principle actionable by recruiting new editors (or suppressing current ones) but in practise difficult for individuals to enact; while still others, such as the number of wikilinks[6] or images, might be more directly addressable by individual editors.

[6]Links to other Wikipedia articles.

In order to identify reverts to calculate *Volatility*, we applied the approach of Priedhorsky et al., which uses regular expressions to match edit comments [Pri+07]. Edit comments are a text field used by contributors to describe the changes they have made in a revision. While this approach does not correctly identify all reverts [FVS12], in May 2011 when we collected our dataset more resilient approaches would have required downloading the text of all revisions of each article to calculate hash values[7]. We also used Priedhorsky et al.'s approach to identify anti-vandal work and exclude anti-vandal edits from median revert time, as much vandal fighting is now handled by bots and software-assisted humans [GR10] and therefore does not properly reflect article controversy. Bot edits were identified by making a case-insensitive match of the username associated with the edit having a part that ends with "bot", for example "RamBot" and "MiszaBot III". The advantage of this approach is that it is fast, but it will miss bots that do not follow the common naming convention of bot accounts[8]. Checking if the account is a member of the "bot" user group should catch most or perhaps all of the missed bots that are officially registered with Wikipedia.

The *Connectivity* feature is the cardinality of the set of other articles edited by the editors of a specific article, after excluding bots and anti-vandal reverting editors from the set. At the time we gathered our data it was nontrivial to determine how reverts affect an article's history [ER09; FVS12]; thus we did not attempt to remove reverted editors when looking for connected articles.

Some of the data used in Stvilia et al.'s model is power law distributed, e.g., number of edits and number of editors. The paper did not specify whether they chose to log-transform these features, so we tested the model with both non-trans-

[7] SHA1 hash values for all revisions are now available through Wikipedia's API.

[8] `https://en.wikipedia.org/wiki/WP:BOTACC`

formed and log-transformed. Non-transformed data had higher performance so we report it here.

We test this set of features using the C4.5 decision tree classifier and our training and test datasets described earlier. The overall classification results are listed in the "2005 model" column in Table 3.2. We report the following measures: True Positive Rate (**TPR**) for each class as well as an overall weighted average, which allow us to judge the classifier's ability to correctly predict classes; **Precision** and **Recall**, which are widely used to judge performance when one class is more important; $\mathbf{F}_1$ (or F-measure), which represents a harmonic mean between precision and recall; and the area under the **ROC** (Receiver Operating Characteristic) curve, which is commonly used to judge relative performance between classifiers for the trade-off between true positive and false positive rates.

Because we defined our "GoodEnough" class to include Featured Articles, Good Articles, and A-class articles, while Stvilia et al. classified Featured Articles versus Random with Stub-class articles removed, we expect to see somewhat lower performance compared to theirs. As we see from Table 3.2, overall prediction performance comes in at 76.1%, while in their work they successfully classified over 90% of their articles.

3.1.3 New potential features

In addition to the features used in Stvilia et al.'s model, we are interested in introducing new features, including actionable features that suggest specific improvements and features that have become more common in Wikipedia since 2005. For instance we know that Wikipedia articles require sources for claims[9]. Previous research has

[9] `https://en.wikipedia.org/wiki/Wikipedia:Verifiability`

Table 3.1: Feature list of all four models

2005 model	**Full model**	**Hybrid model**	**Actionable model**
Authority/Reputation	Authority/Reputation	Authority/Reputation	
Completeness	Completeness*	Completeness*	Completeness*
Complexity	Complexity		
Informativeness	Informativeness*	Informativeness*	Informativeness*
Consistency	Consistency		
Currency	Currency		
Volatility	Volatility		
	ArticleLength	ArticleLength	ArticleLength
	Diversity	Diversity	
	Tenure	Tenure	
	NumHeadings	NumHeadings	NumHeadings
	NumRefs/Length	NumRefs/Length	NumRefs/Length
	NumReferences		
	NumHeadings/Length		
	NumImages/Length		
	NumWikilinks/Length		
	HasInfobox		

Features marked * are modified as described in section 3.1.4.

Table 3.2: Classification results of all four models

	2005	**Full**	**Hybrid**	**Actionable**
GoodEnough TPR	0.839	0.849	0.899	0.898
NeedsWork TPR	0.683	0.868	0.854	0.833
Overall TPR	0.761	0.859	0.876	0.865
Precision	0.767	0.859	0.877	0.867
Recall	0.761	0.859	0.876	0.865
F_1	0.760	0.859	0.876	0.865
ROC AUC	0.792	0.863	0.884	0.883

TPR = True Positive Rate. The classification performance measures are described in section 3.1.2.

shown that when readers judge the trustworthiness of Wikipedia articles, references to sources play an important part [LS10b]. To capture the extent to which claims in the article are sourced we propose *NumReferences*, a measure of the number of citations, by counting the number of <ref>-tags which are used for footnote citations.

We also add a feature to capture the extent to which an article has been organised into sections (*NumHeadings*). Appropriate article structure and organisation is a common theme in the article assessment criteria and many Wikipedia articles have sections such as "See also" for linking to other relevant Wikipedia articles and "References" for listing the article's sources. Research has suggested that organising content in a wiki can help structure future contributions [SW12], meaning this feature can both reflect current article quality and improve future contributions.

Some of these added features might be good metrics by themselves, but it could also be that there is a relationship to the length of the article. For instance the raw number of cited claims is likely to be lower for a short article, but it might be that relative to its length it has an appropriate number of citations. We therefore add features to capture the relationship with article length, as in de la Calzada and Dekhtyar [DD10]:

1. NumReferences/ArticleLength
2. NumImages/ArticleLength
3. NumWikilinks/ArticleLength
4. NumHeadings/ArticleLength

Because many good articles have an infobox, we add a binary (0/1) categorical feature for that. Lastly we add features for the number of templates and categories an

article has (*NumTemplates* and *NumCategories*). High quality articles are likely to use templates for formatting of content and following Wikipedia conventions, whereas low quality articles might lack these. Similarly we suspect that high quality articles will be assigned to a number of categories, whereas low quality articles may be less likely to be categorised well.

We also propose an editor tenure metric to replace Stvilia et al.'s administrator edit share because the proportion of administrators to other contributors on English Wikipedia is now much lower [Stv+08b]. While this is not an actionable feature, we are interested in understanding its effect on performance as previous research suggests that edits by experienced editors positively affect article quality [NGL11; SH07]. We want to capture a notion of total editor experience accumulated across all edits to an article, in both age (time since they registered) and number of edits. This leads us to log-transform the edit count, because it is known to be power-law distributed, and then linearly combine them for each edit a user makes to a specific article as follows:

$$tenuretime(t,i) = t - t_{reg,i} \tag{3.1}$$

$$tenureedits(t,i) = log(nedits_{i,now} * t/(t_{now} - t_{reg,i})) \tag{3.2}$$

$$tenure(t,i) = tenuretime(t,i) + tenureedits(t,i) \tag{3.3}$$

In the formulae above, t is the time user i edited the article, $t_{reg,i}$ is the time user i registered their account, while $nedits_{i,now}$ is user i's edit count as of when the calculation was done (t_{now}). We then sum $tenure(t,i)$ for all registered non-reverting, non-bot editors of a given article to get our proposed metric *Tenure*.

3.1.4 Building new models

We now turn our attention to investigating how different feature sets and machine learning technologies affect classification performance. We first modify two of Stvilia et al.'s features and add our proposed ones to create and evaluate a large model with 17 features. Then we describe how we iteratively tested and removed specific features to create a hybrid model with eight features, ending up with a model that only contains five actionable features. An overview of the different features used in the four models we developed is shown in Table 3.1, and their performance is listed in Table 3.2. Lastly we evaluate the performance of other types of classifiers such as neural networks and random forests.

Full model

We start by modifying the seven dimensions so that the features become more clearly separated between the actionable and non-actionable, then add our proposed features. Separating *ArticleLength* from *Completeness* leaves the latter a measure of the number of wikilinks (0.4∗NumBrokenWikilinks + 0.4∗NumWikilinks). Removing *Diversity* from *Informativeness* leaves the "Diversity" feature a measure of textual noise and number of images (0.6∗InfoNoise + 0.3∗NumImages). The resulting model contains 17 dimensions, as previously defined unless noted, and will be referred to as the "full model". Table 3.3 lists all features ranked by their overall gain ratio as calculated by WEKA using cross-validation on the training set. Gain ratio is the measure used in a C4.5 decision tree to determine which feature to use when splitting between classes [TSK06].

Training WEKA's C4.5 decision tree classifier using these 17 features results in

Table 3.3: Overall gain ratio evaluation for all 17 features

Rank	Feature	Gain ratio	Actionable
1	NumReferences/ArticleLength	0.205 ± 0.018	Yes
2	NumReferences	0.190 ± 0.012	Yes
3	ArticleLength	0.159 ± 0.015	Yes
4	Diversity	0.135 ± 0.006	No
5	Tenure	0.123 ± 0.005	No
6	NumHeadings	0.114 ± 0.007	Yes
7	NumHeadings/ArticleLength	0.105 ± 0.004	Yes
8	Informativeness*	0.101 ± 0.005	Yes
9	Completeness*	0.101 ± 0.003	Yes
10	NumImages/ArticleLength	0.099 ± 0.002	Yes
11	NumWikilinks/ArticleLength	0.091 ± 0.003	Yes
12	Authority/Reputation	0.081 ± 0.003	No
13	Consistency	0.055 ± 0.002	No
14	Volatility	0.043 ± 0.002	No
15	Currency	0.025 ± 0.002	No
16	HasInfobox	0.018 ± 0.002	Yes
17	Complexity	0.016 ± 0.002	Yes

Features marked * are modified as described in section 3.1.4.

Table 3.4: Classification results for all classifiers on 2-class problem

	True Positive Rate						
Classifier	**GE**	**NW**	**Overall**	**Prec.**	**Rec.**	**F_1**	**ROC**
RandomForest	0.889	0.856	0.872	0.873	0.872	0.872	0.939
C4.5	0.898	0.833	0.865	0.867	0.865	0.865	0.883
MultiLayerPerceptron	0.889	0.824	0.857	0.858	0.857	0.856	0.904
JRip	0.882	0.800	0.841	0.843	0.841	0.841	0.871
LibSVM	0.886	0.662	0.774	0.789	0.774	0.771	0.774
SimpleLogistic	0.824	0.708	0.766	0.769	0.766	0.765	0.843

All classifiers use the actionable model with five features, and are ranked by their F_1-score. *GE* and *NW* are True Positive Rate for the *GoodEnough* and *NeedsWork* class, respectively.

a tree of size 153 with 77 leaves. It correctly classifies 1,951 articles, or 85.9%, as shown in the "Full model" column in Table 3.2. This large increase in performance comes mainly from the *NeedsWork* class, which the seven feature model only correctly classified 68.3% of the time, while the full model correctly predicted 86.8% of the articles in that class.

Hybrid model

One of the reasons for choosing to use a decision tree was the ability to inspect the tree to understand how the features were used and whether some would be good candidates for removal. Inspecting the tree trained on the full model, we found that one feature was never used (*NumWikilinks/ArticleLength*) while some features (e.g., *Authority*, *Complexity*, and *Currency*) were mainly used in deep branches to distinguish between a small number of articles. We saw these features as likely candidates for removal to prevent over-fitting without a large impact on performance.

We also iteratively added and evaluated specific features or combinations of these as an alternative to a large feature set that leaves the classifier to figure out which ones are useful. The complete process is omitted for brevity, consisting of testing more than 30 models with various combinations of features. We kept features that created fairly simple trees, indicating they had good information gain, while performing on par with classification performance using the full feature set. *Complexity*, *Volatility*, and *Currency* were removed without impacting performance. The *Consistency* feature was dropped in favour of *Tenure*. The result is our "hybrid model" with eight features, combining actionable and non-actionable ones: *Authority/Reputation*, *Completeness*, *Informativeness*, *Diversity*, *Tenure*, *ArticleLength*, *NumHeadings*, and *NumReferences/ArticleLength*.

The "Hybrid model" column in Table 3.2 shows the overall performance of this hybrid model being slightly better than the one trained on the full list of features. It is correctly identifying more *GoodEnough* articles (89.9% compared to 84.9%) at the cost of misclassifying some additional *NeedsWork* articles (14.6% compared to 13.2%).

Actionable model

Because of our interest in actionable features, we next looked at the impact of removing all remaining non-actionable features from the model, resulting in our "actionable model" which contains only five dimensions[10]:

1. Completeness = 0.4*NumBrokenWikilinks + 0.4*NumWikilinks
2. Informativeness = 0.6*InfoNoise + 0.3*NumImages
3. NumHeadings
4. ArticleLength
5. NumReferences/ArticleLength

The "Actionable model" column in Table 3.2 shows that this model has comparable performance to the full and hybrid models. Our actionable model incorrectly regards a slightly larger proportion of *NeedsWork* articles as high quality. This could be due to a lag in the assessment process, as discussed earlier: it may be that articles edited by high-profile editors are more likely to be reassessed. It could also be that those articles are more likely to be of high quality, as we argued when defining our *Tenure*

[10]For the definition of *InfoNoise*, see section 3.1.2.

metric. Since our five-feature model does not contain features for editor experience, it will instead regard articles as high quality based purely on content features.

3.1.5 Alternative classifiers

The decision tree was useful for exploring and selecting features, and though it provided good performance, other classifiers might outperform it. We used some of WEKA's other available classifiers, including libSVM (Support Vector Machine), MultilayerPerceptron (neural network), JRip (rule-based), SimpleLogistic (logistic regression), and RandomForest with 100 trees. All classifiers used WEKA's default options, with the exception of the random forest, which was tested with sizes from 10 (the default) up to 1000. We report results based on a random forest size of 100 as it had the best performance.

We tested both the full model with 17 features and the 5-feature actionable model[11]. The results for both feature sets were comparable, with only minor improvements in both cases, so we report results for the actionable model in Table 3.4. These results indicate that we might need a different set of features to tease out the benefit of specific classifiers, something which future research could look into.

3.1.6 Predicting all assessment classes

Our investigation of actionable features is motivated by our interest in using those features to help contributors increase the quality of articles. Being able to distinguish between all seven assessment classes could support other quality-related use cases. We might be able to identify articles that need reassessment (e.g., candidates to become

[11]The SVM classifier was not run on the large feature set as the high dimensionality leads to poor performance.

Featured Article), allow users to focus on particular quality levels (e.g., avoiding Stub-class articles or looking for articles near the borderline of quality classes), or highlight ways the classes differ on specific features. Distinguishing between all seven classes has also received relatively little research attention, despite its interestingness as a problem. While the difference between a Featured Article and a Stub-class article may be large enough to make it straightforward to differentiate between them, the boundaries between some of the other classes (e.g., between C-class and B-class) are likely to be less well-defined because there are smaller differences in the assessment criteria and because of errors and lag in assessment.

In these evaluations we reuse our existing training and test datasets, but do not collapse them into two classes. We again evaluate both the full model with 17 features and the actionable model with five features; as before, the results are comparable. We also tested all of the classifiers described in the last section, and again the random forest classifier was the highest-performing classifier. Thus, below we report only on the results for the random forest classifier using the five-feature model. We then discuss how the results differ depending on the classifier and feature set.

Table 3.5 shows the performance of a random forest classifier with 100 trees using the actionable model with five features to classify all seven classes. In this table we also report the false positive rate (**FPR**), which is the proportion of other articles predicted to belong to a given class and allows us to judge the confusion between classes. Overall the classifier only correctly classifies 42.5% of the articles, showing that this is a very difficult classification problem. Some of the classes are easier to predict than others, with performance on Featured Article (FA) and Stub-class of 60.3% and 57.7%, respectively. As we speculated above, results are worst in the middle for A-, B-, and C-class articles.

Table 3.5: Classification results for all seven assessment classes

Class	**TPR**	**FPR**	**Precision**	**Recall**	**F_1**	**ROC**
FA	0.603	0.165	0.439	0.603	0.508	0.857
A	0.289	0.079	0.388	0.289	0.331	0.733
GA	0.433	0.126	0.424	0.433	0.428	0.806
B	0.327	0.096	0.327	0.327	0.327	0.764
C	0.292	0.102	0.290	0.292	0.291	0.772
Start	0.405	0.088	0.398	0.405	0.401	0.825
Stub	0.577	0.021	0.796	0.577	0.669	0.934
Overall	0.425	0.101	0.436	0.425	0.425	0.813

These results stem from using the actionable model with five features and a Random Forest classifier with 100 trees. TPR = True Positive Rate. FPR = False Positive Rate.

Table 3.6: Confusion matrix for classification of all seven assessment classes

	FA	A	GA	B	C	Start	Stub	N
FA	**241**	62	83	11	2	1	0	400
A	123	**97**	63	19	18	12	4	336
GA	128	58	**173**	21	17	3	0	400
B	29	18	40	**93**	74	27	3	284
C	22	9	37	72	**83**	51	10	284
Start	6	6	11	50	71	**115**	25	284
Stub	0	0	1	18	21	80	**164**	284
N	549	408	250	284	286	289	206	2,272

Results from using the actionable model with five features and a random forest classifier with 100 trees. Rows show correct class, columns show predicted class. The highlighted diagonal shows correctly classified articles. Rightmost column and bottom row shows total number of articles per class.
Abbreviations: FA = Featured Article. GA = Good Article.

The full confusion matrix is shown in Table 3.6. Two important patterns emerge from this matrix. First is that there is a lot of confusion between FA, GA, and A-class articles. Both FA and A-class articles are defined as "complete", thus they should mostly differ by what comes out of the FA review process. Our model does not appear to capture that difference, with 123 of the 400 A-class articles (30.1%) predicted to be Featured Articles.

The second pattern is that the classifier is pretty good at getting within one class, and tends to err on the high side. If we allow the classifier to be off by one class[12], it correctly identifies 1,747 articles, or 76.9%. This still might be useful for human-in-the-loop tasks such as reviewing quality assessments, but probably does not perform well enough for automatic tasks such as filtering articles out of suggestion lists based on quality class.

If we relax the requirement that features be actionable, and test the full 17-feature model, we see a gain in overall performance from 42.5% to 48.3%. FA and GA are the classes with large gains, improving their true positive rate to 74% and 57%, respectively. The other classes see little or no improvement, indicating that distinguishing between the remaining classes requires other types of features. This suggests that use cases which do not need actionability, such as assessment, would benefit from using a richer feature set.

When it comes to performance differences between different types of classifiers, we found that some perform very well on certain classes. The neural network and rule-based classifiers performed well on the Featured Article class, but the latter struggles with C-class articles. We see this as a good opportunity for future research to look at ensemble methods to exploit the advantages of some of the classifiers when it comes

[12]Where FA=FA or A; A=FA, A, or B; Stub=Stub or Start.

to predicting specific assessment classes.

This completes our initial exploration into building an actionable quality model for Wikipedia. We have found that a simple set of five numeric features[13] provides good performance for assessing Wikipedia article quality using a decision tree or random forest classifier. Our initial problem divided English Wikipedia's seven assessment classes into two classes depending on whether the articles appeared to need more attention or not, but we also saw promising performance on predicting all seven classes, a problem that we tackle in the following section.

3.2 Improving the Actionable Quality Model

In this section we describe how we built our classifier used to study quality improvement projects in our paper published at the 2015 CSCW conference [War+15b]. We extend the work described in the previous section in several ways. First of all, we use a Random Forest classifier as our chosen technology since it showed promising performance in our previous work. Our data gathering process is also improved by inspecting the history of assessment ratings for each article to find the right version of the article to include, a process that is explained in more detail below. In addition, we improve our previous work in four specific ways:

1. A much larger dataset (N=29,828), which requires us to address the class imbalance problem imposed by Wikipedia's low number of A-class articles.

2. A larger set of quality features extracted from each article.

[13]The specific features are listed at the end of section 3.1.4.

3. Each feature is tested six times using 10-fold cross-validation to determine how each feature most strongly relates to article quality (raw metric, log-transformed, and four variants of proportions relative to article length).

4. Classifier parameter tuning (again using 10-fold cross validation) to determine forest size, the number of features to use in each tree split, and terminating node size.

