Towards a Better Understanding of Peer-Produced Structured Content Value

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

Andrew Hall

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

Loren Terveen

July, 2019

Acknowledgements

I would first like to thank my Ph.D. advisor, Dr. Loren Terveen, for his assistance over the last five years. Throughout the Ph.D. program, I had countless informal meetings with Loren in which he taught me how to be a more logical and critical thinker. Rather than directly providing answers to many of the questions that I would have for him, Loren helped me develop the necessary tools to derive such answers myself. He would also frequently push me to get out of my comfort zone to pursue challenging research questions/directions that had large potential impacts.

I would like to thank Dr. Aaron Halfaker who is a Principle Research Scientist at the Wikimedia Foundation. I started actively working with Aaron during the spring of 2017 as a research intern at the Wikimedia Foundation. However, our collaboration did not finish once the internship ended, and for the last two years Aaron has served in a role that is essentially that of a co-advisor. I greatly appreciate Aaron's willingness to discuss both low-level technical details and high-level research questions and anything in between during our weekly meetings.

I would also like to thank many of the professors at the University of Minnesota for helping me succeed with both coursework and research. Specifically, I would like to thank Drs. Haiyi Zhu and Daniel Keefe for being a part of my thesis committee and for providing feedback when I proposed the thesis. Additionally, I'm grateful to Dr. Eric Van Wyk who introduced me to research as an undergraduate student and helped me realize that graduate school could serve an important role towards achieving my career goals.

Many current and formers members of GroupLens Research have helped me succeed in the projects I performed. GroupLens is a very collaborative lab, and I believe that this has strengthened my research tremendously. I would especially like to thank Dr. Jacob Thebault-Spieker, Allen Lin, Sarah McRoberts, Dr. Isaac Johnson, Dr. Brent Hecht, and Dr. Shilad Sen.

Finally, I would like to thank my parents and brother for all the support they have provided over a challenging five years of my life.

Dedication

To my parents, grandparents, and brother.

Abstract

Over the last 30 years, peer production has created everything from software (e.g. Linux) to encyclopedia articles (e.g. Wikipedia) to geographic data (e.g. OpenStreetMap). In recent years, peer production has increased its focus on the production of structured (key-value pair) content. This content is designed to be consumed by applications and algorithms. This thesis explores two challenges towards generating content that is as valuable as possible to these applications/algorithms. The first challenge is unique to the context of peer-produced structured data and is focused on a tension between the core peer production ethos of contributor freedom and the need for highly-standardized data in order for applications/algorithms to effectively operate. To explore this tension between freedom and standardization, I qualitatively analyze the ways in which it surfaces and then quantitatively analyze its impact. For the second challenge, I compare how different levels of automation affect content value. Contributions in peer production come from manual editing, semi-automated tool editing, and fully-automated bot editing. I use two important lenses to study the value provided by these different types of contributions. Specifically, I study value by considering 1) the relationship between content quality and demand, and 2) problematic societal-level content biases (e.g. along male versus female, Global North versus Global South, and urban versus rural lines). While peer-production research has explored these two lenses of value in the past, it has not sought to develop a robust understanding in the context of structured content. To ensure that automated and manual contributions are effectively differentiated, I also develop a bot detection model. Finally, I provide implications based on my results. For example, my work motivates socio-technical tools that can reduce the manual effort required to contribute structured data and tools that direct effort towards in-demand content.

Table of Contents

List of Tables ..... **vii**

List of Figures ..... **viii**

Notes on the Content in this Thesis ..... **ix**

1 Introduction **1**
1.1 Defining the Problem Space **1**
1.1.1 Challenge 1: Contributor Freedom versus Data Standardization. 1
1.1.2 Challenge 2: Understanding the Different Roles that Manual and Automated Contributions Play in Affecting Content Value 3
1.2 Research Questions **3**
1.3 Summary of Challenge 1 Studies **3**
1.3.1 Exploring the Tension Between Freedom and Data Standardization 3
1.3.2 Measuring Contributor Freedom's Effect on Data Standardization 4
1.4 Summary of Challenge 2 Studies **4**
1.4.1 Unidentified Bot Detection 4
1.4.2 Comparing Content Value Produced by Manual and Automated Contributions Along Three Intuitive Dimensions 4
1.5 Thesis Organization **5**

2 Background and Related Work **6**
2.1 Background on Peer Production **6**
2.2 Brief Background on Communities Studied in this Thesis **6**
2.2.1 OpenStreetMap 6
2.2.2 Wikidata 7
2.3 Performing Contributions in Peer Production **8**
2.3.1 Manual Contributions 8
2.3.2 Automated Contributions 8
2.4 Research Studying Contributor Freedom within Peer Production **9**
2.5 Studies of Content Value in Peer Production **10**

3 Understanding the Causes of a Tension Between Freedom and Standardization in Peer-Produced Structured Content **12**
3.1 Introduction **12**
3.2 Related Work **12**
3.3 Method **13**
3.4 Results and Interpretations **14**
3.4.1 Theme 1: Freedom vs. Metadata Completeness 15
3.4.2 Theme 2: Project-Specific Freedom and Metadata Correctness: Humanitarian OpenStreetMap (HOT) 17
3.4.3 Theme 3: Cultural Differences Make Global Metadata Correctness Standards Difficult to Achieve and Maintain 18
3.4.4 Theme 4: Community-Management Obstacles to Achieving Consensus 22
3.4.5 Theme 5: Data Representation Prevents Conceptual Correctness 23
3.4.6 Theme 6: Data Entry Tools May Harm Metadata Correctness and Privilege Certain Users 24
3.5 Reflecting on Correctness, Community, and Code **26**

4 Understanding the Effect of a Tension Between Freedom and Standardization in Peer-Produced Structured Content 28
4.1 Introduction 28
4.2 Related Work 29
4.3 Additional Relevant Information on OpenStreetMap Tags and Tagging Standards 29
4.4 Motivation for Analyzing Chain Business Standardization 30
4.5 Methods 31
4.5.1 Clustering Algorithm 31
4.6 Determining a Metadata Taxonomy 33
4.7 Results 34
4.7.1 Measuring Standardization 34
4.7.2 Detailed Results 36
4.8 Discussion 40
4.9 Limitations and Future Work 42

5 Bot Detection in Wikidata Using Behavioral and Other Informal Cues 44
5.1 Introduction 44
5.1.1 Contributions 45
5.2 Background and Related Work 46
5.2.1 Bot Detection 46
5.2.2 Activity Session Behavioral Patterns 46
5.3 Methods 47
5.4 Results 49
5.4.1 Formal Model Evaluation on Registered User Edits 49
5.4.2 Qualitative Model Evaluation on Anonymous User Edits 49
5.4.3 Summary of Model Evaluations 51
5.5 Applying the Model To Registered and Anonymous Users 52
5.5.1 Implications on Behavioral Research in Peer Production 52
5.5.2 Implications for the Wikidata Community and Applications using Wikidata 53
5.6 Discussion 54
5.6.1 Model and Feature Implications 54
5.6.2 Model Improvements 55

6 Exploring the Effects of Manual and Automated Contributions on Content Value in Wikidata 56
6.1 Introduction 56
6.1.1 Contributions 57
6.2 Related Work 58
6.2.1 Consumer-Level Value Analyses 58
6.2.2 Societal-Level Value Analyses 58
6.3 Methods and Data 59
6.3.1 Background on Wikidata Revision Data Used 59
6.3.2 Determining Contribution Strategy Types of Revisions 59
6.3.3 Analyzing Wikidata Contribution Strategy Behavior Longitudinally 60
6.3.4 Measuring Item Quality 60
6.3.5 Measuring Item Demand 61
6.3.6 Data Generation for Consumer-Level Value Analyses 61

6.3.7 Data Generation for Societal-Level Value Analyses 62
6.3.8 Additional Details of Our Analytic Methods 63
6.4 Results 64
6.4.1 Value from a Consumer-Level Perspective 64
6.4.2 Value from a Societal-Level Perspective 70
6.5 Discussion, Implications, and Future Work 74
6.5.1 Should Bots – and Wikidata – adjust editing behavior? 74
6.5.2 Potential Mechanisms to Support Aligned Content Editing 75
6.5.3 Using Automated Contribution Strategies to Reduce Rural Content Disparities 76
6.5.4 Future Work 76
6.6 Limitations 77
7 Conclusion 78
7.1 Summary of Studies 78
7.2 Summary of Contributions 79
7.3 Implications and Future Work 80
8 Bibliography 84
9 Appendix 92
9.1 Study 2 Sampling Contingent Metadata 92
9.2 Study 3 Details of Data Preparation 93
9.2.1 Extracting Revision Data and Sessionization 93
9.2.2 Ground Truth Bot Account Data 94
9.3 Study 3 Bot Prediction Model Features 95
9.4 Study 3 Detailed Anonymous Contributor Qualitative Coding Results 96
9.5 Study 4 Additional Details of Determining Contribution Strategy Types of Revisions 98
9.5.1 Identifying Revisions from Bots 98
9.5.2 Identifying Revisions from Semi-Automated Tools 98
9.5.3 Identifying Revisions from Manual Effort 98
9.6 Study 4 Additional Tables and Figures 99

List of Tables

Table 3.1: Participant Information 14

Table 4.1: Chain Businesses Used in Standardization Analyses 32

Table 4.2: Chain Business Metadata (Key) Role Classes 34

Table 4.3: Business-Universal Pairs with 2 or More Aligned Values 39

Table 5.1: Bot Prediction Model Fitness on Registered User Contributions 49

Table 5.2: Qualitative Analysis Codes 52

Table 5.3: Bot Prediction Model Features 95

Table 5.3: Overview of the Results of the Anonymous Contributor Qualitative Coding Summary 96

Table 6.1: Content Edits Sampled from Each Contribution Strategy and Period 99

Table 6.2: Content Edits Sampled for Male and Female Items 99

Table 6.3: Content Edits Sampled for Global North and Global South Items 99

Table 6.4: Content Edits Sampled for Urban and Rural Items 100

List of Figures

Figure 1.1: Foursquare's City Guide using OpenStreetMap Data 2

Figure 4.1: Universal Business-Key-Value Triples 37

Figure 5.1: Precision-Recall and Receiver Operating Characteristic Curves 49

Figure 5.2: Maximum Monthly Anonymous User Session Lengths 53

Figure 6.1: Mean Item Demand (Expected Quality) Percentile of Content Edits 65

Figure 6.2: Density Function Breaking Down Demand (Expected Quality) Percentiles for Bot Content Edits in the 2016-2017 Period 65

Figure 6.3: Mean Quality Difference (Actual Minus Expected) for Items within Content Edits 66

Figure 6.4: Item Mean Quality Difference (Actual Minus Expected), Broken Down by Demand (Expected Quality) Quantile. Plots represent the beginning and end of study. 66

Figure 6.5: Item Mean Quality, Broken Down by Demand Quantile. Plots represent the beginning and end of study. 68

Figure 6.6: Mean Absolute Error (MAE) Based on Item Quality Difference (Between Actual and Expected Item Quality) 68

Figure 6.7: Normalized Human Female Item Content Edit Proportion 71

Figure 6.8: Normalized Global South Item Content Edit Proportion 73

Figure 6.9: Normalized Rural Item Content Edit Proportion 73

Notes on the Content in this Thesis

I led all research within this thesis. As of the time of the submission of this thesis, the thesis incorporates the following three pieces of published content, included as chapters 3, 4, and 5.

- Hall, A., McRoberts, S., Thebault-Spieker, J., Lin, Y., Sen, S., Hecht, B., & Terveen, L. (2017, May). Freedom versus standardization: structured data generation in a peer production community. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 6352-6362). ACM.
- Hall, A., Thebault-Spieker, J., Sen, S., Hecht, B., & Terveen, L. (2018, August). Exploring the Relationship Between Informal Standards and Contributor Practice in OpenStreetMap. In *Proceedings of the 14th International Symposium on Open Collaboration* (p. 10). ACM.
- Hall, A., Terveen, L., & Halfaker, A. (2018). Bot Detection in Wikidata Using Behavioral and Other Informal Cues. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 64.

Since much of the work in this thesis was collaborative, I often use the pronoun “we” in the text.

1 Introduction

Built upon large-scale, decentralized collaboration, the phenomenon of *peer production* has been described by Benkler as "the most significant organizational innovation that has emerged from Internet-mediated social practice" [8]. It is easy to see why peer production is considered such an innovation if one looks at the success of its various communities. For instance, Linux is a highly-popular open source operating system. Photo sharing platforms like Flickr have enabled people to share tens of millions of photos with each other [101]. Citizen science platforms like Zooniverse and Foldit have produced large quantities of useful scientific information (e.g. [49]). Finally, the online encyclopedia, Wikipedia, is the world's fifth-most-visited website [113], and Wikipedia's content is also used in Google Knowledge Graph and other third-party services (e.g. [67]).

The focus of my thesis research has been peer production communities that produce repositories of *structured data*. The structured data produced in these communities are key-value pairs and are designed to be used by applications and algorithms. Indeed, such data is widely used. For example, data from Wikidata, Wikipedia's sister project, is used in Google Knowledge Graph, Apple's voice assistant Siri, and in Wikipedia infoboxes. Further, OpenStreetMap (OSM), "The Wikipedia of Maps" [17], produces structured map data used by Mapbox, Craigslist, Foursquare's City Guide (pictured in Figure 1.1), and in humanitarian efforts.

1.1 Defining the Problem Space

It is important that the structured data generated by peer production communities are as *valuable* as possible to the applications and algorithms running on machines. My work has focused on understanding the *challenges* associated with creating valuable content. There are two broad themes in this work. I introduce these next.

1.1.1 Challenge 1: Contributor Freedom versus Data Standardization.

Contributor freedom is a foundational principle of peer production communities and is responsible for the incredible amounts of content that communities have produced. This principle suggests that contributors "go for it" [114], "make changes as they see fit" [33], and "be bold" [114] and allows contributors to not be overburdened with complicated guidelines or rules. In the context of OpenStreetMap, this principle even indicates that "[OpenStreetMap will never] force any of its mappers to do anything" [115].

While contributor freedom within the community helped Wikipedia and OpenStreetMap create millions of encyclopedia articles and hundreds of gigabytes of geospatial data, it is not clear whether it will be as helpful as communities begin major pushes into producing *highly standardized structured data*. Wikipedia's