There is no gold-standard dataset on which to train a classifier for this task. To gather a suitable set of candidate articles we copied the behaviour of WP 1.0 Bot[14], the software robot that gathers statistics on Wikipedia article assessments. Using Wikipedia's category system to find articles in a specific assessment class we collected 29,828 article assessments, 5,000 from each class with two exceptions: the Featured Article (FA) class had 4,062 articles at that time, and we only found 766 A-class articles. Official statistics listed 1,279 A-class articles and the discrepancy is likely due to duplicates.

The low number of A-class articles creates what is known as a *class imbalance problem* [JS02]. Random Forest classifiers require reasonably balanced classes, so without remedial action, this would result in poor classifier performance on A-class articles. Typical approaches are oversampling the smaller class, or undersampling the larger classes. We tested both of these approaches and found that they led to lower classifier performance. The 766 articles accounted for only 0.018% of the total number of articles in the English Wikipedia at that time, so, statistically speaking, this article class simply is not used in the encyclopedia. Given the low usage of this class, the probability of an article in our datasets belonging to it is very low,

[14] `https://en.wikipedia.org/wiki/User:WP_1.0_bot`

which means that removing it does not significantly impact our study. Therefore, we decided to ignore A-class articles altogether, and we confirmed this significantly increased classifier performance.

WikiProjects "claim" – and thus assess – articles, and multiple projects can claim the same article (e.g., the Barack Obama article is claimed by 14 projects). How do we select an assessment class for an article if different projects disagree on its assessment? We looked at two methods – (1) choose the **highest** class, (2) choose the **majority** class – and found that these two methods disagreed on only 150 out of 29,828 articles (Cohen's Kappa = 0.967 with two raters, p-value $\ll 0.001$). Therefore, we chose to use the highest assessment class as an article's correct class.

For each article in our training dataset, we went through the article's assessment history to find the point in time where it first belonged to a given class. If that revision is not available, for instance revisions sometimes get deleted due to copyright issues, we used the first available more recent revision. Assessments are mainly done by volunteers, which means that there's potentially a delay between an article reaching a certain quality level and its rating being updated to reflect that. By examining the assessment history we ensure that we train our classifier on a version of the article that is either the same as or as close to the rated version as possible. We then retrieved the revision content (text and wiki markup) and extracted the following 11 features:

1. article length in bytes (log-transformed)
2. number of references (log-transformed)
3. number of links to other articles (log-transformed)
4. number of citation templates

Table 3.7: Quality Improvement Project classifier confusion matrix

	FA	GA	B	C	Start	Stub	**N**
FA	546	167	47	9	1	0	770
GA	252	655	54	83	5	0	1049
B	81	151	374	261	129	11	1007
C	25	128	201	471	189	18	1032
Start	1	12	71	201	600	138	1024
Stub	0	0	0	14	166	818	998

Rows are true (assessed) class, columns are predicted class. Last column (**N**) is the total number of articles in each class.

5. number of non-citation templates (log-transformed)
6. number of categories linked in the article text
7. number of images / article length
8. information noise score (as defined by Stvilia et al. [Stv+05a])
9. has an infobox template (binary variable)
10. number of level 2 section headings
11. number of level 3+ section headings

In order to verify classifier performance on this dataset, we chose to split the dataset using random selection to get a training dataset (80%) and a test dataset (20%). Using 10-fold cross-validation on the training set we validated our features and identified optimal classifier parameters. A forest with 501 trees and terminating node size 8 showed the best performance. Training a classifier on the entire training

Table 3.8: Quality Improvement Project classifier prediction error

Distance	**CotW**		**WEP**		**WikiCup**	
	N	%	N	%	N	%
5					1	0.1
4			2	0.8		
3	2	3.5	7	2.7	26	2.4
2	8	14.0	38	14.8	40	3.7
1	12	21.1	104	40.5	223	20.8
0	27	47.4	96	37.4	699	65.1
-1	7	12.3	7	2.7	49	4.6
-2	1	1.8	3	1.2	31	2.9
-3					4	0.4
Total	57	100.0	257	100.0	1,073	100.0

Prediction errors by distance between reassessed and predicted class for the Collaboration of the Week (CotW), Wikipedia Education Program (WEP), and Wikicup. Positive distance means the prediction was a higher quality class. "N" is number of articles, proportions are measured within each improvement project.

dataset and validating its performance on the test set results in the confusion matrix shown in Table 3.7.

The difference in number of articles per class in Table 3.7 is due to fewer Featured Articles (FA) and the random selection. We see that the overall error rate is 41.08%. Similarly as in our previous work, the classifier is often off by one class. If we allow one class leeway the error rate drops to 10.5%. The classifier also often errs on the high side, for instance more Start-class articles are predicted as C than Stub.

While the performance of the classifier on the test dataset is promising, we also wanted to verify its performance on articles that were part of the quality improvement projects we study in chapter 4 as that would enable us to understand its performance on data specifically associated with the goals of that research. Three datasets of

Table 3.9: Comparison between OLR models for reassessments and predictions

Dataset	**Reassessments**	**Predictions**
CotW	N collaborators negatively associated with quality. Significant ($p < 0.01$).	N collaborators negatively associated with quality. Significant ($p < 0.01$).
WEP	Not statistically significant.	Not statistically significant.
WikiCup	N collaborators negatively associated with quality. New articles positively associated with quality. Significant ($p < 0.001$).	N collaborators negatively associated with quality. New articles positively associated with quality. Significant ($p < 0.001$)

Results from building Ordinal Logistic Models on each dataset based on reassessments or classifier predictions of the quality at time of reassessment. Project abbreviations: CotW: Collaboration of the Week; WEP: Wikipedia Education Program.

articles that were reassessed after a quality improvement project's completion were gathered, one each from each of the projects we study in Chapter 4: the Collaboration of the Week (518 articles), the Wikipedia Education Program (987 articles), and the WikiCup (1,617 articles).

Many of these reassessments suffer from the delay between quality improvement and the subsequent rating update we described earlier. For instance in the CotW dataset the median time to reassessment is 157.6 days. In the intervening time the article may have gone through substantial changes. We therefore restricted these datasets to reassessments that occurred within 10 edits, and where the article has changed by less than 100 bytes. When checking some of our prediction errors described below, we also confirmed that this edit/size limitation led to articles only going through minor changes, e.g. copy edits. After applying this limitation we were left with 57 CotW articles, 257 WEP articles, and 1,073 WikiCup articles.

For each article, we then predicted their quality class as described earlier in order

to enable the comparison of predictions against human assessments post quality improvement. As we were interested in learning specifically to what extent the classifier makes prediction errors, and when it does, how severe these errors are, we chose to measure the error as the distance between predicted and assessed class along the ordinal scale listed in Table 4.3 (e.g. a B-class article predicted to be Start has an error of -2). We then summed these errors across all classes. The distribution of prediction errors for each reassessment dataset is shown in Table 3.8.

For the WikiCup, the classifier shows stronger performance than on our test set, while the performance is less for the other two. The large proportion of one-class errors in the WEP dataset led us to investigate further, finding that the majority of the errors come from articles reassessed as C-class (23 articles) and Start-class (64 articles). A random sample of 23 Start-class articles and all 23 C-class articles were selected and verified that they had all gone through only minor changes (e.g. link fixes or minor copy-editing). The English Wikipedia's criteria for Start-class assessment states in part that the article "most notably, lacks adequate reliable sources." Inspection of the Start-class articles by an expert Wikipedia contributor indicated that the vast majority (20 articles) appeared to have several, if not many, reliable sources, suggesting that this subset of articles were not correctly reassessed, an issue that is also discussed in our "Limitations" section and that opens to future work.

The classifier's predictions are strongly correlated with Wikipedia's own article assessments, more so than using a crowdsourcing approach. This is the case across all four evaluation datasets: the test set ($r_s = 0.86$), CotW ($r_s = 0.57$), WEP ($r_s = 0.58$), and the WikiCup ($r_s = 0.82$). In all these cases we have a higher correlation than what was reported when crowdsourcing was used ($r_s = 0.54$) [KK08].

Using the post-improvement reassessments to perform the same analysis that

forms the basis for the results described in the next chapter leads to the exact same conclusions. To determine this we used the same modelling approach with Ordinal Logistic Regressions for each of the three reassessment datasets, one each using the reassessment rating and the predicted quality rating as the dependent variable. This set of models can then be checked for agreement and the results are listed in Table 3.9.

Aside from the lack of statistical significance for WEP, we see in Table 3.9 that in all cases the pair of models agree with each other. When significant, number of contributors is negatively associated with post-improvement quality, and creating a new artefact is positively associated with quality. Based on the classifier's performance as established in this section and its agreement with post-improvement manual assessments, we therefore concluded that our classifier-based results reported in the next chapter hold when we are analysing the entire datasets, which also allows us to gain statistical significance for WEP. That concludes our research into building an actionable quality model for Wikipedia, and we round off this chapter by describing the usage and impact this research has had.

3.3 Usage and Impact

The research described in this chapter has had a reasonable impact on the research community in the time since publication of our first paper at WikiSym/OpenSym in 2013. Google Scholar reports 28 citations to that paper as of November 4, 2016, of which only one is our own. Our collaboration with Aaron Halfaker at the Wikimedia Foundation since our CSCW paper was published in 2015 has led to further improvements as described below, and that classifier has been referred to as "the state-of-the-art" [DI16] in peer-reviewed research. When it comes to the Wikipedia

community, the impact has been significant:

1. The classifier has been used to deliver quality predictions to Wikipedia contributors who use SuggestBot [Cos+07; Yua+09] since 2011.

2. A Python library called Wiki-Class[15] has been developed in collaboration with Aaron Halfaker at the Wikimedia Foundation. During this work the classifier performance has been further improved and currently reports 61.9% accuracy[16], an improvement of almost three percentage points.

3. The Wiki-Class library has been incorporated into the Objective Revision Evaluation Service[17] (ORES), thus allowing anyone access to quality predictions of Wikipedia articles through an API. While our work focused on the English Wikipedia, the ORES team has also trained prediction models for the French and Russian editions.

4. The datasets used for training and testing our classifiers have been published on figshare[18].

[15] https://github.com/wiki-ai/wikiclass

[16] https://meta.wikimedia.org/wiki/Objective_Revision_Evaluation_Service/wp10#enwiki

[17] https://meta.wikimedia.org/wiki/Objective_Revision_Evaluation_Service

[18] http://figshare.com/articles/English_Wikipedia_Quality_Asssessment_Dataset/1375406

Chapter 4

Understanding Quality Improvement Projects

How do peer production communities like Wikipedia work to improve the quality of their content? In this chapter we study five different quality improvement projects in the English Wikipedia to identify factors and mechanisms associated with successful quality improvement. Our findings can help inform future projects and tools to support them in order to ensure they are effective at positively improving content quality.

For us to be able to understand the process of quality improvement, we initiate the process of developing a coherent framework to describe, analyse, and evaluate quality improvement projects in peer production communities. Comparisons between studies of quality improvements projects has been hindered by the usage of different evaluation methods, as we described in chapter 2. Our work uses a single evaluation method across all projects, some of which are new and some of which have been previously studied, thus bringing them all underneath the same umbrella.

4.1 Unified Descriptive Framework

In order to make comparisons across a diverse set of quality improvement projects, it is first necessary to identify a unified descriptive framework in which to understand these projects. We considered several candidates from the research literature, be-

fore settling on Preece's "Online Communities: Designing Usability and Supporting Sociability" [Pre00]. Preece divides the social side of online communities into three components: **People**, **Purpose**, and **Policies**. We use each component as a major theme in our analysis and further explore the components as follows:

People: What *recruitment* method is used to find project participants, and is the work done by *individuals* or *groups*? Recruitment can either be *internal* – participants are already members of the community; *external* – participants are recruited from outside the community; or the project is open to anyone at all.

Purpose: The primary purpose for all the projects we examined was to improve the quality of Wikipedia. We are more interested in dimensions on which projects *differ* and our specific analyses will thus focus on a project's secondary purpose, e.g. that students in the Education Program achieve academic course credit.

Policies: These comprise the governing structure for a project. Since Wikipedia itself has many policies and guidelines that influence all the projects we study, to avoid confusion, we use the term **structure** to describe the governing rules for individual projects.

4.2 The Quality Improvement Projects Studied

We sought a diverse set of improvement projects for our study. The five projects we study, and how they fit into our descriptive framework, are listed in Table 4.1. Below, we provide additional details:

Table 4.1: Characteristics of the studied quality improvement projects

		CotW	**WikiCup**	**WEP**	**CP**	**TAFI**
People	*Recruitment:*	Internal	Internal	External	Anyone	Anyone
	Individual or group work:	Group	Individual	Both	Individual	Group
Purpose	*Purpose:*	Group achievement	Scoring points, having fun	Course credit	Improve articles	Improve articles
Structure	*Structure:*	Group collaboration	Gamification	Academic coursework	None	None
	Duration:	Weeks	Months	Months	Hours	Day
	Study period:	2006–2009	2009–2013	2010–2013	Dec 2012	2012–2013
	Project size:	852	4,858	2,914	8,246	249

Project abbreviations as follows: CotW: WikiProjects' Collaboration of the Week; WEP: Wikipedia Education Program; CP: Community Portal; TAFI: Today's Article for Improvement. Project size is measured in number of articles.

- **Collaboration of the Week** (**CotW**). Some of the WikiProjects organise what is known as a **Collaboration of the Week**, where they focus on improving a specific article or a set of articles.

  As we see in Table 4.1 the **people** in CotW are *internal* as nearly all of them are already Wikipedia contributors and members of a specific WikiProject, and the work is done as a *group*. The collaboration's **purpose** is to achieve the group goal of improving a specific article or set of articles. Similarly the **structure** is a *group collaboration*. The vast majority of the collaborations last *a week or two*, on par with the name, but some last as long as a month.

- The **WikiCup**[1] is a competition for Wikipedia contributors. Since 2009, the cup's organisation has been fairly stable, with four initial rounds followed by a

[1] `https://en.wikipedia.org/wiki/Wikipedia:WikiCup`

final round, each round lasting approximately two months. There are comprehensive rules, and three judges award points. In each round contestants score points for achieving specific tasks. For example, contributing significantly to an article that successfully passes peer review for Featured Article status (the highest-quality Wikipedia article status) is awarded 100 points. In addition to the competitive aspects, it also is emphasised that the most important rule of the cup is "*just a bit of fun*" (emphasis theirs).

As with CotW, WikiCup **people** are *internal* to Wikipedia, with contestants most likely already experienced members of the Wikipedia community. Work is done through (and assessed in terms of) *individual* effort. The **purpose** of the WikiCup is described on the cup pages. Of course, its primary purpose is to improve the encyclopedia, but as noted *scoring points* and *having fun* also are called out. Given the point scoring system, the cup **structure** involves *gamification* [Det+11], and the duration is *months*.

- The **Wikipedia Education Program**[2] (**WEP**) started as an organised effort connecting U.S. and Canadian university instructors and students with ambassadors from the Wikipedia community. The original intent of the program was for students to improve the content of public policy articles as part of class assignments. It has since expanded to other subject areas, countries, and languages. The Wikimedia Foundation says that there have been over 6,500 participants who have added "the equivalent of 45,000 printed pages of quality content"[3].

[2] `https://outreach.wikimedia.org/wiki/Education`
[3] `https://outreach.wikimedia.org/w/index.php?title=Education/About&oldid=66258`

The **people** in WEP are *external* to Wikipedia, specifically students at colleges and universities. In some courses, the students work individually on articles while in others they do group work, so we consider the project as having both. Since the work is done as part of college courses, we define the **purpose** of WEP to be *course credit*. Given the shared context of post-secondary education and Wikipedia, the **structure** is *academic coursework*. Like the WikiCup, the duration of WEP courses is on the order of *months*, typically a U.S. semester of three to four months.

- The English Wikipedia's **Community Portal**[4] (**CP**) serves several purposes, such as helping visitors learn what Wikipedia is about and how to do various Wikipedia tasks. However, what is relevant to our purposes is that it also features a list of articles that need improvement. The CP is easily accessed through a link in the menu on the left-hand side of any page on the English Wikipedia and is typically viewed about 10,000 times per day.

  While the **people** who visit the Community Portal are already on the Wikipedia site, they might not be members of the Wikipedia community (i.e., editors). Most Wikipedia articles do not require a registered account to be edited, and the Community Portal is a general call to action, which leads us to define the recruitment target as *anyone*. The CP does not feature any group collaboration or awareness mechanisms, so we regard it as *individual* work. The **purpose** of the list of articles that need improvement is simply article improvement. The CP provides no **structure**, and articles typically are promoted for one hour.

[4] `https://en.wikipedia.org/wiki/Wikipedia:Community_portal`

- **Today's Article for Improvement**[5] (**TAFI**) is a WikiProject started in July 2012 with the goal of identifying "an undeveloped or underdeveloped article", which would then be promoted through various channels in Wikipedia. As of late May 2014 the project had 109 listed members.

  The **people** who participate in TAFI are recruited on Wikipedia through posts on project members' talk pages[6] and on the Community Portal, but also externally. For instance, some TAFI articles have been promoted on the official Wikipedia Twitter account. Thus, we define this project's recruitment target as *anyone*. Due to TAFI being organised by a WikiProject we see it as primary *group work*, but there likely also are individual efforts being made. There is no obvious secondary **purpose**, as the project is so clearly organised on improving a given article. No **structure** is provided, and the duration of TAFI is a single day.

4.3 Datasets

4.3.1 Collaboration of the Week

We began with the collection of WikiProjects and articles studied by Zhu et al. [ZKK12] We removed deleted articles, collaborations that targeted categories, and collaborations where it was unclear which article(s) they worked on. The result is a dataset of 852 articles spanning from 2006 to 2009.

[5] `https://en.wikipedia.org/wiki/Wikipedia:TAFI`
[6] Every Wikipedia user has a talk page where they can be contacted.

Table 4.2: WikiCup participation

Year	Ours	Official	Prop. (%)
2009	25	81	30.9
2010	53	135	39.3
2011	61	117	52.1
2012	51	111	45.9
2013	66	127	52.0
Total	256	571	46.4

"Ours" is the number of participants per year in our dataset, "Official" is the number of WikiCup participants reported for that year.

4.3.2 WikiCup

Each WikiCup contestant has a page where they submit the work they have done for scoring review. We mined these pages for contestants in the cups from 2009 through 2013, as those cups have had the same format and a fairly stable scoring system. The result is a dataset with 256 contestants and 4,858 articles. This number of contestants is lower than the "official number" listed on the relevant WikiCup pages, Table 4.2 gives a yearly overview. We suspect this difference is because some users sign up but withdraw from the competition during a round or get disqualified. Therefore, we do not suspect this results in a distorted sample for our analysis.

4.3.3 Wikipedia Education Program

We mined three sources to gather a dataset covering 258 courses, 2,914 articles, and 2,870 students:

1. The U.S. and Canadian Education Program list of courses[7], which includes the

[7] `https://en.wikipedia.org/wiki/Wikipedia:Education_program/Courses`

Public Policy Initiative.

2. The Education Program extension's database, which covers the more recent courses in the program.

3. The APS Wikipedia Initiative's Wikipedia page[8]. The APS Wikipedia Initiative is to some extent a separate project, but it still fits with the Education Program since some of the APS Wikipedia Initiative courses are included in the Education Program lists of courses.

We included data only from courses where individual students selected specific articles to work on, thus yielding an explicit record of the work done.

There also is an Indian Education Program that has worked on articles in the English Wikipedia. We did not include this project in our dataset for two reasons. First, the Wikimedia Foundation published a report[9] that described early contributions as "poor quality and/or ridden with copyright violations", and second, the remaining WEP courses form a fairly coherent group. There are now education programs in several countries and we plan to study these in future work.

4.3.4 The Community Portal

The list of articles that need improvement on the CP is updated automatically by a bot[10] roughly every hour. We mined the edit history of the Community Portal to gather a dataset of articles listed from December 4, 2012 to January 4, 2013. The bot updates 40 articles every time, and during the given time span 741 updates were

[8] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Psychology/APS-Wikipedia_Initiative`

[9] `https://en.wikipedia.org/wiki/Wikipedia:India_Education_Program/Analysis/Quantitative_Analysis`

[10] Software robot, ref `https://en.wikipedia.org/wiki/Wikipedia:Bots`

made. Some articles were featured multiple times, resulting in a total of 8,246 unique articles.

4.3.5 Today's Article for Improvement

Our dataset of TAFI articles was collected from the project's archived schedule[11] as well as any article having the template "Former TAFI" applied to it. This resulted in a dataset containing 249 articles from July 2012 through December 2013.