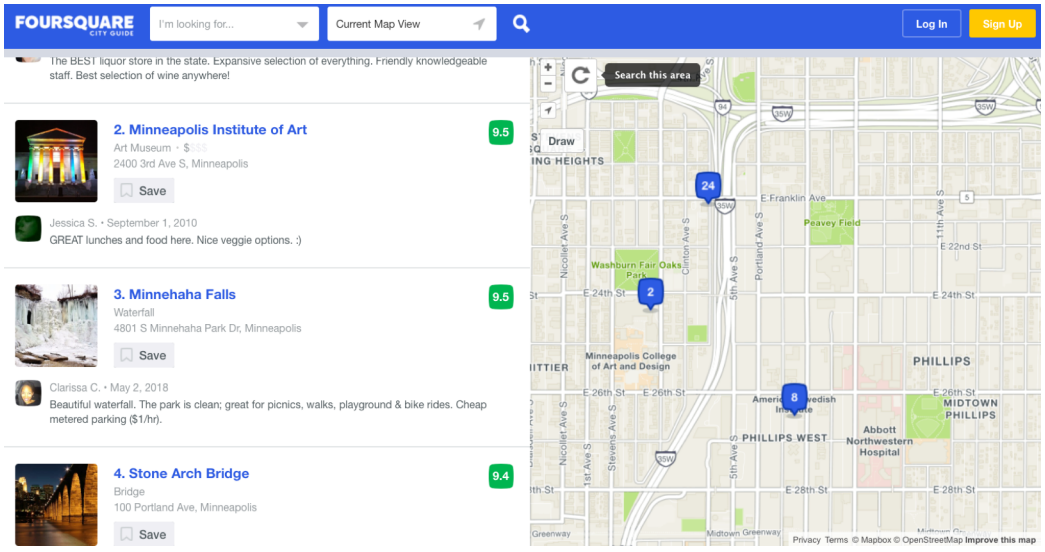

Figure 1.1: Foursquare's City Guide using OpenStreetMap Data

https://foursquare.com

structured data efforts are manifest through Wikidata. OSM's structured data creation occurs in its tagging infrastructure, which lets editors specify the semantics of a geospatial entity using key-value pairs (e.g. its name, whether it is a restaurant or a hospital, etc.). Both these initiatives are motivated by a desire to help computers understand real world semantics (e.g., moving towards computing over "things not strings", as Google puts it [116]).

For Wikidata and OSM to support computing applications most effectively, their structured data must have a high degree of standardization. This means that similar entities must be represented in syntactically similar ways within these communities or else applications will not use the datam effectively. For example, popular tools developed to make maps from OSM data (e.g. Mapbox, CartoCSS) will not use the right color or icon for entities if the entities are not "tagged" with the proper structured data. Similarly, standardizing entities like roads is especially valuable for projects such as the Humanitarian OpenStreetMap Team (HOT) [117], which seeks to provide aid in humanitarian disasters. The story is similar for Wikidata. Tools developed to generate natural language Wikipedia articles from Wikidata (e.g. Reasonator [118]) work properly only if the data conforms to standards. The same is true for Wikipedia itself, which pulls in data from Wikidata for language alignment and infoboxes [119].

Hence, our first challenge focuses on a tension that is novel to peer production communities producing structured data. Specifically, contributors need freedom to produce large amounts of data, but applications/algorithms need data standardization to effectively operate. The first two studies in my thesis explore 1) how this tension manifests itself and 2) the impact on data standardization that occurs as a result.

1.1.2 Challenge 2: Understanding the Different Roles that Manual and Automated Contributions Play in Affecting Content Value

While initial contributions to peer production were largely manual work (e.g. manual Wikipedia article writing), automated contributions have played an ever-increasing role in these communities. Wikipedia, OpenStreetMap, and Wikidata all have taken advantage of automation to perform work at a rate and scale exceeding that of manual contributors. Automated contributions are particularly prevalent in Wikidata. In fact, 88% of Wikidata edits come from fully-automated bots [86] and an increasingly large number of edits are also coming from semi-automated contributor tools. Given the large prevalence of automated contributions in creating structured content, it is important to understand how they *compare* to manual contributors in their effect on content value. The last two studies in my thesis carry out such a comparison.

We compare value along two intuitive and important dimensions. First, we consider the relationship between content quality (as assessed by tools that are based on community-defined standards) and content demand to obtain a metric of value representing consumer data needs. Second, we consider value through the lens of problematic societal-level biases. For example, we consider male versus female, Global North versus Global South, and urban versus rural biases in peer-produced structured content. Pertaining to this second challenge, all of the dimensions in which we consider value have been studied in the context of peer production (e.g. [26,38,44,57,87]), or even more broadly (e.g. [2,48]). However, none of these dimensions have been studied to a significant degree in the context of peer-produced *structured content*.

1.2 Research Questions

The work in this thesis answers the following three research questions.

- How does peer production's strong commitment to contributor freedom affect its efforts to produce standardized structured data? (answered via study 1)
- Given the fundamental ethos of contributor freedom, to what extent does *actual contributor practice* follow a community's *guidelines and best practices*? (answered via study 2)
- How do manual and automated contributions affect content value? (answered via studies 3 and 4)

1.3 Summary of Challenge 1 Studies

1.3.1 Exploring the Tension Between Freedom and Data Standardization

As noted above, peer-produced structured data is used effectively by applications only if it follows standards. For the first piece of research in this thesis, we did an interview study focused on OpenStreetMap's knowledge production processes to investigate how – and how successfully – this community creates and applies its data standards. Our study revealed a fundamental tension between the need to produce structured data in a standardized way and OpenStreetMap's tradition of contributor freedom. We extracted six themes that manifested this tension and three overarching concepts, *correctness*, *community*, and *code*, which help make sense of and synthesize the themes. We also offered suggestions for improving OpenStreetMap's

knowledge production processes, including new data models, sociotechnical tools, and community practices (e.g. stronger leadership).

1.3.2 Measuring Contributor Freedom's Effect on Data Standardization

For our second study, we wanted to measure the effect of contributor freedom on data standardization. To do so, we carried out a study in OpenStreetMap to investigate adherence to the community's geographic structured data guidelines[1]. We found that most applied structured data was consistent with the community's standards; however, we also found that the standards identified many opportunities for applying data that were not achieved. In addition, when we situated the standards in the context of OpenStreetMap's data model, we found a significant amount of ambiguity; the syntax allowed only one value, but everyday meaning -- and the standards themselves -- called for multiple values. Our results suggested significant opportunities for OpenStreetMap to improve content value.

1.4 Summary of Challenge 2 Studies

1.4.1 Unidentified Bot Detection

Understanding the ways in which humans and bots behave and add value in peer production communities is an important topic, and one that relies on accurate bot recognition. Yet, in many cases, bot activities are not explicitly flagged and could be mistaken for human contributions. For our third study, we developed a machine classifier to detect previously unidentified bots using implicit behavioral and other informal editing characteristics. We showed that this method yields a high level of fitness under both formal evaluation (PR-AUC: 0.845, ROC-AUC: 0.985) and a qualitative analysis of "anonymous" contributor edit sessions. Our findings indicate that, most of the time, unidentified bots do not perform a significant portion of edits. However, we also identified some cases where unflagged bot activities can significantly misrepresent manual behavior in analyses. Our model has the potential to support future research and community patrolling activities.

1.4.2 Comparing Content Value Produced by Manual and Automated Contributions Along Three Intuitive Dimensions

For our final study, we explored how Wikidata's automated and non-automated contributions differ in the value they produce. In performing this exploration, we define content value through two important and intuitive lenses. These lenses consider 1) the relationship between content quality and consumer demand and 2) problematic societal-level biases. Our results indicate that automated contribution mechanisms are less effective than manual contributions at targeting work based on consumer demand. However, automated mechanisms also appear effective in improving the quality of underrepresented content (e.g., pertaining to rural areas and the Global South). Based on our findings, we provide actionable insights for Wikidata and other peer production communities.

---

[1] http://wiki.openstreetmap.org

1.5 Thesis Organization

The remainder of the thesis is organized as follows. The next chapter covers background information and a broad summary of relevant related work. The following four chapters each cover one of the four studies that I have performed. Each of these chapters also includes a discussion of related work specific to the respective study. The final chapter provides conclusions, a summary of contributions, and implications of results.

2 Background and Related Work

2.1 Background on Peer Production

Peer production has been described as an "open collaborative innovation and creation, performed by diverse, decentralized groups organized principally by neither price signals not organization hierarchy, harnessing heterogeneous motivations, and governed and managed based on principles other than the residual authority of ownership implemented through contract" [8]. The roots of peer production stem from the FOSS (free and open-source software) community, which originated in the mid-twentieth century [111]. Building upon the ideals of data and software openness held by the FOSS community, peer production came into being in the 1990s due to the Internet's ability to make extremely large-scale, remote collaboration possible. Such collaboration allowed for the creation of new types of socio-technical systems/communities.

Linux was one of the earliest examples of peer production and other systems/communities soon followed. For instance, the online encyclopedia, Nupedia was launched in the year 2000 and evolved shortly thereafter into the peer production community, Wikipedia. Motivated by founder Jimmy Wale's desire for "a world in which every single person on the planet is given free access to the sum of all human knowledge." [120], Wikipedia would eventually become the largest encyclopedia ever produced.

Openness and freedom have always been important in peer production and FOSS, both with regards to the content produced and to the ability of contributors to generally create what they wish. Wikipedia is defined as "the free encyclopedia that anyone can edit" [108] and other peer production communities have similar views. Given such views, questions have understandably arose about the reliability of peer-produced content. Take Wikipedia for example. The credibility of Wikipedia has been consistently brought into question ever since it was launched nearly 20 years ago. Numerous media stories have noted inaccuracies and inappropriate content in Wikipedia (e.g. [13,110,121]) and some of these stories have even been published quite recently (e.g. [13]). However, work has also shown that, in a general sense, Wikipedia is remarkably accurate. One prominent 2005 study [22] in the journal *Nature* noted that the reliability of Wikipedia content is approximately the same as that in the highly-regarded Encyclopedia Britannica. The study provides a compelling argument for the "wisdom of the crowd" [112] approach taken by Wikipedia and peer production versus traditional "expert-only" approaches to content production.

2.2 Brief Background on Communities Studied in this Thesis

The studies in my thesis take place in the context of two peer production communities that generate structured data, namely OpenStreetMap and Wikidata. We briefly describe relevant details of each community next.

2.2.1 OpenStreetMap

OpenStreetMap (OSM) was founded in 2004 and has been described as "The Wikipedia of Maps" [17]. Its data is used in humanitarian mapping initiatives and in popular applications like Mapbox and Apple Maps.

In OSM, contributors can map out any geographic entities including businesses, roads, parks, bodies of water, and more. OpenStreetMap contributors apply metadata to characterize the semantics of geographic entities. Specifically, they apply *tags* – tags are key-value pairs, where the *key* represents a concept and the *value* indicates a specific *instance* of the concept. For example, a fast food restaurant might include the tag "cuisine = burger".

OSM's tag data model effectively limits contributors to applying one tag with a given key (e.g. one "amenity" key) for any given mapped entity. Although technically semi-colons can be used to separate multiple values for a given key, semi-colon use is discouraged because applications don't always handle this syntax appropriately [122]. A proposed solution is to provide only the '"primary"' value for an attribute [122]. However, as we will see (e.g., for Dairy Queen cuisine), often there is no single '"primary" value.

The OSM community also created and maintains a wiki[2] that identifies "certain key and value combinations for the most commonly used tags, which act as informal standards." In other words, the wiki tells contributors how to tag entities. The wiki specifies information such as relationships between tags – for example, that an entity tagged as "amenity = restaurant" also should include a value for the key "cuisine" to indicate the type of food served. It also characterizes appropriate values for a key – http://wiki.openstreetmap.org/wiki/Key:cuisine recommends the value "burger" for "e.g. McDonald's, Wendy's, Jollibee (Philippines)" and the value "coffee_shop" for places that serve "mainly coffee, [and] may have some light cold snacks such as cakes".

Thus, to the extent that OSM contributors can produce standardized metadata, it is by consulting the OSM wiki and understanding how it applies to the geographic entities they are editing. In addition, several editing tools (e.g. JOSM, Potlatch, iD) help editors by suggesting metadata to apply. We will discuss these tools in more detail in the next chapter.

2.2.2 Wikidata

The success of Wikipedia has helped spur the development in 2012 of its sister project, Wikidata, a major Wikimedia Foundation project described as "a free linked database that can be read and edited by both humans and machines." [123]. Whereas the focus of Wikipedia is on article content, the focus of Wikidata is on machine-readable structured content. Current applications using Wikidata are Apple Siri, Google Knowledge Graph, and even Wikipedia infoboxes.

The structured content in Wikidata describes all manners of concrete and abstract concepts. The Wikidata representations of these concepts are called *items* and exist for everything from people (e.g. "Nelson Mandela") to emotions (e.g. "happiness") to literature (e.g. "The Great Gatsby"). As of this writing, more than 50 million Wikidata items have been created. As in OSM, *key-value pair* structured data are applied to

[2] http://wiki.openstreetmap.org

define entity attributes. Pertaining to say, *The Great Gatsby*, Wikidata contributors have added information to tell us when the book was first published (`publication date=10 April 1925`), who it was written by (`author=F. Scott Fitzgerald`), and what its ISBN number is (`ISBN-10=0-7432-7356-7`).

Unlike in OpenStreetMap, the Wikidata data model has been structured to accommodate multiple values per key. For instance, the "occupation" attribute for the item corresponding to Barack Obama includes both "politician" and "lawyer".

2.3 Performing Contributions in Peer Production

2.3.1 Manual Contributions

Prototypically, peer production contributions have come from direct, manual editing by humans. In this thesis, we will use the phrases "manual editing" or "human editing" to describe this type of contribution. Communities like Wikipedia, Wikidata and OpenStreetMap all allow manual editing and provide user interfaces in which to do so. In communities such as Wikipedia and Wikidata, manual contributions can be performed through two different methods, either via 1) registered contributors or 2) logged-out "anonymous" contributors. While the latter group performs a small proportion of edits [86], research [124,125] has shown that such anonymous manual contributions actually provide significant value to peer production communities. Given this, the work in this thesis intentionally studies both registered and anonymous manual contributions when applicable.

2.3.2 Automated Contributions

While manual contributions do still occur in Wikipedia and other peer production communities, *automated* mechanisms play a large and increasing role as communities have matured and technology has improved. Bot editing has become prevalent in Wikidata particularly quickly due to the relatively straightforward nature of extracting structured data from existing knowledge bases. Broadly, two types of automation are used in peer production: semi-automated tools and fully-automated bots. As a general rule, bots require less manual input to run than tools. However, the line between tools and bots can sometimes be blurry, since these strategies perform many of the same tasks and the level of automation may vary between individual tools and bots. For example, in Wikipedia, AutoWikiBrowser [107] is a semi-automated tool that can perform quite rapid automated editing [19].

Automation has many roles in peer production. Halfaker and Riedl [34] defined a set of four primary uses for bots in the context of Wikipedia. First, bots are '"force multipliers"' of manual work: they can edit faster and longer than manual contributors. Second, bots "monitor and curate...content". For example, they can fix spelling errors or attempt to repair broken links. Third, bots can augment the boilerplate editing software to aid contributors. Lastly, bots identify and counteract malicious behavior such as vandalism.

As bots have begun to play a more prominent role in peer production, they have caught the attention of researchers, resulting in a number of studies (e.g. [18,20,34,78]) seeking to understand their effects on

community dynamics and outcomes. Some of this work has noted complex interactions and tensions that have begun to exist between humans and automation. There has been much debate about the roles appropriate for bots. For example, bots can enforce policy by editing community discussions to include member signatures [18]. However, this use of bots resulted in community pushback against bots correcting human behavior. Skepticism towards bots is justified given some unfortunate experiences with them in peer production. A notable example occurred in OpenStreetMap, with a large-scale import of US TIGER[4] geographic data [126]. This import provided basic map features like highways and rivers as a foundation for OpenStreetMap's US map to build upon [126]. However, the import is notorious as a questionable use of automation since it introduced large amounts of "outdated and erroneous" data [99]. Instead of saving human editors years of manual work, the import has taken years of work to fix [127]. A similar problem occurred in Wikipedia with *rambot* imports of geographic data to thousands of articles on cities [103].