4.3.6 Common properties of all datasets

For each article in each of the datasets, we gather the source (text and wiki markup) of the article at the start and end of every project. For the WikiCup and WEP, the end of the project is defined as the last edit by any project participant during the project. This is to ensure that we do not also capture additional work done by other editors. The remaining improvement projects are time-bound, e.g. the end of TAFI is the end of the day an article was selected.

We also gather data on the number of contributors working on each article between the start and end of each project. In the Education Program students assign themselves to specific articles, which provides an explicit mapping for us to use. For TAFI, CotW, and the WikiCup, we search the edit history of each article. We remove three categories of contributors: bots, because those are automated tools; those who were reverted by a bot or through common anti-vandal tools since they were likely vandalistic edits; and those who made reverts using common anti-vandal tools as that is maintenance work. The remaining set of contributors should be those who tried to make productive edits to an article.

[11] `https://en.wikipedia.org/wiki/Wikipedia:Today%27s_articles_for_improvement/Archives/Schedule`

Table 4.3: English Wikipedia's seven assessment classes

Class:	Stub	Start	C	B	GA	A	FA
Quality:	*Low*						*High*

Abbreviations: GA=Good Article, FA=Featured Article.

4.4 Measuring Project Performance

The most important thing to measure about quality improvement projects is how much they improve the quality of the articles within their scope. This means we need a way to assess the quality of the articles in the datasets. There are several possible approaches to doing so, with the majority having appeared in the literature:

1. Using Wikipedia's own article quality assessments.
2. Gathering expert human assessment of randomly sampled articles (e.g. [Rot; WMF]).
3. Crowdsourcing human assessment of a random sample of articles (e.g. [KK08]).
4. Using proxy measures for quality, e.g. words added and word survival (e.g. [FK13; Hal+09]).
5. Leveraging machine learning techniques for predicting article quality (e.g. [WCR13]).

Each one of these approaches comes with benefits and drawbacks. Wikipedia's own assessments are done by Wikipedia contributors using the seven-class scale shown in Table 4.3. This notion of article quality in Wikipedia has been shown to correspond well with existing notions of encyclopaedic quality [Stv+08a]. However, because these

assessments are done by people, there is a potential time lag between substantial changes to an article and its subsequent (re)assessment. As we are interested in measuring the immediate effect of article improvement work, the lag makes us unable to use the reassessments without further analysis.

The drawback of using experts to assess random samples is that it limits the number of samples, which can reduce statistical power. We came across a similar problem in one of our validation datasets described in the previous chapter, where 257 articles were not enough to tease out the effects we seek to understand. While crowdsourcing assessments has been shown to be significantly correlated with Wikipedia's own assessments [KK08], the correlation ($r_s = 0.54$) also suggested disagreements. As we reported in the previous chapter, we were able to produce higher correlations using a machine learning approach. Using proxy measures for quality would mean we would end up not capturing many features associated with article quality (e.g. the presence of references to sources or illustrative images). A machine learning approach enables measuring the entirety of the datasets, but will make prediction errors, requiring analysis of where prediction errors occur and how they affect overall results.

The following section describes our results using the machine learning model to predict quality. As we saw in the previous chapter these findings are consistent across the two approaches we used, both machine learning and human assessments.

4.5 Results

We focus on three main findings in this section. The first relates to the **people** component of our framework, and the second to the **policies/structure** component. Our third finding can be seen as relating either to **purpose** or **policies** depending

on the improvement project's design. To complete our framework, we discuss this finding under the **purpose** component.

4.5.1 People

Result: More People, Less Quality

A fundamental question facing the designer of any effort to improve quality is: does it pay off to have contributors working individually on each artefact in the effort, or should they work in groups? Three of the studied projects have varying number of contributors per artefact, allowing us to investigate this question. Our results suggest that an *increased number of contributors per artefact* is associated with a *lower rate of increase in artefact quality.*

We examine the relationship between number of contributors and quality in the Collaboration of the Week (CotW), the Wikipedia Education Program (WEP), and the WikiCup. In all of these datasets, we have predicted the quality of each article at the start and end of the project using our quality machine learning model that we developed in section 3.2. We also calculated the number of contributors to each article during the project. The distribution of number of contributors is highly skewed in the CotW and WikiCup datasets. This is not uncommon for contributions to online communities. We therefore choose to log-transform these variables. The WEP dataset does not have the skewness issue. Group size in college classes is limited, so the most common size for WEP efforts is 2-5, and only a few outliers have more than 6 people.

To model the relationship between number of contributors and predicted quality we use an Ordinal Logistic Regression [LA05; Yee10] (OLR) with the assessment classes in the order shown in Table 4.3. We have a variable $n\_contributors$ for the

Table 4.4: Model coefficients for Collaboration of the Week

	Estimate	**Std. Error**	**P-value**
Intercept: $Stub\|Start$	-2.57	0.25	***
Intercept: $Start\|C$	-0.12	0.15	
Intercept: $C\|B$	1.30	0.16	***
Intercept: $B\|GA$	3.07	0.19	***
Intercept: $GA\|FA$	4.36	0.23	***
$log_2(n\_contributors)$	-0.51	0.06	***

P-values: *** < 0.001

number of contributors per artefact and add a binary variable $from\_scratch$ in the WEP and WikiCup dataset to control for articles that did not exist prior to the start of the project (thus having an unknown prior quality).

During our model building we also want to control for two additional issues: the *proportional odds assumption* and whether there is an *interaction effect* between our independent variables. The former is a fundamental assumption upon which OLRs are commonly built. In our case, it means the coefficients explaining the relationship between Stub-class ($P(Stub)$) and higher than Stub-class ($P(\geq Stub)$) also explain all other classes (e.g. $P(C)$ and $P(\geq C)$). We have verified that this assumption holds in all our OLR models. Second, we also verified that there is no interaction effect between our independent variables, which would have indicated that the strength of the effect of starting an article from scratch would be altered by the number of contributors to the article.

The results of our OLR models, one for each effort, are listed in Tables 4.4, 4.5, and 4.6. All predictors are statistically significant in all models. In the CotW model, the intercept (cutpoint) between Start and C-class is not significant. Because this cutpoint is only an estimate of the borderline between the two classes and the

Table 4.5: Model coefficients for Wikipedia Education Program

	Estimate	**Std. Error**	**P-value**
Intercept: $Stub\|Start$	-2.94	0.10	***
Intercept: $Start\|C$	-1.26	0.07	***
Intercept: $C\|B$	0.44	0.07	***
Intercept: $B\|GA$	2.18	0.08	***
Intercept: $GA\|FA$	3.25	0.11	***
$from\_scratchTRUE$	0.47	0.08	***
$n\_contributors$	-0.10	0.03	**

P-values: ** < 0.01, *** < 0.001

Table 4.6: Model coefficients for the WikiCup

	Estimate	**Std. Error**	**P-value**
Intercept: $Stub\|Start$	-5.35	0.26	***
Intercept: $Start\|C$	-1.23	0.09	***
Intercept: $C\|B$	0.27	0.09	**
Intercept: $B\|GA$	0.43	0.09	***
Intercept: $GA\|FA$	3.15	0.11	***
$from\_scratchTRUE$	1.89	0.08	***
$log_2(n\_contributors)$	-0.63	0.04	***

P-values: ** < 0.01 *** < 0.001

predictor's P-value < 0.001, this issue does not invalidate the model.

Across all three efforts the number of contributors has a negative sign indicating that larger numbers of contributors per artefact is associated with slower increase in quality. We also built additional models where we controlled for the quality at the start of the effort, to make sure that our model was not influenced by (for example) a larger proportion of articles starting from a certain quality level. Pre-effort quality was generally also a significant predictor in those models, but did not cancel out the effect of number of contributors. This means that consistently across these projects, an *increase in number of contributors per artefact* is connected with a *negative impact on the rate of quality increase.*

Discussion

We find it particularly interesting that the negative effect of additional contributors per artefact is consistent across all three projects, even though the nature of the "group" is different: in the WEP, participants are explicitly connected to an article, while in the other two projects we count all likely productive editors as participants. Wikipedia articles are of course open to anyone to edit, but WEP students are directed to work in a "sandbox", a personal space where they can draft an article before publishing it, as described in the template syllabi[12]. This usage of personal work spaces likely isolates many of the WEP articles from contributions from non-WEP contributors until they are published.

As we will see in the next section, WEP students seldom take articles above B-class quality, supporting the findings of the Wikimedia Foundation's studies on WEP quality [Rot; WMF]. This could be due to a lack of experience with writing Wikipedia articles, but it could also be due to satisficing [Sim56], they are doing just enough for a reasonable grade but nothing more. Groups of students might also be experiencing social loafing [KW93], e.g. that some of the group members are trying to free ride their way through the course while other members do the work. Future research on the WEP could try to tease these effects apart.

The groups of contributors in the WikiCup and CotW datasets are more implicit, and the extent to which participants in these efforts use sandboxes to edit articles before publication is unknown. It may be that additional contributors to those two efforts are not aware that they are taking part in an improvement project, which could alter their edit behaviour. These contributors may also differ in experience

[12] `http://outreach.wikimedia.org/w/index.php?title=Education/Syllabi&oldid=70162`

levels and engagement with the Wikipedia community, previous research has shown that power users in Wikipedia produce higher quality edits from early on [PHT09]. Additional contributors could also be positive as long as only a few contributors are doing the majority of the work, as found by Kittur and Kraut [KK08], otherwise they just cause more maintenance overhead, similar to how adding people to late software development projects make them even later [Bro75].

These results also beg the question of whether it is better for groups creating artefacts to work individually and sequentially. André et al. [AKK14] found simultaneous work to be less effective than a sequential structure, but the effect was mitigated by assigning specific roles to participants. In Wikipedia there are few formal roles. Some users are promoted to become administrators, a role that is supposed to be janitorial and "not a big deal"[13] (yet research indicates it is an increasingly bigger deal [BK08]). Instead, users assume informal roles, which they may seek to use to their advantage in conflicts, for instance by questioning other contributors' expertise [Kri+07].

In order for a peer production community to be successful, there needs to be collaboration. These results suggests some degree of conflict between individual and group work, when does one approach benefit the community more than the other? We see investigations into how contributor roles, work organisation, conflict, coordination, and concentration of contributor effort affect artefact quality in improvement projects as a promising venue for future research.

[13] `https://en.wikipedia.org/wiki/Wikipedia:DEAL#History`

Table 4.7: Wikipedia Education Program Quality

	Stub	**Start**	**C**	**B**	**GA**	**FA**
NA	3.93	14.88	46.79	18.81	12.50	3.10
Stub	21.43	20.44	37.68	9.61	9.61	1.23
Start	0.95	25.79	44.94	17.25	8.07	3.01
C	0.00	1.49	54.29	25.00	12.87	6.34
B	0.00	1.21	16.92	65.86	6.95	9.06
GA	0.00	0.00	10.58	11.54	61.54	16.35
FA	0.00	0.00	3.08	6.15	7.69	83.08

Prior (rows) and post (columns) predicted quality. Proportions are relative to prior quality (rows). NA=Article did not exist prior to start of a course.

Table 4.8: WikiCup Quality

	Stub	**Start**	**C**	**B**	**GA**	**FA**
NA	0.90	34.11	37.95	3.60	21.46	1.98
Stub	1.18	28.10	40.26	2.09	26.67	1.70
Start	0.00	24.25	17.55	2.08	51.96	4.16
C	0.00	1.18	47.14	3.16	39.05	9.47
B	0.00	0.00	5.74	36.89	47.54	9.84
GA	0.00	0.19	1.23	0.85	89.26	8.48
FA	0.00	0.33	0.33	0.66	7.95	90.73

Prior (rows) and post (columns) predicted quality. Proportions are relative to prior quality (rows). NA=Article did not exist prior to start of a cup round.

4.5.2 Purpose

Result: New Artefacts, Higher Quality

Is it more effective to have participants in a quality improvement project create new artefacts or work on improving existing ones? Two of our efforts, the WikiCup and the Wikipedia Education Program include both types of work. Across both of these, our results indicate that artefacts created from scratch end with a higher final quality.

Table 4.9: Collaboration of the Week Quality

	Stub	**Start**	**C**	**B**	**GA**	**FA**
Stub	29.46	40.31	17.83	6.20	4.65	1.55
Start	1.74	46.52	27.39	18.26	6.09	0.00
C	0.00	0.65	69.48	17.53	10.39	1.95
B	0.00	1.45	14.49	72.83	6.52	4.71
GA	0.00	0.00	4.88	2.44	80.49	12.20
FA	0.00	0.00	0.00	9.09	9.09	81.82

Prior (rows) and post (columns) predicted quality. Proportions are relative to prior quality (rows).

To investigate this effect, we first look at what level of quality articles reach at the end of a project. Table 4.7 (WEP) and Table 4.8 (WikiCup) show the relationship between predicted quality at the start of an improvement project (rows) and at the end (columns), where Good Article is abbreviated "GA" and Featured Article "FA". For the WikiCup, the end of the project is the last edit done by a specific participant on the article that participant submits for scoring, and for the WEP the end is the last edit done by any student assigned to a specific article. For convenience, we have also included the same type of table for the Collaboration of the Week (Table 4.9) but note that in that project no articles were created from scratch.

In the WEP (Table 4.7), 65.6% of the articles started from scratch (the "NA" row) reach an intermediate level of quality (C- or B-class). This is not the case for the WikiCup (Table 4.8), where instead more than one fifth of every new article is predicted as Good Article (GA) or Featured Article (FA) status. We can also see that to a certain degree in the WikiCup, and to a much larger degree in the WEP, many articles do not improve enough to change their predicted quality class.

More generally, our OLR models in Tables 4.5 and 4.6 show that the $from\_scratch$ variable is a significant predictor with a positive relationship to end quality. This sug-

gests that in both the WikiCup and WEP projects, *new artefacts have higher end quality* compared to existing artefacts.

Discussion

Here again we have found an effect that is consistent across vastly different improvement projects. As previously noted, many WEP articles are likely isolated from contributions from non-WEP editors due to the extensive use of sandboxes. In contrast, the WikiCup has an "In the news" category for articles that are featured in that section on the English Wikipedia's front page, with the contestant scoring 10 points for each article featured. This will likely lead to the cup containing some breaking news articles [KGC13], newly created articles where the particularly high interest and resulting traffic could lead to quicker improvements in quality.

The result also is interesting because both projects have a long duration, namely months. With that amount of time available, one would not expect there to be a significant difference in quality improvement between new and existing articles, particularly one in favour of new articles. Producing high-quality Wikipedia articles requires access to resources, for instance sources for claims and illustrative images. For some types of content these might be more difficult to find, particularly using online resources, and in the case of existing artefacts resources might already have been exhausted. The lack of online sources could to some extent explain the WikiCup result where participants might strongly prefer them, but it seems unlikely to explain the WEP result, since students should have access to good library resources.

Existing artefacts are also more likely to have some contributors monitoring them. Research on Wikipedia has shown that editors assume ownership of content [Hal+09; TCG09] although Wikipedia's own policy states noone owns an arti-

cle[14]. This type of territoriality also occurs outside of Wikipedia, expert contributors to a museum tagging system were found to more strongly express ownership of content than novices [TCG10]. When participants in a quality improvement project try to make changes, territoriality by existing contributors is likely a barrier to entry, resulting in reduced quality gain through coordination overhead.

It is also not obvious that peer production communities should always focus on creating new artefacts. If the community already has good coverage (e.g. a large number of articles), it would perhaps instead benefit the most if work was concentrated on improving existing artefacts. Community managers could combine the understanding of this trade-off between coverage and quality with information on audience attention to guide contributions to the areas where they are most needed in order to ensure the community's resources are utilised most efficiently.

Perhaps people work differently if they start with a blank slate than if they have to modify an existing piece of work. In their study on the effect of seeding wikis with content, Solomon and Wash [SW12] found that not seeding led to significantly more content added, while those who started with seeded content would instead use that as a model. We do not know to what extent this finding also is present in the work WikiCup and WEP participants do on existing articles. There is an opportunity here for both qualitative analysis of live data as well as lab studies to understand the effects that are in play and how to most efficiently produce high quality artefacts.

We also found interesting differences between improvement projects in the patterns of change in predicted quality. The WikiCup results (Table 4.8) show that few articles move into the B-class. Instead, the cup participants push articles upwards to GA/FA status, a behaviour we interpret to be clearly in line with the cup's incentive

[14] `http://en.wikipedia.org/wiki/Wikipedia:Ownership_of_articles`

mechanism. Successful Good Article nominations score 30 points and Featured Articles 100 points, while getting an article only to B-class scores zero. This is similar to how badges steer user behaviour in Q&A systems [And+13]: when users have nearly reached a badge threshold, they will modify their behaviour to achieve the badge as quickly as possible.

The results for WEP (Table 4.7) and the Collaboration of the Week (Table 4.9) show that for many articles quality does not appear to improve. The CotW's short duration, usually a week or two can explain the effect in that project. Most improvements in CotW occur in low quality articles, confirming Zhu et al.'s description of those articles being the typical collaboration targets [ZKK12]. That the Education Program also to a large extent leads to improvements that appear to not substantially change the article quality is more concerning. Students in the program have more time available to affect change, thus we wonder if they are struggling with learning how to write articles in the context of Wikipedia, for instance how to correctly source content with footnotes and citation templates.

4.5.3 Policies/Structure

Result: Structure is Required

Two of the improvement projects we study, Today's Article for Improvement (TAFI) and Wikipedia's Community Portal, do not have a well-defined structure. For example, they are open to anyone, have a very general purpose, and lack a clear incentive mechanism. Our results indicate that unlike the other projects, neither TAFI nor the Community Portal is particularly successful at improving artefact quality.

First, let us investigate the TAFI project. We predicted the quality of each article

at the beginning and end of the effort, in this case the day the article was promoted for improvement. Only 9 out of 249 articles (3.6%) saw an improvement in predicted class, and of those all but one moved up a single class, the exception being a Start-class article improving to B-class.

Is the problem lack of participation? TAFI started in mid-2012 and at the end of 2013 the project's member list contained 103 usernames. Still, of the 249 articles in the TAFI dataset 56.2% had no contributors during the day of the effort. We investigated whether the degree of participation changed over the course of the project and found that in the first three months, 1 out of 16 articles saw no contributors, while in the last three months it was 33 out of 47. This is a statistically significant difference (Fisher exact count test $p < 0.001$); the project has seen a significant decline in participation as its membership has increased.

We found a similar problem with participation in the Community Portal. Our dataset covers Dec 4, 2012 to Jan 4, 2013, during which time the portal, according to data from the Wikimedia Foundation[15], saw 314,534 views, for a daily average of 10,146 views. One would hypothesise that these views would directly affect listed articles as visitors to the portal follow links to edit them. To investigate this, we calculated average views/day 14 days prior to being listed for the portal articles and removed those articles that were listed twice on the same day due to our view data having a granularity of one day. We sorted the articles into buckets based on average views/day, using exponential buckets since article popularity follows a power-law distribution. Lastly we calculated views on the day of listing, as well as average views/day up to 14 days after.

Articles are typically listed for only one hour, so one would expect the portal to

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

Table 4.10: Excerpt of view statistics for the Community Portal

Bucket	Prior Views	Listing Gain/Loss (%)	Post Gain/Loss (%)
$0 \leq x < 2$	1.6	126.08	80.34
$2 \leq x < 4$	3.1	33.12	12.10
$4 \leq x < 8$	5.6	6.74	-4.36
$8 \leq x < 16$	11.2	-3.68	-8.50

Articles are placed in buckets based on prior mean views. "Prior Views" is the mean views of said bucket. "Listing Gain/Loss" shows the increase/decrease in views on the day of being featured, while "Post Gain/Loss" shows the equivalent for the 14 days after listing.

affect article views less as popularity increases. This is also seen in Table 4.10, which shows an excerpt of the view results. The remaining part of the table (up to x $\geq$ 16,384 views/day) is left out for brevity as the trend of a decrease in views on the listed day as well as in the period after continues. Based on the results in Table 4.10 it seems clear that few views appear to come from the Community Portal.