Due to the potential of bots to cause harm, communities have built bureaucratic infrastructures to govern bot deployment. English Wikipedia developed the Bot Approvals Group (BAG) [105] to "[oversee] most areas and processes dealing with bots" [102]. Bots cannot run without BAG approval [105]. Wikidata has adopted a similar approvals process. As will be discussed in the next chapter, OpenStreetMap has a policy called "mapping for locals" intended to ensure automated imports are used only if approved by contributors in the geographic region the imports affect [128]. These approval processes are burdensome, so some contributors seek to circumvent them. To the extent they succeed, they violate policy and hide their bot activities from analysis. The third study in this thesis explores the topic of bot detection. Of particular use in that study are implicit behavioral attributes of bots that have been discovered both in past peer production work [19,31] and in work in other domains [45,46,89,91] such as malware and video games where bots often actively seeks to avoid detection.

2.4 Research Studying Contributor Freedom within Peer Production

Peer production communities have created sets of core community ethos. For example, in Wikipedia, these ethos are known as the "Five Pillars" [104]. Work by Benkler [8] has attributed the ethos of *contributor freedom* as being particularly vital to the success of these communities. This ethos has allowed contributors to concentrate on adding content "as they see fit" [33] without excessively thinking about community guidelines or rules.

Particularly relevant to the first challenge studied in this thesis is work that has shown the contrast in how OSM and Wikipedia treat community norms. The Wikipedia community has developed hundreds of policies (including essays, guidelines, and "official policies") to "[manage…] diverse views" [53]. Palen et al. [75] discussed how OSM differs from Wikipedia: OSM seeks to keep "bureaucracy at bay" as an effort to support "diverse participation". They also note that as OSM grows, "social and technical approaches" for governance are preferred over bureaucratic ones, while Wikipedia managed "community growth through the creation and

clearer articulation of policies". While policy enforcement varies by language edition in Wikipedia (see [88]), Wikipedia, in general, is more policy-oriented than OSM[3].

Policies can have a negative effect on a community. Halfaker et al. wrote that "Wikipedia has changed from 'the encyclopedia that anyone can edit' to 'the encyclopedia that anyone who understands the norms, socializes him or herself, dodges the impersonal wall of semi-automated rejection and still wants to voluntarily contribute his or her time and energy can edit'." [30] Strictly enforced norms certainly have positive effects, e.g. managing vandalism. However, they also may have negative side effects; for example, Halfaker et al. [30] found that Wikipedia's strong and strictly enforced norms reduced retention of new editors. Moreover, Lin [63] has noted that OSM's "lack of established rules" has advantages: it makes for a low "entry barrier" and provides a chance to become more of a community member. In chapter 3, we explore the ways in which OSM's substantial scope for contributor freedom affect its ability to produced standardized data, where following standards is essential for automatic processing. In chapter 4, we quantify the impact of this tension.

2.5 Studies of Content Value in Peer Production

Broadly, many studies have considered the *value* of peer-produced content and have done so in diverse ways. This section provides a high-level summary of the most relevant of such work.

A common method in Wikipedia, Wikidata, and OpenStreetMap to consider content value is by studying content *quality* (e.g. [24,26,29,35,40,44,51,54,64,96]). In the context of OpenStreetMap, many studies have considered content quality based on spatial attributes such as completeness and positional accuracy. Some have also considered the structured tags that are applied to geographic content. A frequent technique has been to compare tags to government and commercial data sources, e.g. [24,65,98]. Unfortunately, comparing OSM tag data with sources like this is not always possible and does not scale well. Hence, other work has sought to develop *intrinsic* measures of the quality of tagging, e.g. [7,41]. One such method [7] considers the mean tag count per OSM record. With this metric, higher averages indicate higher quality tagging and vice-versa.

Data that is less standardized provides less value to applications and algorithms, and is arguably lower quality. The OpenStreetMap community has defined a global set of tagging recommendations[4] that serve to encourage standardized tagging. Work [4] has studied these community-defined recommendations and identified issues including cultural-related problems that stem from using a global set of road tagging recommendations. We study the community-defined tagging recommendations in detail in the first two studies in this thesis.

[3] For the discussions in this thesis, we focus on English Wikipedia.

[4] https://wiki.openstreetmap.org/

In general, work has indicated that tag quality in OpenStreetMap needs improvement. OpenStreetMap's data is characterized by Ballatore and Bertolotto [3] as being "spatially rich, but semantically poor". To facilitate better-quality tagging in OpenStreetMap, some researchers have introduced tag recommendation systems (e.g. [47,93]).

In the context of Wikidata, some studies have considered content quality. For instance, Piscopo et al. [78] used a machine-learning framework called ORES [129] to predict content quality in order to understand how contributions coming from different sources (bots, registered manual contributors, and anonymous manual contributors) associate with content quality. They found that the highest-quality content tends to be that in which both bot and manual contributors provide a significant proportion of edits. They also found that anonymous manual contributions tend to be associated with lower-quality content.

Some peer production research has defined value by incorporating content demand into their metrics. Warnacke-Wang et al. [95] took a consumer-focused approach to considering Wikipedia content value by exploring the relationship between article quality and demand. Per their value metric, high-demand articles should also be high-quality. In other words, quality and demand should be aligned. To obtain a holistic sense of alignment, Warnacke-Wang et al. applied this metric to four large versions of Wikipedia: English, Portuguese, French, and Russian. They found that for the majority of articles, content quality and demand did align. However, they also found that in-demand content that was viewed 2 billion times total a month had significantly lower quality than if the metric were adhered to.

The reason for Warnacke-Wang et al.'s results are likely due to the fact that content demand tends to not influence where contributors edit. As Warnacke-Wang et al. [95] summarized., "consumer…demand is generally not a large consideration in how contributors decide to allocate their work." Aside from Warnacke-Wang et al.'s work, some Wikipedia studies have found content editing and content demand to be positively associated [39] while others have not indicated this is so [25,60]. The final study in my thesis seeks to better understand the relationship between content quality and content demand in structured data found in Wikidata.

Some studies have also defined content value based on a community's ability to represent minority or protected-populations. For example, work in the context of Wikipedia [57,68] and OpenStreetMap [87] found better representation of *male*-related content compared to *female*-related content. Further, peer-produced information quality has been shown to be associated with the socio-economic development status of a region [26,85]. Finally, work in the context of OpenStreetMap and Wikipedia [42,66,100] found that rural areas tend to have poorer quality content then urban areas. The final study in my thesis builds upon such work to explore problematic biases in structured content found in Wikidata.

3 Understanding the Causes of a Tension Between Freedom and Standardization in Peer-Produced Structured Content

3.1 Introduction

As discussed in chapter 1, there is an inherent tension between the standardization needs of structured data and peer production communities' ethos of contributor freedom, which encourage contributors to just "go for it" and employ "trial and error" [114]. OpenStreetMap (OSM) particularly emphasizes contributor freedom, doing so even more than Wikipedia. Where Wikipedia tempers contributor freedom with a set of policies (such as "Neutral Point of View") that are strictly enforced by the community, OSM's "Good Practice" [115] says "Nobody is forced to obey [the OSM guidelines], nor will OSM ever force any of its mappers to do anything."

This tension led us to articulate a central research question: How does OSM's strong commitment to contributor freedom affect its efforts to produce standardized data?

To answer this question, we performed an interview study of OSM contributors. The study focused on the production of OSM metadata ("tags"), investigating community practices that lead to standardization successes and failures.

Our contributions are as follows:

- We show *why* the large degree of contributor freedom affects the ability of peer production communities to be standardized. For example, some contributors – through greater technical skill or dedication to a cause – were able to influence standards. Cultural differences also caused standardization problems – for example, a "highway" can have different definitions in different regions.
- Based on our results, we offer several new sociotechnical strategies and tools to improve standardized data creation in peer production communities. For example, our work problematizes OSM's 1:1 tagging structure, motivates the need to be able to link similar entities, and informs the design of tools that can improve standardization without increasing the effort required to contribute.

We next discuss related research and then our research methods. The heart of the chapter consists of our results, interleaved with discussion of their meaning and implications. We conclude with a synthesis.

3.2 Related Work

OSM has been the subject of considerable research since its inception in 2004. Much of this research has focused on comparing OSM spatial entities to ground truth data (e.g. [24,26,70]). This work mostly has looked at spatial dimensions such as positional accuracy and completeness of entities mapped.

While the predominant focus in the OSM literature is on the spatial entities themselves, some researchers have examined OSM metadata. Some of this work (e.g. [24,70,93]) touched on challenges in ensuring metadata standardization; these challenges helped motivate our research. For example, Girres and Touya compared OSM highway tag data to French BD TOPO® ground truth data in a small portion of France and found roadway standardization problems [24].

External OSM tag editing applications and the algorithms they leverage have played an increasingly important role in shaping the OSM tagging folksonomy. These algorithms do not necessarily enforce the 'informal standards' of the wiki since they factor in observed tagging practice (which may or may not follow the wiki standards). Vandecasteele and Devillers [93] point out the large degree of "semantic heterogeneity" in OSM and propose a solution to mitigate this problem through a tag recommender, OSMantic. This tool uses existing community tagging practices such as tag co-occurrence to recommend new tags for OSM records being edited. Karagiannakis et al. created a similar recommender, OSMRec [47]. While these systems may well help contributors follow community practice, there is no guarantee that practice actually follows standards; if not, these systems end up reinforcing suboptimal practices. Determining how standardization fails is thus of great importance, and we seek to understand it. Our study builds upon standardization work in OSM and seeks to understand the *causes* of standardization problems in data.

A small amount of work has considered the OSM community's process of standard (as specified in the wiki) creation. Ballatore and Mooney considered the negotiation process of the creation of standards found on the wiki [5]. They did so by textual analysis of the OSM wiki and OSM mailing lists. They find some similar impediments to standardization attempts that we find (but through a different method). For instance, they found that cultural differences in road representation have resulted in the community's inability to arrive at a global roadway tagging standard.

3.3 Method

To answer our research question, we performed 15 semi-structured interviews. Our questions focused on the process of creating OSM's metadata standard and applying it while entering specific geographic data, with an eye to identifying reason for standardization failures. The interviews occurred during March and April 2016 via Skype or Google Hangouts. Recruitment was conducted in three ways: two OSM mailing lists ("Tagging" and "Talk"), an OSM forum, and snowball sampling. I and a collaborator performed the interviews (mostly separately, but occasionally together) in English following predefined protocols. All participants consented electronically and verbally in line with our IRB approved-protocol. We compensated participants with a $10 USD Amazon gift card. Participants came from a number of countries, mainly from Western and developed countries (~33% from each of North America, Europe, and Asia). This distribution is generally consistent with OSM demographics [10]. There was a wide-range of OSM experience, from several months to 10 years (see Table 3.1). 80% of participants were male, broadly consistent with general OSM demographics [27]. Most participants edited OSM in a non-professional capacity (i.e. did not edit OSM

Table 3.1: Participant Information

Participant	Sex	Time in OSM	Map Changes
P1	M	4 years	>1,500,000
P2	M	7 years	>250,000
P3	F	3 years	>1000
P4	M	10 years	>2,500,000
P5	M	3 years	>1000
P6	M	2 years	>1,000,000
P7	M	8 years	>250,000
P8	M	7 years	>250,000
P9	M	< 1 year	>250,000
P10	M	7 years	>500,000
P11	F	< 1 year	>100,000
P12	M	6 years	>2,500,000
P13	M	5 years	>1,500,00
P14	M	6 years	>100,000
P15	F	1 year	>100,000

as part of their jobs); a minority edited due to their role in a humanitarian or other organization. See Table 3.1 for additional participant information gathered from a web tool called "How did you contribute to OpenStreetMap?" [130].

To analyze the interviews, we first transcribed the audio recordings and then employed a Grounded Theory approach [71]. I and two collaborators applied open coding to the transcripts. Then I and three collaborators collaboratively reached consensus on general themes of the codes using an affinity-diagramming approach, involving constant comparison amongst codes. This resulted in six themes we describe next.

3.4 Results and Interpretations

Our themes all relate to how the OSM community's attempts to produce standardized data mesh with its commitment to contributor freedom. To help understand the themes more holistically, we introduce a set of concepts that cut across many of the themes.

- **Correctness**. For metadata to be standardized, there must be definitions of "correct" and "incorrect" metadata. We observed two types of standardization issues associated with the notion of correctness: (a) how *complete* must metadata for an entity be for it to be considered correct? (b) how *consistent* must metadata be across different contexts for it to be considered correct?

- **Community**. How does the OSM community manage itself to produce standardized metadata? To what extent are standards enforced? How does the community reach *consensus* on standards? As we consider these questions, it is instructive to compare OSM to the most prominent peer production community, Wikipedia.
- **Code**. To what extent do OSM's data ontology and data entry tools facilitate – or impede – the production of standardized data?

We conclude each of our themes with a discussion to situate it with respect to these concepts and draw implications for design, practice, or research.

3.4.1 Theme 1: Freedom vs. Metadata Completeness

Freedom is fundamental. A number of contributors (P1, P4, P7, P8, P10, P11) explicitly noted that lack of rules is fundamental to OSM's ethos and differentiates it from Wikipedia. While freedom may harm standardization by letting contributors vary the amount of content they provide, contributor freedom often is precisely why participants preferred OSM (P8, P10). Four participants noted that – compared to Wikipedia, where articles must adhere to strict quality standards such as verifiability and neutrality – OSM has less rigid standards and less enforcement (P4, P7, P8, P10). Participant P10's remark was illustrative:

> *Wikipedia to me feels like Germany, too many rules and regulations. (P10)*

However, too much freedom may cause problems. One participant characterized another OSM contributor as taking freedom to an absurd degree by proposing excessive detail:

> *...he suggested that certain stores like supermarkets, have a list of things they sell, brands, and did we wanna put the brands they sell in OpenStreetMap. Well, for Christ's sake. Imagine that! How many brands? You'd have a list of thousands, and they would change all the time! (P1)*

Several other participants explicitly described the lack of clearly defined and enforced standards as problematic. For example, P11 expressed confusion over what tags should be applied to a McDonald's fast food restaurant.

> *Do you need to list, like everything that a McDonald's will ever give you? (P11)*

While OSM seeks to minimize use and enforcement of policies [75], and the community values this, we see that too little direction also can be problematic. Thus, a balance must be struck. We next suggest several possibilities for doing so.

Region-specific information maturity can suggest appropriate contributor freedom. One participant (P4) noted that actual contributor freedom in OSM was a function of the completeness of the map in a given region.

> *…both in Wikipedia and OSM…freedom allows growth to happen quickly, allows iteration to happen fast. With the iteration, you improve and you figure out a system for that place…And this freedom is kind of contextual. Like in the UK for instance [in OSM], it doesn't seem like there's much freedom for things new, and it would be very wrong if they would impose such restrictions in growing communities like India and other places where they say "oh this is how it should be done and this is the right thing". That'll completely stifle growth of the community…(P4)*

Of course, a notion of *localized* freedom also can be problematic; as we discuss later, others desired *global* standards. However, P4 raises an issue that has become urgent for Wikipedia's health – an increase in policies and restrictions aimed at standardization can effect community growth negatively [30]. Palen et al. characterized OSM as of 2015 as similar to the Wikipedia of 2007: it is growing fast and looking to "find ways to cope with and maximize this opportunity." [75] They argue that it is too early to tell if OSM's less policy-oriented approach to handle growth "will address problems of policy rigidity and unfriendliness to newcomers that Wikipedia has faced." (e.g. those pointed by Halfaker et al. in [30]).