Not surprisingly, since including an article on the Community Portal did not increase how much it was viewed, it also didn't increase participation. We selected portal articles which had no edits in the two weeks prior to being listed, because those articles are most likely getting contributions from the portal. Out of 4,410 articles only eight of them were edited during the time they were listed. This extends data from the Wikimedia Foundation during a redesign of the portal in late 2012, where 220,000 portal views led to 46 saved edits [16]. Since the portal does not lead to participation, there obviously can be no improvements in quality. Therefore it is not an example of a successful improvement project.

[16] `https://meta.wikimedia.org/wiki/Research:Community_portal_redesign/Opentask`

Discussion

Both of the unstructured projects studied were largely unsuccessful. The short duration of these projects, an hour in the case of the Community Portal and a day for the TAFI project, might be posited as the explanation for the lack of success. However, the Community Portal is easily accessible from the left-hand menu of any Wikipedia page, and as we saw exposes a lot of readers to its call to action. In our related work we referenced several successful projects with a much similar approach: Cosley et al. [Cos+06] suggested edits of movie data on a movie recommender site; a general call to action solicited contributions in the Cyclopath geo-wiki [PMT10]; Halfaker et al. [HKT13] asked Wikipedia readers to submit article feedback. In all three cases more structure and guidance was supplied when necessary, for instance Cosley et al. had a form for inputting data, and the Cyclopath experiment provided volunteers with clear instructions for the work needed.

Comparing TAFI and the Community Portal to the other projects, we also see that these two unsuccessful projects lack a clear purpose, perhaps it is unclear to potential participants what the benefit is to both them and the encyclopaedia. In contrast, many of the WEP courses aim to extend Wikipedia's content in areas where it is lacking (e.g. public policy or psychology), and the WikiCup's scoring system appears to steer participant behaviour, as seen in their movement of articles to the higher quality classes to score points. Neither TAFI nor the Community Portal implements similar incentive mechanisms. Where our initial investigation has pointed to a lack of participation, future work could look at how much structure and what kind of incentive mechanisms are needed to trigger increased participation to cross the border into a successful improvement project.

4.6 Limitations

This research has several known limitations. First, while the English Wikipedia is the largest peer production community in existence, results from this community might not generalise. For example, a Q&A system like Stack Overflow is also a peer production community with some wiki-like features. Research to determine to what extent our findings also are present there (or in other peer production communities) would be valuable.

Second, our analysis uses Wikipedia's own assessment classes. Wikipedia's notion of article quality has been shown to correspond well with existing notions of encyclopaedic quality [Stv+08a]. In our analysis of prediction errors, we discovered that in some cases Wikipedia contributors failed to apply the assessment criteria correctly, leading to articles being assessed into a lower class. This suggests that there is room for improvement in the understanding of how Wikipedia contributors apply the assessment criteria, as well as how these correspond to assessment of quality by non-Wikipedians, and we plan future work in this area.

This paper brings together a diverse set of improvement projects, which means we must also consider limitations imposed by them. There is likely a clear difference in skill levels between some of the efforts. Contestants in the WikiCup and WikiProject members participating in the Collaboration of the Week are probably skilled members of the Wikipedia community, while students in the Education Program have little prior experience with writing for Wikipedia. One way to control for this would be to introduce measures of tenure, for instance the number of edits a contributor has or the amount of time since account registration.

We are also limited by how we define a contributor to a specific article. In the

Education Program we use the course pages' explicit definition of which students worked on a specific article, and in the WikiCup we use contestants' submission pages to track which articles they worked on as part of the cup. In the other efforts, we instead use an implicit method of defining participants. This method could potentially be improved by algorithmic content analysis, for instance to account for different categories of contributors (e.g. newly registered users, users without an account, etc).

4.7 Conclusion

In this chapter we studied factors associated with the success and failure of quality improvement projects in peer production communities. We used Preece's three components of online communities (people, purpose, and policies) as building blocks for a coherent analytic framework to study five diverse quality improvement projects in the English Wikipedia. In summary, our findings and their implications for the design of quality improvement projects were as follows:

1. **People:** Increasing number of contributors per artefact is associated with slower increase in quality. Consideration should be given to when working individually can be more effective than group work.

2. **Purpose:** Artefacts created during the improvement project are connected to a higher quality level than existing artefacts worked on during the project. There may even be cases where deleting an old artefact to start over is preferred, although more research is needed.

3. **Policies/Structure:** Unstructured efforts are less likely to succeed. Our results suggest that new efforts should provide a carefully designed socio-technical structure, for instance through incentive mechanisms appropriate for the desired work and the knowledge level of the participants.

Chapter 5

Misalignment Between Supply and Demand

In peer production communities, individual community members typically decide for themselves where to make contributions, often driven by factors such as "fun" [Nov07] or a belief that "information should be free" [LW05]. 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 chapter we study the relationship between how quality content is consumed and produced. We introduce an analytical framework we call the *Perfect Alignment Hypothesis* and use it to identify a great deal of misalignment between the production of high quality content and its consumption in two successful peer production communities: Wikipedia and OpenStreetMap. Our framework also allows us to measure the *impact* of the misalignment, and we find that almost two billion monthly article views in Wikipedia go to articles that would be of much higher quality if editors optimally distributed their work according to reader demand. Applying this framework to OpenStreetMap we also find extensive misalignment between supply and demand, and estimate that 7.3 million people in the US live in areas where the map quality is much lower than demand would suggest.

We also study the nature of misalignment in both communities. In Wikipedia we find that reader demand for certain topics (e.g. LGBT issues) far exceeds the 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). Additionally, we find that a majority of the articles in the highest demand appear to be continuously in high demand. This means that high demand is not only driven by short-term trends (e.g. breaking news). In OpenStreetMap we find that areas of low quality but high demand are more likely in small towns and have lower socioeconomic status. These findings will be put into context and discussed in more detail in Section 5.6, where we also suggest sociotechnical solutions to assuaging the misalignment.

The Wikipedia portion of this research was published in the proceedings of ICWSM in 2015 [War+15a]. Our study of OpenStreetMap is new work for this thesis. We rewrote the ICWSM paper to also include the OpenStreetMap results, and that extended version has been submitted to the journal ACM Transactions on Computer-Human Interaction. This chapter is an adapted version of said journal submission.

5.1 Research questions

The background research covered in Chapter 2 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/areas are over-represented amongst artifacts that are low quality/high demand, and high quality/low demand?

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 our two results sections, one each studying this phenomenon in Wikipedia and OpenStreetMap. 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.

5.2 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 and OpenStreetMap.

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.

While OpenStreetMap does not have a quality scale similar to Wikipedia, geographic information quality is handled quite differently as we will discuss in more detail in our methods section, the PAH describes a similar relationship between supply and demand for quality content. Those areas that have the highest demand would be completely mapped while those that are accessed more rarely will contain only a basic set of information. Due to how both supply and demand can be measured in OpenStreetMap, we model the PAH as the linear relationship between number of OSM contributors (quality supply) and population (quality demand). This means the PAH conditions result in a map where high population areas have high quality and areas that are scarcely populated have lower quality.

In the following sections, we use the Perfect Alignment Hypothesis to understand exactly how far away our peer production communities are 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.

5.3 Methods and Datasets

To enable the study of (mis)alignment in peer production communities, we needed examples of successful communities. We 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.

In order to establish generality of misalignment, we sought to study a different type of peer production community and found OpenStreetMap (OSM) to be a good fit. Where Wikipedia's content is primarily text, OSM is "volunteered geographic information" (VGI) [Goo07], and the community is mapping the entire planet. Started in 2004, the community has about 36,000 active members[1]. It is also a community that has been widely studied in the research literature [NZ14].

5.3.1 Wikipedia Datasets

Supply of Quality Content

Each of these four language editions we studied 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[2], 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.

[1] `http://osmstats.neis-one.org` retrieved June 29, 2016.
[2] `https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment`

Table 5.1: Overview of Wikipedia editions

Language	No. articles	Rated articles	Quality Classes
English	4.67M	3.6M	7
French	1.58M	929k	6
Russian	1.18M	170k	7
Portuguese	862k	444k	6

An article counts as rated if it has at least one quality rating. "Classes" shows the number of classes used in the rating scale. Abbreviations: M=million, k=thousand.

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 [Stv+08a], and research has also shown that these quality ratings correlate relatively well with reader judgement of article quality [KK08]. 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. [Leh+14].

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[3]. 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 5.1.

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

Demand for Quality Content

We measure demand using Wikipedia article pageviews as made available by the Wikimedia Foundation[4], the best available source for per-article view data. Following Hill and Shaw's [HS14b] 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-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 [Thi+12; YSK12], 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.

5.3.2 OpenStreetMap Datasets

Supply of Quality Content

Previous work on quality in OSM has mainly focused on positional accuracy [Hak10; Hak+10; ZZ10; LVK11; Ars+13], comparing the position of objects in OSM such as the road network against those in a gold standard dataset.

We were interested in ways of measuring *overall* quality in OSM. This is similar to how we in Wikipedia decided to use assessment ratings, because they capture a more human-oriented concept of quality than a proxy like "number of words". Quality in

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

geographic data is defined in ISO 19157 [13] and has five parameters: completeness, logical consistency, positional accuracy, temporal accuracy, and thematic accuracy.

The number of contributors to an area can be used as reasonable proxy for overall quality in OpenStreetMap. Haklay et al. [Hak+10] found that as the number of OSM contributors to an area increased the error in the positional accuracy also decreased towards an asymptotic minimum. A similar relationship between number of contributors and other aspects of content quality (e.g. completeness in the form of number of objects) were found by Girres and Touya [GT10]. Barron, Neis, and Zipf [BNZ14] proposed a quality framework for OpenStreetMap quality analysis where they suggest that "a high and increasing number of people who have ever created or edited OSM data within an area indicates a possibly better data quality."

To measure the number of contributors, we gathered a dataset of the entire history of OSM in North America[5]. This dataset (also called a "full history dump") is a geographically filtered subset of all contributions to OSM from its start until February 2014. Analysing geographic data is complicated and previous research has also limited their studies to single countries as the largest unit. We further divide the country into smaller units, in our case the focus is on census tracts. Census tracts are "small [and] relatively permanent subdivisions"[6] with a limited population (optimum size of 4,000) that are used as a part of the US Census. Their boundaries are defined with the intention of allowing comparisons from one census to another, meaning they tend to be stable over time. There are also many additional datasets available that can further our understanding of underlying phenomena, for instance socioeconomic

[5]See `http://wiki.openstreetmap.org/wiki/Planet.osm/full` for general information. Our dataset is from `http://osm.personalwerk.de/full-history-extracts/latest/continents/`, which is linked from the OSM wiki page.

[6]`https://www.census.gov/geo/reference/gtc/gtc_ct.html`

status indicators and measures of the urban-rural spectrum.

The OSM datasets consists of three types of objects: nodes, ways, and relations. Nodes are points with a latitude and longitude, and are used to denote points of interests (e.g. restaurants, museums) as well as where line segments start and end (e.g. parts of a road or corners of a house). Ways are collections of nodes that are used for line-type objects such as roads and the perimeter of buildings. Relations are a meta-type and can contain combinations of other objects, including other relations. We omit relations from our analyses because their impact on our calculation of the number of contributors should be negligible for two reasons: first, they only account for 0.05% of all objects and 0.08% of all edits; and second, we also count contributors to a relation's members (child objects such as nodes and ways).

We counted the number of contributors to an area A by first identifying all nodes N that are contained within A (using PostGIS' `ST_Contains` function) and all ways W that are either within A or intersecting it (using PostGIS' `ST_Intersects` function). Ways were allowed to intersect an area so that lines that cross two areas (e.g. roads) will be counted as members of both, instead of not belonging to any area or having to apply heuristics to determine the parent area. We then took the union of all contributors to objects in N and W and defined that set as the contributors to A. Some of these contributors will have made bulk imports of data into OSM, for instance as part of the TIGER/Line import that was completed in 2008 [ZHN13].

While this type of activity is atypical of VGI contributor behaviour as it spans a large area and makes many edits over a short period of time, it has also been described as "single-user" [QMC14]. These bulk edits can make up a large part of the activity in OSM, but since we measure the number of *contributors* they will not skew our data because they only show up as a single contributor. This means that

activity such as bulk imports are unlikely to bias our dataset, thus we retained these data for our analyses.

Demand for Quality Content

Measuring demand in Wikipedia is relatively easy; as previously mentioned the Wikimedia Foundation makes datasets of article views readily available. The content is also primarily accessed in a central location, readers go to Wikipedia's website to read it. In contrast, measuring demand in OSM is very difficult due to is decentralisation. OSM content is to a large extent used by external tools, for example Mapbox[7], Apple Maps, and Craigslist[8]. Therefore, no central resource of view data for OSM is currently available.

We use population as a proxy for demand; this is reasonable since geographic interactions occur close to where people live [Ste48; Häg68; FA82]. This means that the more people who live in an area, the higher the demand will be for geographic tools. These tools require geographic data such as that supplied by OSM.

This is also reflected in how OSM data is used. Craigslist is a high-traffic site that uses OSM maps, and Craigslist is localised to specific areas[9] where its introduction has been found to affect the classified ad business of local newspapers [SZ14]. OSM content is also used in Apple Maps, meaning it will be used for information about local points of interest (e.g. restaurants) as well as for routing between locations, further establishing that demand for this type of VGI content is localised.

In addition to population, we also explain some additional factors that affect de-

[7] `https://www.mapbox.com`

[8] `http://arstechnica.com/business/2012/08/craigslist-is-on-board-openstreetmap-continues-soaring-to-new-heights/`

[9] `http://www.craigslist.org/about/expansion`

mand. First of all, we account for *spatiotemporal dynamics* by adding datasets from Flickr and Twitter. Flickr images will capture tourism-related activity where contributions occur at long distances from a home location [LGX13; Woo+13]. Tweets from Twitter will capture short-term mobility patterns such as commuting [Nou+12; LGX13; Jur+15]. Secondly, we use median household income as an indicator of *socioeconomic status* (SES). Median household income has been shown to be an important SES factor and is commonly used in this type of research (e.g. [YNS15; TTH15]). In previous research, SES has been shown to correlate with OSM contributions in such a way that lower-SES areas have lower participation [Hak10]. Lastly, we know that activity in social media and user-generated content communities varies across the *urban-rural spectrum* [HS14a; Joh+16a]. We therefore incorporate the urban-rural spectrum in our models through the United States Department of Agriculture's Rural Urban Commuting Area Codes [MCH99][10]. These codes categorise how far along the urban-rural spectrum a census tract is using 10 different categories based on whether it is an urban area (core population above 50,000), urban cluster (core population 10,000 to 49,999), small town (core population 2,500 to 9,999), or outside of these, and whether the population in a given tract commutes to other areas (e.g. to an urban area). Using this type of urban/rural classification is a common approach in this type of research, e.g. [YNS15].

We used the US Census 2012 population estimates[11] as our source of population data. The 2012 estimates were the most recent estimates at the time of our analysis. For Flickr data we used the Yahoo! 100M CC dataset [Tho+16], which contains 50 million geolocated images. This dataset only contains images that are published

[10] `http://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/documentation.aspx`

[11] `http://www.census.gov/popest/`

using the Creative Commons license[12]. It is unknown to what extent this biases the dataset. Our dataset from Twitter consists of 66.8M geolocated tweets from October 10, 2014 to November 19, 2014, where 51M of them were located within US census tracts. The one-month timespan of our Twitter dataset is different from our OSM and Flickr datasets, which both span several years. However, we regard a whole month as a sufficiently long timespan. This specific dataset has also been used in previous research on localness in social media [Joh+16b].

Data on median household income was gathered from the US Community Survey[13]. For data on the urban/rural classification of census tracts we used the 2010 RUCA dataset available from the US Department of Agriculture[14]. Lastly we gathered shapefiles for the census tracts from the US Census[15].

5.3.3 Appropriately Modelling Geographic Data

Wikipedia's quality assessment is, as described earlier, done using an ordinal set of categories. In OpenStreetMap we have a continuous variable (number of contributors) as our proxy for quality, and we seek to understand the relationship between this variable and other continuous variables (e.g. population). A standard approach when studying these types of relationships are linear regressions.

Our analysis of OSM data is based on delineated geographic areas, which in many cases share borders with each other. As people tend to organise into communities with other people who are similar to themselves ("birds of a feather flock together" [Mil01]), our data will contain adjacent areas with similar characteristics. This phenomenon,

[12] `https://creativecommons.org`
[13] `https://www.census.gov/topics/income-poverty/income.html`
[14] `http://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.aspx`
[15] `https://www.census.gov/geo/maps-data/data/tiger-cart-boundary.html`

called "spatial dependency" (or "autocorrelation") [Tob70], invalidates the "independent and identically distributed" assumption of an ordinary least squares regression.

We can solve the problem of spatial dependency by applying an appropriate spatial modelling strategy. The *spatial lag model* (also known as the "spatial auto-regressive model") is our primary approach for modelling spatial data. A variation of this model also incorporates spatial dependency for the explanatory variables, the so-called "Durbin correction", which Yang et al. has argued is the best approach for modelling spatial data [YNS15]. We have confirmed that the Durbin correction leads to a significantly better fit to our data and will report results based on it. All our spatial modelling was done using the "spdep" R library [BHK13; BP15].

Lastly, we define an area's neighbourhood by selecting adjacent areas (also known as "queen's neighbours"), a commonly used approach in this type of research. Defining it using k-nearest neighbours is an alternative approach, and we have confirmed that the queen's approach provided the best fit for our data.

5.4 Misalignment in Wikipedia

5.4.1 The Extent of Misalignment

To understand the amount of misalignment in each of our four Wikipedia 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

Table 5.2: Alignment in the English Wikipedia

	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

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

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 5.2-5.5. 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",

Table 5.3: Alignment in the Russian Wikipedia

	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

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

for which Wikipedia has an opportunity to provide important information about a sensitive topic to readers.

Also visible in each of Tables 5.2-5.5 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 5.4), 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 5.6, 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

Table 5.4: Alignment in the French Wikipedia

	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

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

Table 5.5: Alignment in the Portuguese Wikipedia

	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

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

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 5.2, 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 5.6 also reveals how these communities have been producing content of the

Table 5.6: Overall Proportions of (mis)alignment

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

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).

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 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 Sections 5.4.3 and 5.4.4 where we answer each of the two parts 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.

5.4.2 The Impact of Misalignment

Priedhorsky et al. [Pri+07] 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 5.2-5.5, 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 5.7 provides an overview.

Whereas we previously found large proportions of overall alignment, the results shown in Table 5.7 make it clear that the misalignment that does exist has an

Table 5.7: Impact of Misalignment

Language		Str ExQ	Mod ExQ	Aligned	Mod InQ	Str InQ	Total
English	**N**	89.9	202.9	858.5	1,072.1	1,696.9	3,920.2
	%	2.3	5.2	21.9	27.3	43.3	100.0
Russian	**N**	7.0	10.4	29.9	41.8	112.3	201.5
	%	3.5	5.2	14.8	20.7	55.8	100.0
French	**N**	6.7	18.8	105.3	132.9	120.9	384.7
	%	1.8	4.9	27.4	34.6	31.4	100.0
Portuguese	**N**	2.5	7.9	41.4	45.5	60.6	158.0
	%	1.6	5.0	26.2	28.8	38.3	100.0

Number (N) of article views in million 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. Abbreviations: "Str": Strong; "Mod": Moderate; "ExQ": Excessive Quality; "InQ": Insufficient Quality.

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.

5.4.3 Characterising Misalignment Through 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 [NS08; Hec13]. 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 [KPK09; CRR10; Cho+10; For+12; ZKK12; Mor+14], 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 5.2, 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" dataset.

The second group of articles are those that have reached the highest quality, but

Table 5.8: Overrepresented topics in the Needs Improvement dataset

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

Topics most strongly over-represented in the Needs Improvement dataset, limited to topics with at least 50 articles in the dataset. "N" columns lists number of articles.

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 5.2. 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" 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" [DCT98]. In our case, the relative risk is the probability of encountering a topic in the Needs Improvement/Spent Effort datasets divided by the probability of encountering a topic in the entire English Wikipedia.

Table 5.9: Overrepresented Topics in the Spent Effort dataset

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

Topics most strongly over-represented in the Spent Effort dataset, limited to topics with at least 50 articles in the dataset. "N" column lists number of articles.

Table 5.8 describes the topics that are most over-represented amongst articles in the Needs Improvement 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 5.8 to topics with more than 50 articles in the Needs Improvement 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 5.9 shows the topics most over-represented in the Spent Effort 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 Needs Improvement dataset, for instance the article about the game itself has enough demand to be PAH_7 but is now a middle-quality article[16]. This might be because it is easier or more exciting [LW05; Nov07] to write biographies about cricket players than it is to perfect an article about a more general topic.

Our results also contain examples of how groups of volunteers in peer production communities are not making "efficient" choices about where to supply quality improvements. One of the topics listed in Table 5.9, military history, is a very successful WikiProject with over a thousand active members [For+12]. 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 Needs Improvement 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.

5.4.4 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 dataset, the most popular articles that are not also of the highest quality. As before,

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

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 [EH13], 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 [BD02; HK08].

We first verify that this approach can successfully identify these types of demand changes in our English Wikipedia pageview data. From the Needs Improvement 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 Needs Improvement 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 when we discuss the implications of our findings in Section 5.6. We now turn our attention to studying misalignment in OpenStreetMap in order to establish misalignment as an issue that is not isolated to the Wikipedia community.

5.5 Misalignment in OpenStreetMap

5.5.1 Methods

We use a linear regression to model the relationship between supply and demand of quality content in OpenStreetMap (OSM). Our Perfect Alignment Hypothesis describes perfect alignment as perfect correlation, in other words a linear relationship between these two variables. As we discussed in methods (Section 5.3.2), supply of quality is defined by the number of OSM contributors, since that has been shown to correlate with quality factors such as positional accuracy and completeness. Similarly, demand is defined by population because content usage tends to be localised. These representations of supply and demand mirror the Q- and PAH-classes we used in our examination of misalignment in Wikipedia, but whereas in Wikipedia we had two ordinal sets of categories, in OSM we are comparing continuous variables.

This initial model with population and number of OSM contributors will be called

our "Misalignment" model. In later sections, we will study how additional factors help to explain misalignment. The factors we will look at are: the urban-rural spectrum, because previous research has indicated that the quality and quantity of VGI content varies across this spectrum;spatiotemporal changes in demand (e.g. sports events and tourism), because contribution patterns should align with these types of short- and long-term patterns in demand; and socioeconomic status, because previous research has shown that OSM quality is lower in areas with lower socioeconomic status [Hak10].

We study misalignment in OSM on the census tract level. Larger enumeration units can mask local variation, for instance poorer areas of richer counties. This would make it both more difficult to discover misalignment when it occurs, as well as hinder our examination of explanations for misalignment.

5.5.2 Measuring Misalignment

We model the relationship between supply (number of OSM contributors) and demand (population) using a spatial Durbin model. As previously discussed, spatial dependence is most likely an issue in this dataset. We confirm the presence of spatial dependence by applying the commonly used Lagrange multiplier test, which test for the presence of spatial dependence in the error term as well as whether there is a missing spatially lagged dependent variable. Using this approach, we confirm that there is a statistically significant amount of spatial dependence in the dataset ($P < 0.005$), and conclude that a standard least squares regression is an ill-suited modelling strategy.

We find a positive relationship between population and number of OSM contributors. The estimated coefficients in Table 5.10 are partial derivatives and cannot be interpreted directly [LP09]. Instead we calculate "impacts", which take into account

Table 5.10: Misalignment Model Results

	Estimate	Lag est.
Intercept	0.346	
Population	0.017	-0.012
Rho (spatial lag)	0.74	
AIC	-98,464	

Intercept and population estimates both have $P \ll 0.001$.

Table 5.11: Misalignment Model Impacts

	Direct	**Indirect**	**Total**
Population	0.017	-0.000	0.016

Impacts are estimated based on one million simulations. Direct and Total impacts both have $P \ll 0.001$, Indirect impact has $P = 0.84$.

the spillover effects between tracts and provides interpretable coefficients. Due to the size of our dataset (72,095 census tracts) an exact calculation was unfeasible, and we therefore follow common practice [LP09; Fis11; YNS15] and estimate them using one million simulations. The resulting estimated impacts are listed in Table 5.11, where we see a positive and statistically significant *direct impact*, indicating that if the population in a census tract increases the number of contributors is also expected to increase (positive correlation). A unit increase in population (1,000) results in slightly more than one additional OSM contributor ($10^0.016$). The indirect impact is not statistically significant, meaning that this model cannot determine the effect of an increase in the average population of a census tract's neighbouring tracts. However, as we will see later when additional variables are added, the indirect impact is similar to that of the direct impact.

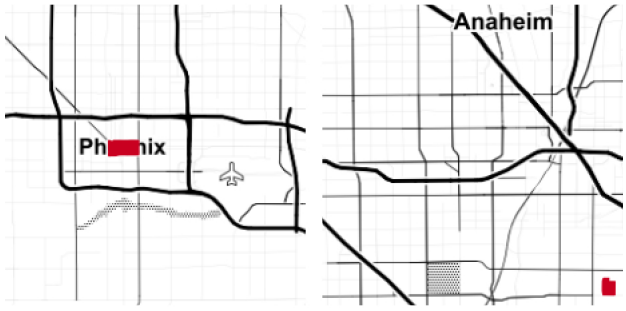

Figure 5.1: Locations of example census tracts. Phoenix, Arizona on the left, a part of the Phoenix/Mesa/Scottsdale metropolitan area. Santa Ana, California on the right, a part of the Los Angeles/Long Beach/Anaheim metropolitan areas.

The spatial dependence in the number of OSM contributors in a census tract is strong. In Table 5.10 we see that "Rho (spatial lag)" is 0.74. This means that if the average number of OSM contributors in neighbouring tracts doubles, the number of contributors in a tract is expected to increase by 67% ($2^{0.74} = 1.67$). Another interpretation of this high rho value is that the variation in number of OSM contributors across the US has a strongly regular pattern.

Misalignment is deviations from perfect alignment in much the same way we measured it in Wikipedia, except we are now operating with a continuous variable. To measure the deviations we use the relative error of the model's predicted values. The variation in number of OSM contributors between tracts is very large, leading to the variable being log-transformed in our models to account for its skewness. Relative error will account for these large differences in magnitude; a prediction of 14 contributors in a tract with 10 has the same relative error as a prediction of 140 contributors in a tract with 100.

To exemplify potential categories of misalignment, Figure 5.1 shows two example census tracts. On the left is a tract in Phoenix, Arizona, and on the right is a tract in Santa Ana, California. The census tract in Phoenix covers much of the

downtown area of the city and has a relative error of 54.4% in the Misalignment Model. A large positive error means the predicted value (29.2 contributors) is much lower than the actual value (64 contributors). As the number of OSM contributors is positively correlated with quality, this tract could potentially be categorised as "excessive quality", similarly to how we used that label in Wikipedia. The other census tract is in Santa Ana, California, a city southeast of Los Angeles, and this census tract is on the opposite side of the spectrum of the one in Phoenix. In the Santa Ana tract the relative error is -139.8% with the predicted number of contributors (28.8) being much higher than the actual (12). If we again carry over the categories from Wikipedia, this tract is one potentially exemplifying "insufficient quality".

The relative error in the Misalignment Model appears to follow a log-normal distribution. This suggests that we might use standard deviations as threshold for categorising misalignment, for example by defining more than two standard deviations as "misaligned". Figure 5.2 shows a histogram of the distribution, where we see the truncated right side due to the maximum relative error potentially being 100% (with a prediction of 0 contributors; the maximum relative error in the Misalignment Model is 96.0%). Based on the shape of the relative error distribution we selected three three potential candidate distributions: Gamma, Weibull, and log-normal. Table 5.12 shows a summary of the results of fitting each of the candidate distributions against the relative error, where we can see that the log-normal distribution has the best fit, which then means that we might be able to categorise misalignment using the standard deviations of that distribution.

The relative errors listed in Table 5.13 show that many areas have large deviations from alignment. On the side of "Excessive Quality" we see that two standard deviations is a relative error of 40.2%, which translates to a census tract with 30

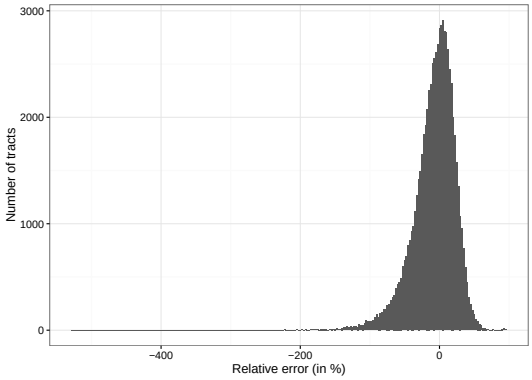

Figure 5.2: Histogram of relative error in percent for the Misalignment Model.

Table 5.12: Goodness-of-fit results for the Misalignment Model

	Gamma	**Weibull**	**Log-normal**
Goodness-of-fit statistics:			
Kolmogorov-Smirnov statistic	0.0335	0.0888	0.0209
Goodness-of-fit criteria:			
AIC	26,089.45	41,405.30	25,736.03

The distributions are fitted against the distribution of relative error shown in Figure 5.2.

contributors being predicted to have about 18, and for three standard deviations that prediction drops to less than 15. Given the log-normal distribution, the deviations are larger on the "Insufficient Quality" side, with a relative error of -81.5% at negative two standard deviations, and -139.8% at negative three. If we again assume a census tract with 30 contributors, it would be predicted to have slightly more than 54 contributors for a relative error of -81.5%, and almost 72 contributors with a relative error of -139.8%. These very large deviations suggest that, similarly to certain topics in Wikipedia, there are certain areas of the map in OSM that have received large amounts of effort, and other areas of the map that appear to be lacking in quality. We will later study the nature of misalignment in more detail as we examine factors that can help explain it, but first we estimate its impact.

Table 5.13: Standard deviations for relative error in the Misalignment Model

	Standard Deviation						
	-3	-2	-1	0	1	2	3
Relative error:	-139.8	-81.5	-37.5	0	21.1	40.2	54.7

Standard deviations are calculated using a log-normal distribution fitted against the relative error, and said error is measured in percent. Mean error is by definition 0%.

5.5.3 Estimating the Impact of Misalignment

We *estimate* the impact of misalignment in OSM by measuring the size of the total population in affected areas. Unlike Wikipedia where we had a direct measure of the number of pageviews, we do not know exactly how many people use OSM and in what capacity. Therefore, we find two standard deviations of the log-normal distribution of relative error to be a reasonable and conservative threshold for labelling census tracts as "misaligned".

Using this labelling strategy, we find that 1,842 census tracts (2.55%) are categorised as "Insufficient Quality", affecting a total population of 7.3 million, and lacking 24,070 contributors. On the opposite side we find 1,452 census tracts (2.01%) are labelled "Excessive Quality", with a total population of 5.7 million, and a surplus of 45,298 contributors.

These results indicate that OSM is more likely to have insufficient quality when misalignment occurs. The difference in proportions (2.55% vs. 2.01%) is statistically significant compared to an expected 50/50 split (Chi-square goodness-of-fit test: $\chi^2 = 46.2, df = 1, p \ll 0.001$). We also see that insufficient quality affects a larger proportion of the US population (7.3 million) than the number of people who live in areas of excessive quality (5.7 million). Lastly, the surplus of contributors in high-

quality areas is much greater than the lack of contributors in low-quality areas. In other words, OSM's contributors appear more eager to make additional contributions to an already crowded area than venture into improving a low-quality area.

These results clearly point out areas where the OSM community has opportunities for improving the VGI quality for a large part of the US population. In the next section, we study these areas in more detail by looking at factors that can help explain why misalignment occurs.

5.5.4 Explanations for Misalignment

Explanation #1: Small Towns Receive Less Attention

The ruralness or urbanness of an area can help explain some of the misalignment in OSM. Like we discussed in our methods section, previous studies on VGI has found higher activity levels and content quality in urban areas [Hak10; ZZ10; HS14a; Joh+16a], although a study focusing on the US state of Florida showed comparatively higher coverage in rural areas [ZH11].

In order to determine the relationship between quality (OSM contributors) and where on the urban-rural spectrum a census tract is, we collapse the RUCA classification into three binary variables: Metropolitan areas (RUCA code 1, 2, and 3; core population above 50,000), Micropolitan areas (RUCA code 4, 5, and 6; core population between 10,000 and 50,000), and Rural areas (RUCA code 10). This leaves small towns (RUCA code 7, 8, and 9; core population between 2,500 and 10,000) as the default label for an area in our model. This approach achieved the best fit and statistical significance for all variables in the model, whereas other approaches did not.

The overall fitness of the model, named "RUCA", is improved when this dataset is added. In other words, adding information about how urban or rural tracts are describes patterns in the distribution of quality in OSM that are not already captured by population data. We can see this reflected in the lower AIC seen in Table 5.14 ($-99,425$) compared to the Misalignment Model's AIC from Table 5.10 ($-98,464$). It is also seen in the slight reduction in the spread of the relative error in the model, shown in Table 5.16 and visualised in Figure 5.4.

We find that how urban/rural a tract is has a curvilinear relationship with OSM quality: urban areas as well as rural areas receive more attention at the cost of small towns (with a core population between 2,500 and 10,000). The direct impacts listed in Table 5.15[17] are positive for all of the urban-rural spectrum variables (metropolitan: 0.046, micropolitan: 0.038, and rural: 0.060). This means that a census tract in either of these areas has higher quality compared to one in small towns (with population held constant). Or conversely, that when it comes to OSM quality, small towns are on average worse off. These findings align with previous research on urban bias and quality in VGI as described earlier. The higher quality in urban (Metropolitan and Micropolitan) areas reflects the higher participation reported in these areas. When it comes to the rural areas, previous work has pointed to data imports as the most likely reason for high coverage in those areas. In our methods section we discussed how these imports would not skew our data, and what we see here is most likely the cleanup efforts that have occurred after the imports, which then draw attention to these rural areas that they otherwise would not have seen.

When it comes to indirect effects of how far along the urban-rural spectrum a

[17]For brevity the table lists impacts for the SES model, which contains all independent variables. We have confirmed that the impacts are similar for the RUCA and STD models.

Table 5.14: Census tract spatial Durbin models

	RUCA		**ST**		**SES**	
	Estimate	Lag est.	Estimate	Lag est.	Estimate	Lag est.
Intercept	0.369		0.396		0.336	
Population	0.017	-0.011	0.010	[1]-0.002	0.009	-0.003
Metropolitan	0.067	-0.090	0.067	-0.088	0.052	-0.066
Micropolitan	0.056	-0.090	0.057	-0.088	0.045	-0.066
Rural	0.071	-0.037	0.065	-0.042	0.057	-0.023
Num. photos			0.062	-0.020	0.061	-0.021
Num. tweets			0.040	-0.047	0.041	-0.044
Median income					0.106	-0.101
Rho (spatial lag)	0.73		0.68		0.70	
AIC	-99,425		-114,750		-116,490	

Abbreviations: RUCA: Urban-rural spectrum; ST: Spatiotemporal dynamics; SES: Socioeconomic status. All estimated coefficients have $P \ll 0.001$ except $^1P = 0.18$. Population is in thousands. Number of photos, tweets, and median income are all log-transformed, as is the dependent variable due to skewed distributions.

Table 5.15: Impacts of spatial Durbin models

	Direct	**Indirect**	**Total**
Population	0.010	0.012	0.021
Metropolitan	0.046	-0.090	-0.044
Micropolitan	0.038	-0.108	-0.070
Rural	0.060	0.051	0.111
Num. photos	0.065	0.066	0.131
Num. tweets	0.038	-0.048	-0.010
Median income	0.100	-0.086	[1]0.014

Impacts are estimated based on one million simulations. All estimated impacts have $P < 0.005$ except $^1P = 0.08$.

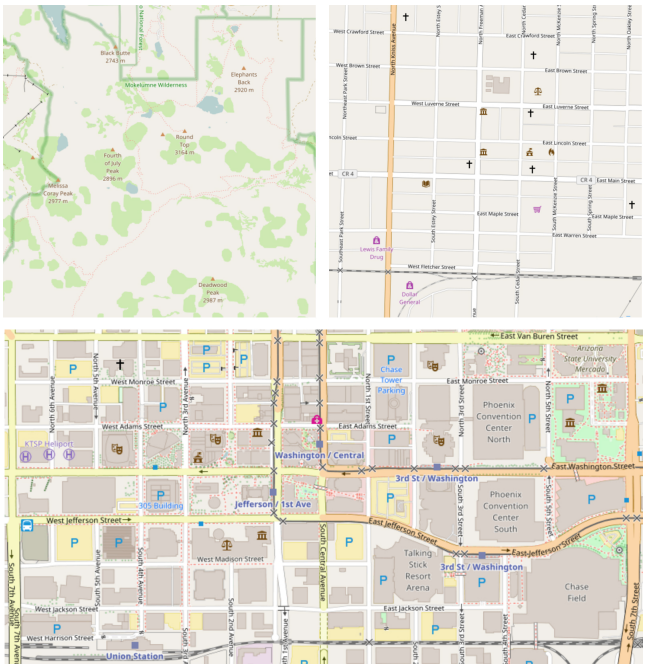

Figure 5.3: Map quality differences across the urban-rural spectrum. Upper left shows a part of Alpine County, California, a rural area with high quality relative to population (e.g. ski lifts, hiking paths, and forested areas). Upper right shows Luverne, Minnesota, a small town with lower quality (e.g. no buildings, only points for churches, a museum, and the post office). Below is Phoenix, Arizona (same as the red area to the left in Figure 5.1), a central urban area with higher quality (e.g. foot paths, trees, heliports).

census tract is, we can see from Table 5.15 that the effects are diverging. Both kind of urban areas have a negative indirect impact (metropolitan: -0.090, micropolitan: -0.108), indicating that if the neighbourhood of a specific census tract is urbanised we would expect the neighbourhood to receive attention and improve quality, at the cost of the given tract. This suggests a competitive relationship for contributor attention on the border between urban areas and small towns. For rural areas we do not see this effect, the indirect impact is instead positive (rural: 0.051), indicating that bordering rural areas is associated with higher quality.

Figure 5.3 exemplifies these differences across the urban-rural spectrum. We iden-

tified three areas with different urban/rural classifications: Alpine County, California, a rural area (RUCA index 10); Luverne, Minnesota, a small town (RUCA index 7); and Phoenix, Arizona, a central urban area (RUCA index 1). In the figure we can see how the rural area in Alpine County has higher map quality than its population would predict, with hiking paths, ski lifts, and forest areas plotted in. The small town of Luverne has its road network entered, but the only buildings shown are there as "points of interest" (e.g. museums and the post office). Contrast that to downtown Phoenix, where we can not only see the outline of all buildings, but also see foot paths, trees, parking ramps, and public transportation.

Coming back to our two example census tracts in Santa Ana and Phoenix, they are both centrally located in a large urban area (RUCA index 1), with all neighbouring tracts also having the same urban classification. Adding data on urban/rural classification to the model has different effects on the two, with the Phoenix, Arizona tract relative error increasing slightly to 55.0%, while the Santa Ana, California tract relative error decreasing slightly to -139.4%. This is in correspondence with our finding that an urban location is associated with higher OSM quality, as in both cases it increases the predicted value. If both tracts were located in small town areas, we would instead see the opposite effect.

Explanation #2: Spatiotemporal Dynamics

Spatiotemporal dynamics can also help explain some of the misalignment. For instance, activity such as attendance of events at the Chase Field baseball stadium in Phoenix, which is located in our example census tract in that city, can explain some of the attention given to that area. There are also areas that receive continued high levels of interests such as tourist destinations. Lastly there are the daily com-

mutes that bring employees to different areas than where they reside. Our Twitter and Flickr datasets can encode these spatiotemporal patterns and thereby help align OSM contributor patterns.

These spatiotemporal patterns are similar to the patterns we previously saw in Wikipedia. Bursts in activity such as sports events can be compared to the surges in demand we studied in Section 5.4.4, although they are likely to occur with more regularity. An example of a less predictable surge in demand for VGI content could be some kind of natural disaster, e.g. flooding. Continued high levels of interests such as tourism are comparable to for instance the pop culture topics we found in Section 5.4.3.

The overall fitness of the model, named "ST", is greatly improved when these datasets are added to the model. Similarly as we found for the RUCA model, this means that these datasets encode patterns in OSM quality that are not captured by the other two independent variables. This is reflected in the lower AIC seen in Table 5.14 ($-114,750$) compared to the RUCA Model's AIC ($-99,425$). We can also see a reduction in the spread of the relative error in the model, shown in Table 5.16 and visualised in Figure 5.4.