Contributors look to "satisfice" – apply "basic" tags and do a "good enough" job (P1, P2, P3, P4, P7, P10, P13, P15). Participants had different definitions of what they took to be "complete enough" when mapping. Adding basic tags to characterize entities was a suggested heuristic. For example:

> *...I always think the main thing is to just get it [a business] on there [OSM], and sort of basically what it is, if it's a restaurant, or you have a cuisine, maybe a website or something, the phone number, or the address. (P7)*

And while contributors believed they put in good faith effort to tag entities, they also acknowledged that they sometimes skipped details – it was too much work or too time consuming to apply all possibly relevant tags to an entity.

> *Yeah, it's too much work to add everything. I just add...minimum I feel is required to uniquely describe, or not uniquely but, to a good extent describe the object to somebody who is new. (P4)*

3.4.1.1 Discussion

Freedom is highly valued by the OSM community and plays a major role in defining and differentiating the community, particularly in contrast to Wikipedia. This is reflected in relaxed attitudes about the degree of metadata completeness required. Participants resolved this in a rough and ready manner, by suggesting a "good enough" heuristic and noting that regions of differing information maturity rightly should expect different levels of completeness.

3.4.2 Theme 2: Project-Specific Freedom and Metadata Correctness: Humanitarian OpenStreetMap (HOT)

HOT is a special interest group in OSM that works to create maps "when relief organizations are responding to disasters or political crises" [131]. HOT and other humanitarian activities play a prominent role in the OSM community: HOT serves as "a driver of OSM's evolution" [75].

At least half our participants (P2, P6, P7, P9, P11, P12, P15) did HOT work, often in Africa. HOT mapping is directed by the HOT Tasking Manager, a tool that lets volunteers do "armchair mapping" (mapping using remote imagery) for predefined regions throughout the world [132] [133].

HOT prioritizes only metadata needed for humanitarian efforts, not metadata completeness. P11 noted that HOT's focus is achieving humanitarian goals, not producing detailed maps. Specifically, providing detail not needed for humanitarian efforts slows contributors down, resulting in fewer objects mapped per unit time. Highways, tracks, and paths were noted as particularly important to HOT, leading to metadata like "highway=path", "highway=track", etc.

> *I feel like I've been told to calm down and like hold back on some of the detail and I think it's because it's specifically for that project [HOT]; like I wanted to [map] every single house, and I wanted to tag whether I thought it was a house or whether it was a commercial building. I was going into all this detail…And someone was like 'just calm down, we're trying to save lives here we don't need like the most, like intricate map'. But I also really want to put in pretty much every bit of information. My theory is that eventually someone's gonna need that information… (P11)*

> *…[HOT wants her to] 'tell us how big the village is and tell us where the roads are'. They're really particular about 'we only want highways and paths and tracks and that's it' and I'm like 'there's so much more detail' (P11)*

HOT may redefine metadata meaning, leading to inconsistency with global standards. Participants observed that HOT contributors both invent new tags and redefine global definitions of tag meaning to achieve their goals.

> *The NGOs in HOT…invent their own tags...so they sort of define what they want (P12)*

> *…we do bend the rules slightly with Humanitarian OpenStreetMap for the humanitarian work. For instance, in the earthquake in Nepal, we were trying to help the aid agencies to reach remote places. So what we were doing online was trying to identify helicopter landing sites. And what we did was, we found the feature for a helicopter-landing site. And what we would do is look for an open area about 30 meters...A clearance of 30 meters that was level land near a village, and we will label that as a helicopter landing site. (P6)*

In other words, HOT appropriated and redefined the tag for a "helicopter landing site" to meet their immediate needs (for the April 2015 Nepal Earthquake). Interestingly, in less than a month, the OSM wiki was updated with a new tag for this purpose, "emergency=landing_site". This is a vivid example of HOT bending OSM's global rules, leading to (at least temporarily) inconsistent semantics, and eventually driving the evolution of the global ontology.

3.4.2.1 Discussion

HOT exploits OSM's freedom to meet its own needs for metadata correctness: only as complete as necessary for humanitarian work and inconsistent with global standards if needed. It is instructive to compare HOT to projects within Wikipedia; Wikipedia includes many *WikiProjects*, dedicated to improving Wikipedia coverage of specific topics. Like HOT (and other OSM projects), WikiProjects can define new information structures as needed. However, WikiProjects must abide by global Wikipedia standards such as the "five pillars" – not doing so, even temporarily, is considered a policy violation [134] and is grounds for immediate reverting. In contrast, as we have seen, HOT may choose not to follow global OSM recommendations for metadata completeness and may adopt inconsistent semantics for global metadata concepts.

HOT's approach helps it achieve important humanitarian aims. However, as P11 noted, OSM data is persistent and potentially of interest over time and in multiple contexts. Thus, there is a tension between HOT's need for focus and speed and desire for generally useful and globally consistent metadata. We will revisit this tension in the next themes.

3.4.3 Theme 3: Cultural Differences Make Global Metadata Correctness Standards Difficult to Achieve and Maintain

Since OSM is a global community, contributors come from different cultures and speak different languages. Our participants often mentioned how language and culture differences make it difficult to achieve globally consistent definitions of metadata correctness.

Must Western definitions be followed globally in OSM? (P1, P6, P10, P11, P15). OSM started in London, and participants noted the strong Western influence on the OSM metadata standards. This was reflected in tag naming conventions, available tags, and tags considered important.

> *Why should we be using British terminology just 'cause it started in Britain? (P1)*
>
> *...a lot of the tags are very Western world-centric, you know, so there's lots of amenities and services for people in the African countries where volunteers are serving and tags don't exist for those kind of things yet...one of our volunteers in Zambia has told me that in Zambia they have what's called a maternity waiting shelter and it's right next to their rural health centers so it's a space for women to stay at the shelter until they give birth... (P15)*

Prior research also found that global road standardization in OSM is a complex problem that has not been solved satisfactorily [5]; our participants likewise complained that the global road classification system did not accurately describe roads in all areas, particularly outside the global West. For example:

> *[In the context of humanitarian mapping in Africa] Someone says "this is a highway", and I'm like I disagree. And I'm really afraid to map that as a highway if I think, for example, like a vehicle can't go down it and that sort of thing. And I think there's a lot of conflict between what sort of roads people use to differentiate between those things and what roads people think are important or not…people see things differently. (P11)*

Different language editions differ in metadata correctness standards (P3, P5, P6, P7, P10). The OSM wiki (like Wikipedia) has language-specific versions. Some tags exist, or are documented, only in certain language versions.

> *[P5 noted that in Japan, navigation can occur by using neighborhood police boxes] This [system] is, as far as I can tell, a totally ubiquitous part of the urban landscape in Tokyo and most Japanese cities, and a really important way-finding mechanism...I think that's because it's just not a phenomenon that exists in too many other countries, and as a result the tagging documentation only exists in Japanese.*

Further, a tag might have different descriptions in different language versions of the wiki, thus, there is no globally agreed upon definition of correctness.

> *…there are Dutch translations, or French translations, or Spanish translations…they don't say the same things (P10)*

As a specific example, P7 (an American) noted that different language versions specified wholly different tagging conventions (e.g., McDonald's could be "amenity= fast_food" in one, and "amenity=restaurant" in another).

> *…last year I found that there's this whole parallel tagging scheme, so I tried to sort of communicate with the guy who has been working on that page, I think it was in German or something, and sort of, we sort of, if you think one of these is better than the other, and we didn't really come to anything… (P7)*

It is interesting to note that all these phenomena also have been observed across Wikipedia language editions [12,36,37,62]. OSM tags that occur in only some languages correspond to "concept-level diversity" [37] in Wikipedia. and differences in tag descriptions correspond to "sub-concept-level diversity" [37]. The parallels with Wikipedia engender promising research possibilities. For instance, the methods used by Hecht and Gergle [37] and Bao et al. [6] could be leveraged to assess and visualize the differences in spatial attributes in different parts of the world. Similarly, work on Wikipedia has shown that the diversity between language

editions surfaces in algorithms that leverage Wikipedia data [37]. It would be interesting to see if that also occurs with algorithms that utilize OSM data.