We find that contribution patterns in OSM align with spatiotemporal dynamics found in our Flickr and Twitter datasets and have a positive correlation with OSM quality. In Table 5.15 we can see that both tweets and images have a positive direct impact. When it comes to indirect impacts, the effect of an increase in photos/tweets in other tracts, the results diverge. Images, which we interpret as encoding tourism, have a positive indirect impact, meaning that more activity on Flickr will always result in higher OSM quality. With tweets, the relationship is instead negative, suggesting a competitive relationship as we also saw for urban areas. Since our study

Table 5.16: Standard deviations for relative error in all models

Model	**Standard Deviation**						
	-3	-2	-1	0	1	2	3
Misalignment	-139.8	-81.5	-37.5	0	21.1	40.2	54.7
RUCA	-138.6	-81.0	-37.3	0	21.0	40.0	54.5
ST	-121.2	-72.1	-33.9	0	18.9	36.9	50.9
SES	-118.6	-70.8	-33.4	0	18.6	36.4	50.4

Abbreviations: RUCA: Urban-rural spectrum; ST: Spatiotemporal dynamics; SES: Socioeconomic status. Standard deviations are calculated using a log-normal distribution fitted against the relative error, and said error is measured in percent. Mean error is by definition 0%.

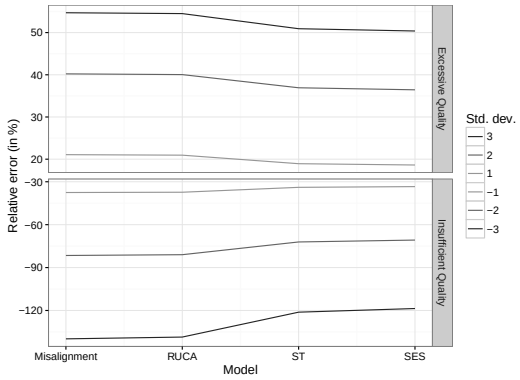

Figure 5.4: Plot of the relative error for each standard deviation on a log-normal scale for each of our four spatial Durbin models.

is correlational we cannot draw causal conclusions, but future work could seek to untangle these effects in more detail.

Spatiotemporal dynamics have a similar effect on our example tracts. The tract in Phoenix, Arizona, which we described as potentially exemplifying "Excessive Quality", had a relative error of 55.0%, which is 3 standard deviations of our log-normal distribution. With the additional data added to our model the relative error decreases to 29.6%. Being located in downtown Phoenix, this census tract has relatively low population (2,406), but is in an area with high dynamism in activity, for instance

because people commute there to work and it has a sports arena. This can explain the very high engagement in the Flickr and Twitter datasets (almost 10,000 images and more than 10,000 tweets) compared to the global averages (28.5 images and 384.4 tweets[18]). Our example tract in Santa Ana, California also benefits from these added datasets with the relative error dropping from -139.8% to -78.2% due to no images and few tweets (45) being located in this census tract. In both cases we see that spatiotemporal dynamics in the population as found in these datasets accounts for some of the excessive or insufficient OSM activity in these areas.

Explanation #3: Richer Areas Receive More Attention

Socioeconomic status (SES) can also explain some of the misalignment found in OSM. It could for instance be that contributors need to invest in technology (e.g. GPS units) thereby requiring a certain amount of disposable income. Participation also requires time available for it. Lastly, it could be that contributors shy away from lower SES areas as they regard them as "unsafe". As we discussed in our methods section, research on OSM quality found that lower SES areas had lower quality [Hak10], and participants in the "gig economy" avoided jobs in lower SES areas [TTH15].

A similar type of sociocultural effect is also seen in Wikipedia, where the proportion of women in the contributor community has been found to be only about 16% [Lam+11; HS13]. This has in turn been connected with a difference in the quality of articles on some topics such as movies depending on the gender balance of their audience [Lam+11]. Volunteer projects have also been started in order to fill a void in the article space on certain topics, for instance about women scientists [Bro16]. While we in OSM study a different effect, if lower SES areas consistently receive less

[18]Geometric means due to skewed distributions, these are close to the medians of both measures.

attention, similar types of targeted efforts could be initiated in order to alleviate the problem.

The overall fitness of the model is again improved slightly, and similarly as for the RUCA and ST models this means that data on socioeconomic status encodes patterns in OSM quality that is not captured in the other models. We can see in Table 5.14 the AIC is lower ($-116,490$) compared to the ST model ($-114,750$). The spread in the relative error also decreases as shown in Table 5.16 and Figure 5.4. The effect of adding data on socioeconmic status is reduced by the fact that our Flickr and Twitter datasets also encode some of that information [LGX13].

We find that socioeconomic status aligns with OSM contributor patterns and has a positive correlation with OSM quality. The direct impact of socioeconomic status is positive as seen in Table 5.15, thereby indicating that if the socioeconomic status of a census tract increases it is expected to see an increase in OSM quality (positive correlation). Similarly as for our Twitter dataset, we see that SES has a negative direct impact, in order words that it is more of a competitive relationship. If the average income in other increases, we would expect the number of OSM contributors in the tract to *decrease*.

In general, these results suggests that OSM in the US, similarly to what Haklay previously found for England, also is likely to suffer from lower quality in low-SES areas. Previously we pointed to there being a larger surplus of contributors in areas with excessive quality, indicating "preferential attachment" as contributors make additional contributions to areas that already have high quality. With OSM content being reused, for instance in smartphone applications, this type of sociocultural bias in the quality of information is a serious issue that the OSM community should seek to combat.

For our example census tracts, data on socioeconomic status follows the overall trend and modestly decreases the relative error. The tract in Phoenix, Arizona is a relatively lower income area (median of USD 22,244 compared to the overall US median of USD 52,250), and the neighbouring tracts also have income below the US median with a mean of USD 36,600. This results in a slight decrease in this tract's relative error from 29.6% to 29.5%. Our example tract from Santa Ana, California has median income close to the overall median, at USD 54,802. In this case the tract's neighbourhood have comparatively much higher income, with a mean of USD 69,095. While the higher income in the neighbourhood is associated with lower OSM activity in a given tract as we would expect the neighbouring tracts to draw attention away, we also have high spatial dependency in the model (ref Table 5.14, $\rho = 0.70$), meaning that higher OSM participation in neighbouring tracts will spill over. This can then explain why the relative error in our example tract increases from -78.2% to -80.4%, as the tract is predicted to have slightly higher participation once socioeconomic status is added.

This concludes our investigation of the extent and impact of misalignment in OpenStreetMap and answers our research questions. We modelled the relationship between supply and demand of quality content in OSM using a spatial model and studied how this relationship differs from perfect alignment. As we have found, several million people in the US live in areas that appear to suffer from lack of OSM quality, and these areas are more likely to have lower socioeconomic status and be in small towns. We now turn our attention to discussing the implications of our findings for both Wikipedia and OpenStreetMap, as well as other peer production communities.

5.6 Discussion

Through our modelling of the *Perfect Alignment Hypothesis*, we investigated misalignment between the supply and demand of quality content in peer production communities. Across all four Wikipedia language editions 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. In OpenStreetMap we found extensive misalignment as well, and when it occurs it is significantly more likely to be in the form of insufficient quality. This misalignment has a big impact on content consumers: in Wikipedia several billion monthly views are to articles that would be of considerably higher quality if quality supply and demand were more in alignment; in OpenStreetMap our conservative estimate is that a total population of several million live in areas of insufficient quality.

Our coverage of previous research pointed to studies showing that contributors are primarily motivated by "fun" [LW05; Nov07] and altruism [BH13]. Previous work has also identified several types of biases in peer produced content [HG09; Hak10; HG10; Lam+11; RR11; Ste13]. 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? The mission of Wikipedia and OpenStreetMap are both similar; Wikipedia aims to give everyone access to the world's knowledge[19],

[19] `https://en.wikipedia.org/wiki/Wikipedia:Purpose`

OpenStreetMap aims to "make the best map data set of the world"[20]. The PAH argues that Wikipedia should focus its attention where the readers are, which for instance means that expansion of coverage in the form of new articles cannot occur before alignment is achieved. In OpenStreetMap, perfect alignment means that quality follows population. Our results pointed to quality issues in small towns, where achieving alignment is challenged by the fact that there are more geographic entities per capita than in urban areas [Joh+16a]. While our analysis has revealed areas in both communities where attention is needed, areas where for instance newcomers who are eager to make a worthwhile contribution could be directed, these communities have to figure out how to best serve their purposes.

We realise that the success of peer production communities has been driven largely by very prolific editors [Kit+07; Pri+07; NZ12] who maintain high levels of activity throughout their lifetime in the system [PHT09]. 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 a direction of future work that emerges from our findings.

What should the locus of such mechanisms be? Self-organised groups of volunteers who express interest in improving content on a particular topical or geographical area, the "WikiProjects" we used in our topic analysis in Wikipedia, or "mapping parties" in OpenStreetMap [HW08; Hri+13], are both intriguing possibilities. Previous work has shown that WikiProjects have goal-setting mechanisms that can motivate contributors towards group efforts [ZKK12], but recall that we found that even within the scope of a project, misalignment occurs. However, some of these groups have cre-

[20] `https://wiki.osmfoundation.org/wiki/Mission_Statement`

ated hundreds of top-quality articles [For+12], meaning their members have acquired domain-specific knowledge that can benefit other groups as well. To address this, a tool like SuggestBot [Cos+07] could be modified to suggest good candidate (i.e. low-quality/high-demand Wikipedia articles). In OpenStreetMap, a similar type of tool could suggest geographic areas and objects in need of improvement, an approach that has been found to be effective in soliciting VGI contributions [PMT10]. If the tool also identifies and suggests candidate sets of editors with the required range of topic/area and community expertise, it could enable 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 [KGC13], 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.

5.7 Future Work and Limitations

We used Wikipedia's own assessment scale to measure article 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 [Stv+08a], 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 re-assessment delays themselves are a form of misalignment, which deserves further study.

To measure demand in Wikipedia, 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[21], 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.

Due to the lack of a central source for demand data such as tile views in OpenStreetMap we used population to estimate demand, and supported this with additional datasets to account for other factors such as spatiotemporal dynamics and socioeconomic status. Similarly, because OSM does not have content quality labels and estimating it through computation is non-trivial, we proxied quality through number of OSM contributors to an area. Future work should use other data sources or different methods in order to confirm our findings, measure the impact of insufficient quality, study short-term patterns in misalignment, as well as answer new research questions in this area.

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

5.8 Conclusion

In this chapter, we studied alignment between supply and demand of quality content in peer production communities in the context of four large Wikipedia editions and OpenStreetMap, reaching the following conclusions:

1. Reader demand and contributor supply for high-quality content in Wikipedia 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 topics, e.g. countries and sensitive topics, are over-represented amongst articles in high demand but of low quality.

4. The misalignment of Wikipedia 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.

5. There is also a great deal of misalignment in OpenStreetMap, where certain areas of the map indicates low demand but very high quality, but also the opposite.

6. We conservatively estimate that a population of at least 7.3 million live in areas labelled "Insufficient Quality", and these areas are more likely to be small towns and with lower socioeconomic status.

We situated these findings in relation to previous research and discussed how we have established the presence of another issue that appears to be prevalent in peer production communities. This opens up the opportunity for several future design and research projects aiming to alleviate a problem that, as we have seen, affects a huge number of people every day. In the next chapter, we will discuss these findings in a larger context and describe some of the future work that is motivated by our work.

Chapter 6

Conclusion

In this thesis, we have studied the production and consumption of quality content in peer production communities. We have furthered our understanding of how these communities work and identified areas where there is room for improvement by making several contributions:

- A quality prediction model for Wikipedia that has high performance and is considered the state of the art. This model focuses on aspects of content quality that contributors can easily improve, so called "actionable features". By focusing on these features we argue for continued work in applying machine learning to quality prediction in ways that emphasise the content itself instead of developing increasingly intricate algorithms that focus on who created the content. Through collaborations with developers at the Wikimedia Foundation this prediction model is now widely deployed and used to build tools to support the Wikipedia community by making quality assessment readily available in multiple languages.

- A framework for describing and analysing quality improvement projects in peer production communities that allows us to discover factors associated with the success or failure of these types of projects. We utilise this framework to study five diverse quality improvement projects in the English Wikipedia, and report

three findings; first, that an increasing number of contributors is associated with a slower increase in quality; secondly, that new artifacts are associated with higher quality than improving an existing artifact; and lastly, that a lack of proper structure (e.g. incentive mechanisms) is connected with failing improvement projects. These findings provide community organisers with useful guidance when designing these types of projects, both within the Wikipedia sphere as well as outside of it.

- An analytical framework for studying the relationship between the production and consumption of quality content in peer production communities. We use this framework called the *Perfect Alignment Hypothesis* to study successful communities and discover a great deal of misalignment between supply and demand. This framework also allows us to measure the impact of the misalignment on consumers, where we find that around two billion monthly article views in Wikipedia go to articles that would be of much higher quality if supply and demand were in alignment. Similarly in OpenStreetMap, we estimate that 7.3 million people in the US live in areas of lower quality than demand suggests. We also study characteristics of areas in which misalignment occurs, finding that certain Wikipedia topics (e.g. LGBT issues) are more likely to have lower quality than demand suggests, and that low-income areas and small towns are more likely to have lower map quality in OpenStreetMap than demand indicates. Lastly, we also show that misalignment in Wikipedia is not driven only by short-term demand, about half of the most popular articles that are not top quality show signs of stable high demand. These findings establish misalignment as an issue that peer production community organisers should be

aware of as it clearly has a significant impacts on consumers. In our work we discuss sociotechnical solutions that can help balance the interest of peer production contributors, who are typically volunteers, with a community's interest of having the greatest impact.

6.1 Broader Implications

The content that is created by peer production communities is widely consumed. As we saw in our analysis of the impact of misalignment in Wikipedia, just views of articles that are of lower quality than demand suggests is on the order of *billions*, and in OpenStreetMap we estimated a population impact of 7.3 million people just in the US. Content from both communities is also widely reused, for example maps in OpenStreetMap being used in smartphone applications through Mapbox. In other words, what happens in these communities *matters* and affects a large audience.

Research on peer production communities has discovered several different types of bias in peer produced content (e.g. gender bias). While misalignment is not a bias per se, it is clearly an issue that affects peer produced content. It is also worth noting that our analytical framework for misalignment has allowed us to calculate (or estimate, in the case of OpenStreetMap) the impact this issue has, whereas for some of the other issues the impact is currently less well-defined.

Peer production is not going anywhere soon. Yochai Benkler has argued that it has clear benefits over other approaches such as traditional firms and markets [Ben02], meaning that it would be irrational to seek to dismantle peer production communities. A question raised by our work is whether it is possible to blend peer production's volunteer-based self-selection with a more traditional notion of supply meeting de-

mand in order for these communities to increase their impact.

Increased impact on content consumers is the underlying assumption behind our work on misalignment. This assumption does have certain biases that need consideration. For instance, in our study of OpenStreetMap we defined demand using population, which results in a bias towards larger cities because they are more densely populated. Our analysis of overrepresented topics amongst highly popular articles in Wikipedia discovered many related to pop culture, suggesting that they should receive more attention from contributors. In essence, this makes a statement about what we should regard as *important* content. Is it more important to have a top quality article about a contemporary popular artist, or should the community develop a top quality article about a historic figure in classical music instead? When we study misalignment, we argue that the content that is *consumed* is more important and that the communities have not historically paid enough attention towards that type of content. At the same time, we return to the fact that these communities are volunteers, meaning a balance between self-interest and the interest of the community is necessary.

This balance between self-interest and the interest of the community also affects sub-communities. Some of the WikiProjects we studied in Wikipedia have a narrow scope but are very successful, and they might not be sustainable unless they can keep their focus. We brought up the example of the military history WikiProject producing a larger number of high quality articles with a narrow audience, but it is also a project that is highly successful. The overall impact of a peer production community's content might therefore again be limited by having a balance between a sub-community's interests and what would benefit the overall community the most.

The communities we have studied are all well established and successful with

thousands of volunteer contributors. It is worth asking how our findings affect smaller communities. For instance, the Perfect Alignment Hypothesis assumes that the most popular content would be of top quality, which in turn means that until there is alignment there cannot be new content added. In Wikipedia, this would mean that a new article cannot be added until there is alignment amongst the existing ones. Historically, Wikipedia had a focus on creating new articles, so much that Jimmy Wales suggested in his opening plenary at the Wikimania conference in 2006 that the community shift its focus towards working on the quality of existing articles[1]. Similarly as we discussed a trade off between coverage and quality in Chapter 4 after finding that new articles achieve higher quality, a community that is just getting started might need to focus on coverage instead of alignment in order to grow and become successful.

Our work focuses on *commons-based* peer production, but likely impacts any community where contributors are volunteers and tasks are self-selected. These communities could be ones that are not commons-based, in other words the content is not shared using an open license like Wikipedia and OpenStreetMap use. It could also be that these communities do not use peer production, meaning that a piece of content would have a limited number of creators. In both of these cases, our work on using machine learning to determine content quality could be applied and used to guide contributors. Our framework for describing and analysing quality improvement projects could be adapted to these types of settings and be used to understand the conditions for success. Lastly, our work on misalignment again applies as the issue is most likely the consequence of task self-selection.

[1] `https://wikimania2006.wikimedia.org/wiki/Opening_Plenary_(transcript)#Quality_initiative_.2833:20.29`

The bottom line is that the work undertaken in this thesis has increased our understanding of peer produced content quality, how these communities work toward achieving it, and how there is a possibility for increasing the community's impact on content consumers by balancing the interests of the contributors with that of the consumers. Our findings also open up important venues for future studies, which we will discuss next.

6.2 Future Work

Our study of misalignment between the consumption and production of quality content has identified an issue, but we do not know much about the root cause of it, nor do we know to what extent it is possible to alter contributor behaviour. We pointed to contributor motivation as a potential cause, but we would like to be more certain. An important first step would therefore be to survey contributors to these communities to understand more about their reasoning around these issues. However, survey responses might be deceiving, meaning we would also want to design experiments that can reveal underlying attitudes and assumptions, similar to the Implicit Association Test in social psychology [GMS98]. This would then hopefully allow us to understand more about how to balance individual autonomy and interests in certain areas with work that has the most benefit to the community, since as we discussed in Chapter 5 because contributors are generally volunteers we cannot simply tell them what to do.

Once we know more about contributors we can design interventions that seek to alter the production side of the misalignment issue. It could come in the form recommendations of high-impact work, thereby making it easier for contributors to find work that would benefit the community. In Wikipedia, we have been in-

volved with recommending articles to contributors for several years through SuggestBot[2] [Cos+07], and that could be used as a platform for experiments seeking to alter contributor behaviour. Secondly, these communities rarely make consumption salient. If you go to an article on Wikipedia, it is not easy to find out how popular it is, but it is possible to discover[3]. Learning how popular an area is in OpenStreetMap is most likely even more difficult due to how decentralised its usage is. Contributors to these communities therefore have a hard time understanding what the impact of their contributions will be, meaning their decisions will most likely not be based on said impact.

Reducing misalignment might also happen through altering the consumption side. Currently we know little about what consumers of peer produced content do as research on peer production communities have been more focused on the contributor side. Some studies are being done to understand more about Wikipedia's readers[4], but we know very little about consumers of OpenStreetMap content. Further work that can provide us with knowledge on the behaviour of content consumers can point out where there might be opportunities to steer behaviour towards content of high quality that is currently being consumed less. The next step would then be to aim to design and run interventions that would point consumers towards those types of content. For example, several of the different language editions of Wikipedia feature a specific article on their front page and we know that this drives traffic to these articles [Thi+12]. While that is one potential venue for future work, there might also be possibilities of cross-community work, for example by using recommender systems

[2] `https://en.wikipedia.org/wiki/User:SuggestBot`

[3] `http://tools.wmflabs.org/pageviews/` is as of November, 2016, one place to find this information.

[4] `https://lists.wikimedia.org/pipermail/wiki-research-l/2016-November/005500.html` points to a presentation of the Wikimedia Foundation's work on Wikipedia's readers.

to promote high-quality Wikipedia content in other communities.

There is also future work to be done in supporting quality improvement projects. Inside the Wikipedia universe, our work has resulted in the availability of an API for article quality prediction. This facilitates experimenting with building sociotechnical tools that can enable individual or groups of contributors to make better decisions about where to improve quality. While we know quite a lot about these types of projects in Wikipedia, much less is known about how they work in other communities. For instance, as mentioned in our discussion in Chapter 5 OpenStreetMap has "mapping parties", which there have been a few studies of. With more research and potentially experiments, we would be able to understand to what extent that is an effective way or improving quality. There might be similar types of approaches in other communities such as those in open source software as well, thus providing us with another possible line of work.

Lastly, there are opportunities for improving our understanding of how quality assessment is done in peer production communities. Wikipedia's quality assessments have been used extensively in the literature to train machine learners as we covered in Chapter 2, but the actual assessments have received little focus. For instance, we know little about how consistent the ratings are, both whether a contributor rates articles of similar quality consistently, and whether different contributors rate the same article consistently. Studying this might provide us with solutions that can help provide us with better training data, which in turn can further improve our machine learners. In OpenStreetMap, studies of quality have focused on quantitative measures that provide a general overview. Ways of assessing quality and presenting those assessment to content consumers and contributors has received little attention. Is there a way to categorise an area of a map as low, medium, or high quality? And if

so, how do you present that to users in a meaningful way? Similarly to how Wikipedia had an explosive growth from 2005 onwards and had to deal with handling quality assurance [Hal+13], OpenStreetMap might reach the same state and have to invent new techniques, providing us with an opportunity for future work.

6.3 In Conclusion

The research presented in this thesis provides us with a deeper understanding of peer production communities. These communities have created software, encyclopedic content, and maps that in many ways improve our everyday lives as well as those of millions of others. At the same time we have identified areas where there's a shortage of quality content and discussed future work that can help reduce this problem. This thesis has thus laid the foundation upon which we can build improved communities and positively impact a large part of the world's population. We look forward to continuing this work and seeing what future peer production communities will bring.

Bibliography

[AD07] B Thomas Adler and Luca De Alfaro. "A content-driven reputation system for the Wikipedia". In: *Proc. WWW*. 2007.

[AS12] Maik Anderka and Benno Stein. "Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia". In: *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*. 2012.

[ASL12] Maik Anderka, Benno Stein, and Nedim Lipka. "Predicting quality flaws in user-generated content: the case of Wikipedia". In: *Proc. SIGIR*. Portland, Oregon, USA, 2012, pp. 981–990. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348413. URL: http://doi.acm.org/10.1145/2348283.2348413.

[And+13] Ashton Anderson et al. "Steering User Behavior with Badges". In: *Proc. of WWW*. 2013, pp. 95–106.

[AKK14] Paul André, Robert E. Kraut, and Aniket Kittur. "Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks". In: *Proc. of CHI*. Toronto, Ontario, Canada, 2014, pp. 139–148. ISBN: 978-1-4503-2473-1. DOI: 10.1145/2556288.2557158. URL: http://doi.acm.org/10.1145/2556288.2557158.

[AN10] Ofer Arazy and Oded Nov. "Determinants of Wikipedia quality: the roles of global and local contribution inequality". In: *Proc. CSCW*. Savannah, Georgia, USA, 2010, pp. 233–236. ISBN: 978-1-60558-795-0. DOI: 10.1145/1718918.1718963. URL: http://doi.acm.org/10.1145/1718918.1718963.

[Ara+15] Ofer Arazy et al. "Functional Roles and Career Paths in Wikipedia". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. Vancouver, BC, Canada: ACM, 2015, pp. 1092–1105. ISBN: 978-1-4503-2922-4. DOI: 10.1145/2675133.2675257. URL: http://doi.acm.org/10.1145/2675133.2675257.

[Ard08] Alexandre Ardichvili. "Learning and Knowledge Sharing in Virtual Communities of Practice: Motivators, Barriers, and Enablers". In: *Advances in Developing Human Resources* 10.4 (2008), pp. 541–554.

[Ars+13] Jamal Jokar Arsanjani et al. "Assessing the Quality of OpenStreetMap Contributors together with their Contributions". In: *Proc. of the 16th AGILE International Conference on Geographic Information Science*. 2013.

[BNZ14] Christopher Barron, Pascal Neis, and Alexander Zipf. "A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis". In: *Transactions in GIS* 18.6 (2014), pp. 877–895. ISSN: 1467-9671. DOI: 10.1111/tgis.12073. URL: http://dx.doi.org/10.1111/tgis.12073.

[Ben02] Yochai Benkler. "Coase's Penguin, or, Linux and "The Nature of the Firm"". In: *Yale Law Journal* (2002), pp. 369–446.

[BHK13] Roger Bivand, Jan Hauke, and Tomasz Kossowski. "Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods". In: *Geographical Analysis* 45.2 (2013), pp. 150–179. URL: http://www.jstatsoft.org/v63/i18/.

[BP15] Roger Bivand and Gianfranco Piras. "Comparing Implementations of Estimation Methods for Spatial Econometrics". In: *Journal of Statistical Software* 63.18 (2015), pp. 1–36. URL: http://www.jstatsoft.org/v63/i18/.

[Blu08] Joshua E. Blumenstock. "Size matters: word count as a measure of quality on Wikipedia". In: *Proc. WWW*. 2008.

[Bra+09] Ulrik Brandes et al. "Network analysis of collaboration structure in Wikipedia". In: *Proc. WWW*. Madrid, Spain, 2009, pp. 731–740. ISBN: 978-1-60558-487-4. DOI: 10.1145/1526709.1526808. URL: http://doi.acm.org/10.1145/1526709.1526808.

[BD02] Peter J Brockwell and Richard A Davis. *Introduction to Time Series and Forecasting*. 2nd ed. Vol. 1. Springer, 2002.

[Bro75] Frederick P Brooks. *The Mythical Man-Month*. Vol. 1995. Addison-Wesley Reading, 1975.

[Bro16] Taylor Kate Brown. *Female scientist fights harassment with Wikipedia*. Ed. by BBC Trending. Online, accessed 1 Sept 2016. Mar. 2016. URL: http://www.bbc.com/news/blogs-trending-35787730.

[BH13] Nama R. Budhathoki and Caroline Haythornthwaite. "Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap". In: *American Behavioral Scientist* 57.5 (2013), pp. 548–575. DOI: 10.1177/0002764212469364. eprint: http://abs.sagepub.com/content/57/5/548.full.pdf+html. URL: http://abs.sagepub.com/content/57/5/548.abstract.

[BK08] Moira Burke and Robert Kraut. "Mopping Up: Modeling Wikipedia Promotion Decisions". In: *Proc. of CSCW*. San Diego, CA, USA, 2008, pp. 27–36. ISBN: 978-1-60558-007-4. DOI: 10.1145/1460563.1460571. URL: http://doi.acm.org/10.1145/1460563.1460571.

[CRR10] Jilin Chen, Yuqing Ren, and John Riedl. "The Effects of Diversity on Group Productivity and Member Withdrawal in Online Volunteer Groups". In: *Proc. of CHI*. 2010.

[Che06] Thomas Chesney. "An empirical examination of Wikipedia's credibility". In: *First Monday* 11.11 (2006), pp. 1–13.

[Cho+10] Boreum Choi et al. "Socialization Tactics in Wikipedia and Their Effects". In: *Proc. of CSCW*. 2010, pp. 107–116.

[Cie+10] Błażej Ciepłuch et al. "Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps". In: *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*. University of Leicester. 2010.

[Cla+08] Kevin A Clauson et al. "Scope, Completeness, and Accuracy of Drug Information in Wikipedia". In: *Annals of Pharmacotherapy* 42.12 (2008), pp. 1814–1821.

[Cos+06] Dan Cosley et al. "Using Intelligent Task Routing and Contribution Review to Help Communities Build Artifacts of Lasting Value". In: *Proc. of CHI*. CHI '06. Montréal, Québec, Canada, 2006, pp. 1037–1046. ISBN: 1-59593-372-7. DOI: 10.1145/1124772.1124928. URL: http://doi.acm.org/10.1145/1124772.1124928.

[Cos+07] Dan Cosley et al. "SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia". In: *Proc. IUI*. Honolulu, Hawaii, USA, 2007, pp. 32–41. ISBN: 1-59593-481-2. DOI: 10.1145/1216295.1216309. URL: http://doi.acm.org/10.1145/1216295.1216309.

[DI16] Quang Vinh Dang and Claudia-Lavinia Ignat. "Quality Assessment of Wikipedia Articles Without Feature Engineering". In: *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. JCDL '16. Newark, New Jersey, USA: ACM, 2016, pp. 27–30. ISBN: 978-1-4503-4229-2. DOI: 10.1145/2910896.2910917. URL: http://doi.acm.org/10.1145/2910896.2910917.

[DCT98] Huw Talfryn Oakley Davies, Iain Kinloch Crombie, and Manouche Tavakoli. "When can odds ratios mislead?" In: *BMJ* 316.7136 (1998), pp. 989–991. DOI: 10.1136/bmj.316.7136.989.

[DD10] Gabriel De la Calzada and Alex Dekhtyar. "On measuring the quality of Wikipedia articles". In: *Proc. WICOW*. Raleigh, North Carolina, USA, 2010, pp. 11–18. ISBN: 978-1-60558-940-4. DOI: 10.1145/1772938.1772943. URL: http://doi.acm.org/10.1145/1772938.1772943.

[Det+11] Sebastian Deterding et al. "From Game Design Elements to Gamefulness: Defining "Gamification"". In: *Proc. of Mindtrek*. 2011, pp. 9–15.

[ER09] Michael D. Ekstrand and John T. Riedl. "rv you're dumb: identifying discarded work in Wiki article history". In: *Proc. WikiSym*. 2009, 4:1–4:10.

[EH13] Eirik Emanuelsen and Martin Rudi Holaker. "Event Detection using Wikipedia". MA thesis. Norwegian University of Science and Technology, 2013.

[FK13] Rosta Farzan and Robert E. Kraut. "Wikipedia Classroom Experiment: Bidirectional Benefits of Students' Engagement in Online Production Communities". In: *Proc. of CHI*. 2013.

[Fis11] Manfred M. Fischer. "A spatial Mankiw–Romer–Weil model: theory and evidence". In: *The Annals of Regional Science* 47.2 (2011), pp. 419–436. ISSN: 1432-0592. DOI: 10.1007/s00168-010-0384-6. URL: http://dx.doi.org/10.1007/s00168-010-0384-6.

[FVS12] Fabian Flöck, Denny Vrandečić, and Elena Simperl. "Revisiting reverts: accurate revert detection in Wikipedia". In: *Proc. HT*. Milwaukee, Wisconsin, USA, 2012, pp. 3–12. ISBN: 978-1-4503-1335-3. DOI: 10.1145/2309996.2310000. URL: http://doi.acm.org/10.1145/2309996.2310000.

[FA82] Robin Flowerdew and Murray Aitkin. "A METHOD OF FITTING THE GRAVITY MODEL BASED ON THE POISSON DISTRIBUTION". In: *Journal of Regional Science* 22.2 (1982), pp. 191–202. ISSN: 1467-9787. DOI: 10.1111/j.1467-9787.1982.tb00744.x. URL: http://dx.doi.org/10.1111/j.1467-9787.1982.tb00744.x.

[For+12] Andrea Forte et al. "Coordination and Beyond: Social Functions of Groups in Open Content Production". In: *Proc. of CSCW*. 2012, pp. 417–426.

[GR10] R. Stuart Geiger and David Ribes. "The work of sustaining order in Wikipedia: the banning of a vandal". In: *Proc. CSCW*. 2010, pp. 117–126. ISBN: 978-1-60558-795-0. DOI: 10.1145/1718918.1718941. URL: http://doi.acm.org/10.1145/1718918.1718941.

[Gil05] Jim Giles. "Internet encyclopaedias go head to head". In: *Nature* 438.7070 (2005), pp. 900–901.

[GT10] Jean-François Girres and Guillaume Touya. "Quality Assessment of the French OpenStreetMap Dataset". In: *Transactions in GIS* 14.4 (2010), pp. 435–459. ISSN: 1467-9671. DOI: 10.1111/j.1467-9671.2010.01203.x. URL: http://dx.doi.org/10.1111/j.1467-9671.2010.01203.x.

[Goo07] MichaelF. Goodchild. "Citizens as sensors: the world of volunteered geography". English. In: *GeoJournal* 69.4 (2007), pp. 211–221. ISSN: 0343-2521. DOI: 10.1007/s10708-007-9111-y. URL: http://dx.doi.org/10.1007/s10708-007-9111-y.

[Gor11] Andreea D Gorbatai. "Exploring Underproduction in Wikipedia". In: *Proc. of WikiSym*. 2011, pp. 205–206.

[Gor14] Andreea D Gorbatai. "The Paradox of Novice Contributions to Collective Production: Evidence from Wikipedia". In: *Available at SSRN 1949327* (2014).

[GMS98] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. "Measuring individual differences in implicit cognition: The implicit association test." In: *Journal of personality and social psychology* 74.6 (1998), p. 1464.

[GBR14] Simon Gröchenig, Richard Brunauer, and Karl Rehrl. "Estimating Completeness of VGI Datasets by Analyzing Community Activity Over Time Periods". In: *Connecting a Digital Europe Through Location and Place*. Ed. by Joaquín Huerta, Sven Schade, and Carlos Granell. Cham: Springer International Publishing, 2014, pp. 3–18. ISBN: 978-3-319-03611-3. DOI: 10.1007/978-3-319-03611-3_1. URL: http://dx.doi.org/10.1007/978-3-319-03611-3_1.

[Häg68] Torsten Hägerstrand. *Innovation diffusion as a spatial process*. University of Chicago Press, 1968.

[HW08] M. Haklay and P. Weber. "OpenStreetMap: User-Generated Street Maps". In: *IEEE Pervasive Computing* 7.4 (Oct. 2008), pp. 12–18. ISSN: 1536-1268. DOI: 10.1109/MPRV.2008.80.

[Hak10] Mordechai Haklay. "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning. B, Planning & design* 37.4 (2010), p. 682.

[Hak+10] Mordechai Haklay et al. "How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information". In: *The Cartographic Journal* 47.4 (2010), pp. 315–322.

[HKT13] Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. "Making Peripheral Participation Legitimate: Reader engagement experiments in Wikipedia". In: *Proc. of CSCW*. 2013, pp. 849–860. DOI: 10.1145/2441776.2441872.

[Hal+09] Aaron Halfaker et al. "A jury of your peers: quality, experience and ownership in Wikipedia". In: *Proc. WikiSym*. Orlando, Florida, 2009, 15:1–15:10. ISBN: 978-1-60558-730-1. DOI: 10.1145/1641309.1641332. URL: http://doi.acm.org/10.1145/1641309.1641332.

[Hal+13] Aaron Halfaker et al. "The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline". In: *American Behavioral Scientist* 57.5 (2013), pp. 664–688. DOI: 10.1177/0002764212469365. eprint: http://abs.sagepub.com/content/57/5/664.full.pdf+html. URL: http://abs.sagepub.com/content/57/5/664.abstract.

[Has+09] Daniel Hasan Dalip et al. "Automatic quality assessment of content created collaboratively by web communities: a case study of Wikipedia". In: *Proc. JCDL*. 2009, pp. 295–304.

[Hec13] Brent Hecht. "The Mining and Application of Diverse Cultural Perspectives in User-Generated Content". PhD thesis. Northwestern University, Mar. 2013.

[HG09] Brent Hecht and Darren Gergle. "Measuring self-focus bias in community-maintained knowledge repositories". In: *Proc. C&T*. 2009, pp. 11–20.

[HG10] Brent Hecht and Darren Gergle. "The tower of Babel meets web 2.0: User-generated content and its applications in a multilingual context". In: *Proc. CHI*. 2010, pp. 291–300.

[HS14a] Brent Hecht and Monica Stephens. "A Tale of Cities: Urban Biases in Volunteered Geographic Information". In: *Proc. of ICWSM*. 2014.

[Hei+11] James M Heilman et al. "Wikipedia: A Key Tool for Global Public Health Promotion". In: *Journal of Medical Internet Research* 13.1 (2011).

[HS13] Benjamin Mako Hill and Aaron Shaw. "The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation". In: *PLoS ONE* 8.6 (June 2013), pp. 1–5. DOI: 10.1371/journal.pone.0065782.

[HS14b] Benjamin Mako Hill and Aaron Shaw. "Consider the Redirect: A Missing Dimension of Wikipedia Research". In: *Proceedings of The International Symposium on Open Collaboration*. OpenSym '14. Berlin, Germany: ACM, 2014, 28:1–28:4. ISBN: 978-1-4503-3016-9. DOI: 10.1145/2641580.2641616. URL: http://doi.acm.org/10.1145/2641580.2641616.

[Hri+13] Desislava Hristova et al. "The Life of the Party: Impact of Social Mapping in OpenStreetMap". In: *Proc. of ICWSM*. 2013.

[Hu+07] Meiqun Hu et al. "Measuring article quality in Wikipedia: models and evaluation". In: *Proc. CIKM*. Lisbon, Portugal, 2007, pp. 243–252. ISBN: 978-1-59593-803-9. DOI: 10.1145/1321440.1321476. URL: http://doi.acm.org/10.1145/1321440.1321476.

[HK08] Rob J Hyndman and Yeasmin Khandakar. "Automatic time series for forecasting: The forecast package for R". In: *Journal of Statistical Software* 27.3 (2008), pp. 1–22.

[13] *Geographic information – Data quality*. Standard. Geneva, CH: International Organization for Standardization, Dec. 2013.

[JS02] Nathalie Japkowicz and Shaju Stephen. "The class imbalance problem: A systematic study". In: *Intelligent data analysis* 6.5 (2002), pp. 429–449.

[Joh+16a] Isaac L. Johnson et al. "Not at Home on the Range: Peer Production and the Urban/Rural Divide". In: *Proc. of CHI*. Santa Clara, California, USA: ACM, 2016, pp. 13–25. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858123. URL: http://doi.acm.org/10.1145/2858036.2858123.

[Joh+16b] Isaac L. Johnson et al. "The Geography and Importance of Localness in Geotagged Social Media". In: *Proc. of CHI*. Santa Clara, California, USA: ACM, 2016, pp. 515–526. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858122. URL: http://doi.acm.org/10.1145/2858036.2858122.

[Jur+15] Raja Jurdak et al. "Understanding Human Mobility from Twitter". In: *PLoS ONE* 10.7 (July 2015), pp. 1–16. DOI: 10.1371/journal.pone.0131469.

[KR16] Gerald C. Kane and Sam Ransbotham. "Research Note—Content and Collaboration: An Affiliation Network Approach to Information Quality in Online Peer Production Communities". In: *Information Systems Research* 27.2 (2016), pp. 424–439. DOI: 10.1287/isre.2016.0622. eprint: http://dx.doi.org/10.1287/isre.2016.0622. URL: http://dx.doi.org/10.1287/isre.2016.0622.

[KW93] Steven J Karau and Kipling D Williams. "Social loafing: A meta-analytic review and theoretical integration." In: *Journal of personality and social psychology* 65.4 (1993), p. 681.

[KG10] Brian Keegan and Darren Gergle. "Egalitarians at the gate: one-sided gatekeeping practices in social media". In: *Proc. CSCW*. Savannah, Georgia, USA, 2010, pp. 131–134. ISBN: 978-1-60558-795-0. DOI: 10.1145/1718918.1718943. URL: http://doi.acm.org/10.1145/1718918.1718943.

[KGC13] Brian Keegan, Darren Gergle, and Noshir Contractor. "Hot Off the Wiki: Structures and Dynamics of Wikipedia's Coverage of Breaking News Events". In: *American Behavioral Scientist* 57.5 (2013), pp. 595–622.

[KK08] Aniket Kittur and Robert E. Kraut. "Harnessing the wisdom of crowds in Wikipedia: quality through coordination". In: *Proc. CSCW*. San Diego, CA, USA, 2008. ISBN: 978-1-60558-007-4. DOI: 10.1145/1460563.1460572. URL: http://doi.acm.org/10.1145/1460563.1460572.

[KPK09] Aniket Kittur, Bryan Pendleton, and Robert E. Kraut. "Herding the Cats: The Influence of Groups in Coordinating Peer Production". In: *Proc. of WikiSym*. Orlando, Florida, 2009, 7:1–7:9. ISBN: 978-1-60558-730-1. DOI: 10.1145/1641309.1641321. URL: http://doi.acm.org/10.1145/1641309.1641321.

[Kit+07] A. Kittur et al. "Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie". In: *Proc. alt.CHI*. 2007. URL: http://www.parc.com/research/publications/details.php?id=5904.

[Kri+07] Travis Kriplean et al. "Community, Consensus, Coercion, Control: Cs*W or How Policy Mediates Mass Participation". In: *Proc. of GROUP*. Sanibel Island, Florida, USA, 2007, pp. 167–176. ISBN: 978-1-59593-845-9. DOI: 10.1145/1316624.1316648. URL: http://doi.acm.org/10.1145/1316624.1316648.

[LW05] Karim R. Lakhani and Robert G. Wolf. "Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects". In: *Perspectives on Free and Open Source Software*. Ed. by J. Feller et al. MIT Press, 2005.

[Lam+11] Shyong (Tony) K. Lam et al. "WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance". In: *Proc. of WikiSym*. 2011, pp. 1–10.

[Lam+12] Cliff Lampe et al. "Classroom Wikipedia Participation Effects on Future Intentions to Contribute". In: *Proc. of CSCW*. 2012, pp. 403–406.

[Leh+14] Janette Lehmann et al. "Reader Preferences and Behavior on Wikipedia". In: *Proc. HT*. ACM. 2014, pp. 88–97.

[LT02] Josh Lerner and Jean Tirole. "Some Simple Economics of Open Source". In: *The Journal of Industrial Economics* 50.2 (2002), pp. 197–234. ISSN: 00221821, 14676451. URL: http://www.jstor.org/stable/3569837.

[LP09] James P LeSage and R Kelley Pace. *Introduction to Spatial Econometrics (Statistics, textbooks and monographs)*. CRC Press, 2009.

[LGX13] Linna Li, Michael F. Goodchild, and Bo Xu. "Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr". In: *Cartography and Geographic Information Science* 40.2 (2013), pp. 61–77. DOI: 10.1080/15230406.2013.777139. eprint: http://dx.doi.org/10.1080/15230406.2013.777139. URL: http://dx.doi.org/10.1080/15230406.2013.777139.

[Lin+05] Kimberly Ling et al. "Using social psychology to motivate contributions to online communities". In: *Journal of Computer-Mediated Communication* 10.4 (2005), pp. 00–00.

[LS10a] Nedim Lipka and Benno Stein. "Identifying featured articles in Wikipedia: writing style matters". In: *Proc. WWW*. Raleigh, North Carolina, USA, 2010, pp. 1147–1148. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772847. URL: http://doi.acm.org/10.1145/1772690.1772847.

[LA05] Ivy Liu and Alan Agresti. "The analysis of ordered categorical data: An overview and a survey of recent developments". In: *Test* 14.1 (2005), pp. 1–73.

[LR11] Jun Liu and Sudha Ram. "Who does what: Collaboration patterns in the Wikipedia and their impact on article quality". In: *ACM TMIS* 2.2 (July 2011), 11:1–11:23. ISSN: 2158-656X. DOI: 10.1145/1985347.1985352. URL: http://doi.acm.org/10.1145/1985347.1985352.

[LS10b] Teun Lucassen and Jan Maarten Schraagen. "Trust in Wikipedia: how users trust information from an unknown source". In: *Proc. WICOW*. Raleigh, North Carolina, USA, 2010, pp. 19–26. ISBN: 978-1-60558-940-4. DOI: 10.1145/1772938.1772944. URL: http://doi.acm.org/10.1145/1772938.1772944.

[LVK11] Ina Ludwig, Angi Voss, and Maike Krause-Traudes. "A Comparison of the Street Networks of Navteq and OSM in Germany". In: *Advancing Geoinformation Science for a Changing World*. Springer, 2011, pp. 65–84.

[MT14] Mikhil Masli and Loren G. Terveen. "Leveraging the Contributory Potential of User Feedback". In: *Proc. of CSCW*. CSCW '14. Baltimore, Maryland, USA: ACM, 2014, pp. 956–966. ISBN: 978-1-4503-2540-0. DOI: 10.1145/2531602.2531686. URL: http://doi.acm.org/10.1145/2531602.2531686.

[Mil01] James M. Cook Miller McPherson Lynn Smith-Lovin. "Birds of a Feather: Homophily in Social Networks". In: *Annual Review of Sociology* 27 (2001), pp. 415–444. ISSN: 03600572, 15452115. URL: http://www.jstor.org/stable/2678628.

[Mor+13] Jonathan T. Morgan et al. "Project Talk: Coordination Work and Group Membership in WikiProjects". In: *Proc. of WikiSym*. 2013, 3:1–3:10.

[Mor+14] Jonathan T. Morgan et al. "Editing Beyond Articles: Diversity & Dynamics of Teamwork in Open Collaborations". In: *Proc. of CSCW*. Baltimore, Maryland, USA, 2014, pp. 550–563. ISBN: 978-1-4503-2540-0. DOI: 10.1145/2531602.2531654. URL: http://doi.acm.org/10.1145/2531602.2531654.

[MCH99] Richard Morrill, John Cromartie, and Gary Hart. "METROPOLITAN, URBAN, AND RURAL COMMUTING AREAS: TOWARD A BETTER DEPICTION OF THE UNITED STATES SETTLEMENT SYSTEM". In: *Urban Geography* 20.8 (1999), pp. 727–748. DOI: 10.2747/0272-3638.20.8.727. eprint: http://dx.doi.org/10.2747/0272-3638.20.8.727. URL: http://dx.doi.org/10.2747/0272-3638.20.8.727.

[NC03] Jeanne Nakamura and Mihaly Csikzentmihalyi. "The construction of meaning through vital engagement". In: *Flourishing: Positive psychology and the life well-lived*. Ed. by Corey L. M. Keyes and Jonathan Haidt. Washington, DC, US: American Psychological Association, 2003. Chap. 4, pp. 83–104.

[NS08] Vivi Nastase and Michael Strube. "Decoding Wikipedia Categories for Knowledge Acquisition." In: *AAAI*. 2008.

[NZ14] Pascal Neis and Dennis Zielstra. "Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap". In: *Future Internet* 6.1 (2014), pp. 76–106.

[NZ12] Pascal Neis and Alexander Zipf. "Analyzing the Contributor Activity of a Volunteered Geographic Information Project – The Case of OpenStreetMap". In: *ISPRS International Journal of Geo-Information* 1.2 (2012), pp. 146–165.

[NGL11] Keiichi Nemoto, Peter Gloor, and Robert Laubacher. "Social capital increases efficiency of collaboration among Wikipedia editors". In: *Proc. HT*. Eindhoven, The Netherlands, 2011, pp. 231–240. ISBN: 978-1-4503-0256-2. DOI: 10.1145/1995966.1995997. URL: http://doi.acm.org/10.1145/1995966.1995997.

[Nou+12] Anastasios Noulas et al. "A Tale of Many Cities: Universal Patterns in Human Urban Mobility". In: *PLoS ONE* 7.5 (May 2012), pp. 1–10. DOI: 10.1371/journal.pone.0037027.

[Nov07] Oded Nov. "What Motivates Wikipedians?" In: *Commun. ACM* 50.11 (Nov. 2007), pp. 60–64. ISSN: 0001-0782.

[Oli+14] Dev Oliver et al. "Disaster Response and Relief, VGI Volunteer Motivation in". In: *Encyclopedia of Social Network Analysis and Mining*. Ed. by Reda Alhajj and Jon Rokne. New York, NY: Springer New York, 2014, pp. 370–380. ISBN: 978-1-4614-6170-8. DOI: 10.1007/978-1-4614-6170-8_57. URL: http://dx.doi.org/10.1007/978-1-4614-6170-8_57.

[PHT09] Katherine Panciera, Aaron Halfaker, and Loren Terveen. "Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia". In: *Proc. GROUP*. 2009, pp. 51–60.

[Pre00] Jenny Preece. *Online communities: Designing Usability and Supporting Socialbilty*. John Wiley & Sons, Inc., 2000.

[PMT10] Reid Priedhorsky, Mikhil Masli, and Loren Terveen. "Eliciting and Focusing Geographic Volunteer Work". In: *Proc. of CSCW*. Savannah, Georgia, USA, 2010, pp. 61–70. ISBN: 978-1-60558-795-0. DOI: 10.1145/1718918.1718931. URL: http://doi.acm.org/10.1145/1718918.1718931.

[Pri+07] Reid Priedhorsky et al. "Creating, Destroying, and Restoring value in Wikipedia". In: *Proc. of GROUP*. 2007, pp. 259–268.

[QMC14] Giovanni Quattrone, Afra Mashhadi, and Licia Capra. "Mind the Map: The Impact of Culture and Economic Affluence on Crowd-mapping Behaviours". In: *Proc. of CSCW*. Baltimore, Maryland, USA: ACM, 2014, pp. 934–944. ISBN: 978-1-4503-2540-0. DOI: 10.1145/2531602.2531713. URL: http://doi.acm.org/10.1145/2531602.2531713.

[Ras+06] Al M. Rashid et al. "Motivating Participation by Displaying the Value of Contribution". In: *Proc. of CHI*. CHI '06. Montréal, Québec, Canada, 2006, pp. 955–958. ISBN: 1-59593-372-7. DOI: 10.1145/1124772.1124915. URL: http://doi.acm.org/10.1145/1124772.1124915.

[RR11] Joseph Reagle and Lauren Rhue. "Gender Bias in Wikipedia and Britannica". In: *International Journal of Communication* 5 (2011). ISSN: 1932-8036.

[Rei11] Antonio J. Reinoso. "Temporal and behavioral patterns in the use of Wikipedia". PhD thesis. Universidad Rey Juan Carlos, 2011.

[Rot] Amy Roth. *Student Contributions to Wikipedia*. https://outreach.wikimedia.org/wiki/Student_Contributions_to_Wikipedia. Retrieved June 2, 2014.

[Sch+06] Kendra L Schwartz et al. "Family Medicine Patients' Use of the Internet for Health Information: a MetroNet Study". In: *The Journal of the American Board of Family Medicine* 19.1 (2006), pp. 39–45.

[SZ14] Robert Seamans and Feng Zhu. "Responses to Entry in Multi-Sided Markets: The Impact of Craigslist on Local Newspapers". In: *Management Science* 60.2 (2014), pp. 476–493. DOI: 10.1287/mnsc.2013.1785. eprint: http://dx.doi.org/10.1287/mnsc.2013.1785. URL: http://dx.doi.org/10.1287/mnsc.2013.1785.

[Sim56] Herbert A Simon. "Rational choice and the structure of the environment." In: *Psychological review* 63.2 (1956), p. 129.

[SW12] Jacob Solomon and Rick Wash. "Bootstrapping wikis: developing critical mass in a fledgling community by seeding content". In: *Proc. CSCW*. Seattle, Washington, USA, 2012, pp. 261–264. ISBN: 978-1-4503-1086-4. DOI: 10.1145/2145204.2145247. URL: http://doi.acm.org/10.1145/2145204.2145247.

[Spo07] Anselm Spoerri. "What is popular on Wikipedia and why?" In: *First Monday* 12.4 (2007).

[SH07] Klaus Stein and Claudia Hess. "Does it matter who contributes: A study on featured articles in the German Wikipedia". In: *Proc. HT*. 2007, pp. 171–174.

[Ste13] Monica Stephens. "Gender and the GeoWeb: divisions in the production of user-generated cartographic information". en. In: *GeoJournal* (2013). 00001, pp. 1–16. ISSN: 0343-2521, 1572-9893. DOI: 10.1007/s10708-013-9492-z. URL: http://link.springer.com/article/10.1007/s10708-013-9492-z (visited on 09/18/2013).

[Ste48] John Q. Stewart. "Demographic Gravitation: Evidence and Applications". In: *Sociometry* 11.1/2 (1948), pp. 31–58. ISSN: 00380431. URL: http://www.jstor.org/stable/2785468.

[SAY09] Besiki Stvilia, Abdullah Al-Faraj, and Yong Jeong Yi. "Issues of cross-contextual information quality evaluation–The case of Arabic, English, and Korean Wikipedias". In: *Library & Information Science Research* 31.4 (2009), pp. 232–239. ISSN: 0740-8188. DOI: 10.1016/j.lisr.2009.07.005. URL: http://www.sciencedirect.com/science/article/pii/S0740818809000954.

[Stv+05a] Besiki Stvilia et al. "Assessing information quality of a community-based encyclopedia". In: *Proc. ICIQ*. 2005, pp. 442–454.

[Stv+05b] Besiki Stvilia et al. "Information quality discussions in Wikipedia". In: *Proc. ICKM*. 2005, pp. 101–113.

[Stv+08a] Besiki Stvilia et al. "Information quality work organization in Wikipedia". In: *J. Am. Soc. Inf. Sci. Technol.* 59 (6 Apr. 2008), pp. 983–1001. ISSN: 1532-2882. DOI: 10.1002/asi.v59:6. URL: http://dl.acm.org/citation.cfm?id=1358262.1358274.

[Stv+08b] Besiki Stvilia et al. "Information quality work organization in Wikipedia". In: *Journal of ASIST* 59.6 (2008), pp. 983–1001.

[SY12] Yu Suzuki and Masatoshi Yoshikawa. "Mutual Evaluation of Editors and Texts for Assessing Quality of Wikipedia Articles". In: *Proc. WikiSym*. 2012.

[TSK06] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Education, 2006, pp. 163–164.

[TTH15] Jacob Thebault-Spieker, Loren G. Terveen, and Brent Hecht. "Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets". In: *Proc. of CSCW*. Vancouver, BC, Canada: ACM, 2015, pp. 265–275. ISBN: 978-1-4503-2922-4. DOI: 10.1145/2675133.2675278. URL: http://doi.acm.org/10.1145/2675133.2675278.

[Thi+12] Marijn ten Thij et al. "Modeling and predicting page-view dynamics on Wikipedia". In: *CoRR* abs/1212.5943 (2012). URL: http://arxiv.org/abs/1212.5943.

[Tho+16] Bart Thomee et al. "YFCC100M: The New Data in Multimedia Research". In: *Commun. ACM* 59.2 (Jan. 2016), pp. 64–73. ISSN: 0001-0782. DOI: 10.1145/2812802. URL: http://doi.acm.org/10.1145/2812802.

[TCG09] Jennifer Thom-Santelli, Dan R. Cosley, and Geri Gay. "What's Mine is Mine: Territoriality in Collaborative Authoring". In: *Proc. of CHI*. Boston, MA, USA, 2009, pp. 1481–1484. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518925. URL: http://doi.acm.org/10.1145/1518701.1518925.

[TCG10] Jennifer Thom-Santelli, Dan Cosley, and Geri Gay. "What Do You Know?: Experts, Novices and Territoriality in Collaborative Systems". In: *Proc. of CHI*. Atlanta, Georgia, USA, 2010, pp. 1685–1694. ISBN: 978-1-60558-929-9. DOI: 10.1145/1753326.1753578. URL: http://doi.acm.org/10.1145/1753326.1753578.

[Tob70] W. R. Tobler. "A Computer Movie Simulating Urban Growth in the Detroit Region". English. In: *Econ. Geography* 46 (1970), pp. 234–240. ISSN: 00130095.

[VWM07] Fernanda B. Viégas, Martin Wattenberg, and Matthew M. McKeon. "The hidden order of Wikipedia". In: *Proc, OCSC*. 2007.

[WI11] Se Wang and Mizuho Iwaihara. "Quality evaluation of Wikipedia articles through edit history and editor groups". In: *Web Technologies and Applications*. 2011, pp. 188–199.

[WCR13] Morten Warncke-Wang, Dan Cosley, and John Riedl. "Tell Me More: An Actionable Quality Model for Wikipedia". In: *Proc. of OpenSym/WikiSym*. 2013, 8:1–8:10.

[War+15a] Morten Warncke-Wang et al. "Misalignment Between Supply and Demand of Quality Content in Peer Production Communities". In: *Proc. of ICWSM*. 2015.

[War+15b] Morten Warncke-Wang et al. "The Success and Failure of Quality Improvement Projects in Peer Production Communities". In: *Proc. of CSCW*. Vancouver, BC, Canada: ACM, 2015, pp. 743–756. ISBN: 978-1-4503-2922-4. DOI: 10.1145/2675133.2675241. URL: http://doi.acm.org/10.1145/2675133.2675241.

[WVH07] Martin Wattenberg, Fernanda Viégas, and Katherine Hollenbach. "Visualizing Activity on Wikipedia with Chromograms". In: *Human-Computer Interaction - INTERACT 2007*. Ed. by Cécilia Baranauskas et al. Vol. 4663. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2007, pp. 272–287. ISBN: 978-3-540-74799-4. URL: http://dx.doi.org/10.1007/978-3-540-74800-7_23.

[WWC12] Robert West, Ingmar Weber, and Carlos Castillo. "Drawing a Data-driven Portrait of Wikipedia Editors". In: *Proc. of OpenSym/WikiSym*. Linz, Austria, 2012, 3:1–3:10. ISBN: 978-1-4503-1605-7. DOI: 10.1145/2462932.2462937. URL: http://doi.acm.org/10.1145/2462932.2462937.

[WH07] Dennis M. Wilkinson and Bernardo A. Huberman. "Cooperation and quality in Wikipedia". In: *Proc. WikiSym*. Montreal, Quebec, Canada, 2007. ISBN: 978-1-59593-861-9. DOI: 10.1145/1296951.1296968. URL: http://doi.acm.org/10.1145/1296951.1296968.

[WMF] WMF. *Spring 2012 United States and Canada student article quality research results*. https://en.wikipedia.org/wiki/Wikipedia:Ambassadors/Research/Article_quality/Results. Retrieved June 2, 2014.

[WP09] Thomas Wöhner and Ralf Peters. "Assessing the quality of Wikipedia articles with lifecycle based metrics". In: *Proc. WikiSym*. 2009, 16:1–16:10.

[Woo+13] Spencer A Wood et al. "Using social media to quantify nature-based tourism and recreation". In: *Scientific Reports* 3 (2013).

[WHC12] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. "Classifying Wikipedia Articles Using Network Motif Counts and Ratios". In: *Proc. WikiSym*. 2012.

[XL11] Yanxiang Xu and Tiejian Luo. "Measuring article quality in Wikipedia: Lexical clue model". In: *3rd Symposium on Web Society*. IEEE. 2011, pp. 141–146.

[Yan+16] Diyi Yang et al. "Who Did What: Editor Role Identification in Wikipedia". In: *Proc. of ICWSM*. 2016. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13077.

[YNS15] Tse-Chuan Yang, Aggie J Noah, and Carla Shoff. "Exploring Geographic Variation in US Mortality Rates Using a Spatial Durbin Approach". In: *Population, space and place* 21.1 (2015), pp. 18–37.

[YSK12] Taha Yasseri, Robert Sumi, and János Kertész. "Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis". In: *PLoS ONE* 7.1 (Jan. 2012), e30091. DOI: 10.1371/journal.pone.0030091.

[Yee10] Thomas W. Yee. "The VGAM Package for Categorical Data Analysis". In: *Journal of Statistical Software* 32.10 (2010), pp. 1–34.

[Yua+09] Y. Connie Yuan et al. "The Diffusion of a Task Recommendation System to Facilitate Contributions to an Online Community". In: *Journal of Computer-Mediated Communication* 15.1 (2009), pp. 32–59.

[ZKK12] Haiyi Zhu, Robert Kraut, and Aniket Kittur. "Organizing Without Formal Organization: Group Identification, Goal Setting and Social Modeling in Directing Online Production". In: *Proc. of CSCW*. 2012, pp. 935–944.

[ZH11] D Zielstra and HH Hochmair. "Digital Street Data: Free versus Proprietary". In: *GIM Int* 25 (2011). Retrieved July 16, 2016, pp. 29–33. URL: `http://www.gim-international.com/content/article/digital-street-data`.

[ZHN13] Dennis Zielstra, Hartwig H Hochmair, and Pascal Neis. "Assessing the Effect of Data Imports on the Completeness of OpenStreetMap–A United States Case Study". In: *Transactions in GIS* 17.3 (2013), pp. 315–334.

[ZZ10] Dennis Zielstra and Alexander Zipf. "A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany". In: *Proc. of the 13th AGILE International Conference on Geographic Information Science*. 2010.