Local tagging practices are preferred over (conflicting) global correctness standards (P1, P3, P5, P7, P8, P9, P15). P3 (an American) noted that the wiki contained conflicting descriptions of what constitutes "correct" metadata. When she sees multiple recommendations for tagging an entity, she looks at how others have mapped that same entity in the area she is mapping to select "locally appropriate" tags.

> *Yeah, I know I have [seen conflicting advice on the wiki pages]…when that happens I guess I just pick a way, whatever I guess, if there's one way that seems to be prevalent in this area. (P3)*

Another American participant noted that she believes that the tags she applies are correct per wiki guidelines, but that she also values *"the local knowledge of the people involved…" (P15)*; thus, her tags correspond to *local participants'* ideas of metadata correctness (this mapping philosophy also has been reported in prior work [52]). In general, our participants – who typically were *not local to the areas they mapped* – emphasized the value of mapping for the sake of local accuracy. P8, who had spent time mapping locally in Thailand, said just this – while also reporting that everyone does not take this approach (similar to the issues Wikipedia has had representing "Indigenous knowledge" [94]).

> *I believe the map has to be usable by locals, so for Thailand, it has to be Thai script [the name tag], but it's not a map for foreigners, made by foreigners, but it should be a map for locals. So that is some deviation, likely other mappers don't do. (P8)*

Working towards a global consensus for tag definition (P1, P4, P5, P6, P7, P8, P12, P14, P15). In tension with the previous point, many participants expressed a desire to develop a globally consistent consensus on tag meanings. This is a challenge, particularly given different language representations of the same entity.

> *...there was a discussion on tagging the streets in a way that, for shop=marine, and people in the US it made sense that that's a place that would sell boating supplies, or gear, things like that, but people in Europe, and the UK were like no, that's chandlery…So there was some discussion, and I think settled on boat supplies or something, it was something that would be easily understandable around the world, so until you get that sense of how worldwide it is, and how diverse all the people are around the world, it's easy to just assume I know the right way, this is it, and completely not understand that there's other concepts, there are other things to call them. (P7)*

P7 noted that the community insists on globally applicable rather than region-specific tags, regardless of the entity being mapped, resulting in these different interpretations of entities across cultures/languages. Similar to the "culture clashes" noted by Lin [63], P6 stated that defining a general tag that makes sense on a global scale is difficult.

> *[related to tagging apartment buildings] The result is that conflicts can occur because a [tag for an apartment] can be different things in different countries…What generally happens [hypothetical scenario] is that the French, if they're trying to set up a specific local feature, would actually come on to the OSM and discuss it. They'd put a proposal on the wiki page, and open it up for discussion on the tagging forum. And then everyone would come in and discuss the possibilities, how that would fit in…and how it can be used. And then after the proposal, they go for a vote to actually vote whose version is going to come up as being the accepted version. And after the vote takes place, then a wiki page is set up which describes that feature. (P6)*

3.4.3.1 Discussion

We saw that the OSM community insists on a global set of tags (i.e. the French do not have their own French-specific tags), reflecting a desire for consistent global standards. However, there also is a preference for local tagging schemes to be given priority over other, more global schemes (i.e. if both "highway=path" and "highway=primary" make sense for a road in Africa, the one preferred by locals should be used). This tension can result in a local-versus-global tug of war over metadata correctness that can be confusing to contributors. This is analogous to the distinction between personal and public tags in the folksonomy literature. In studying the movie recommendation site MovieLens[5], Sen et al. [84] found that personal tags (analogous to locally appropriate OSM tags) were not as valuable to the community (analogous to OSM as a whole) as they are to a given person (or, for OSM, a local community).

This observation is consistent with the commentary of Ballatore and Mooney [5]. They stated that OSM's "global, universalistic scope…clashes with the heterogeneity of its contributors and objects of interest" and also noted that "Recurrently, contributors set off to find a universalistic conceptualisation and, after encountering insurmountable problems, resorted to more contingent and localised approaches."

There are two possible paths for resolving the tension participants reported between global standards and local knowledge. First, OSM could privilege the local, allowing language-specific tags and tagging schemes analogous to language-specific versions of Wikipedia. This would help purely local applications, say routing applications within a specific country or region. On the other hand, intra-regional comparisons and applications would suffer. Second, OSM could insist on a global set of tags and tagging standards. Several possible steps could make this effective across regions and contexts, including: (a) clearer definitions of tag meanings and conditions for applying them, (b) (as suggested by P4) use of pictures on the wiki to help contributors understand the physical entity corresponding to a given tag and minimize misinterpretations due

[5] https://movielens.org

to cultural differences (the use of pictures has been successful for similar purposes in other multi-national online systems [28]).

3.4.4 Theme 4: Community-Management Obstacles to Achieving Consensus

The OSM community uses various online media to discuss proposals for new tags and related topics. These include many forums and mailing lists. According to our participants, these forums have problems that make it hard for the community to achieve a clear and unambiguous metadata standard, and thus make it hard for contributors to produce correct metadata.

Standards proposals lack authority. Some participants did not like the voting process for proposed tags (P3, P10). Specifically, since so few people participated, our participants did not consider the process to be a valid way to reach a community consensus.

> *Think about the thousands of people who contribute to OpenStreetMap and think like it shouldn't just be 12 people deciding on this [proposed tag]. (P3)*

Online discussions are often unproductive (P2, P4, P5, P8, P13). Five participants described the online communication among community members as unproductive. A veteran contributor from Thailand was particularly frustrated with discussions related to how to map in the US.

> *So I tell ya, I see a lot of discussion, particularly in the US map the discussions are just endless and they talk and talk and talk and they don't actually do anything. The US map, I shouldn't say this but I think it's in terrible condition. (P13)*

P5 felt that stronger leadership was important for OSM:

> *[Wikipedia has] arrived at a stronger set of norms, governance. They definitely have their own problems, but they have a strong central leadership organization and have professionalized in a way that OpenStreetMap has not…I think [OSM should be more similar to Wikipedia]. (P5)*

Hostility and toxic behavior online (P3, P4, P5, P9, P15). Five participants said that hostility/toxicity was a problem in the mailing lists and other communication media.

> *Oh boy so the lists can be really great and an awesome way to keep the community connected and supportive of each other…but they [emails] can also be like horrible…I know there's been rifts in the past and sometimes the email list can be very toxic…people feel like they can say things that they wouldn't say to someone else's face. (P15)*

P3 also mentioned that she had "heard of women not being listened to or respected".

Another participant noted that small sets of contributors with rather extreme viewpoints were given more credence than they should. In a particular instance, the effectiveness of HOT was brought into question.

That is a fringe minority opinion that HOT could be doing bad work, and everyone's entitled to spout off their beliefs, but without the kind of strong leadership that can establish a vision for the project, these ideas get way more credence than they deserve. (P5)

3.4.4.1 Discussion

Lam et al. noted that psychological research has shown that while large groups are more prone to conflict, "large and diverse groups can make better decisions than individuals or experts" [56]. Their study of Wikipedia found that more contributors increased quality while warning to "be wary of decisions that are made by groups that are very small" [56]. This research reinforces the idea that OSM decision processes would benefit by involving more people.

Further, as in many online communities, toxic behavior and inefficient communication reduce community effectiveness and member satisfaction. Similar issues exist in Bitcoin, another open source community that lacks strong leadership and enforced norms [135]. Crucial updates needed to handle the increased popularity of Bitcoin were hindered due to ineffective leadership [135]. As P5 stated above, strong, proactive leadership as in Wikipedia could improve communication, likely resulting in increased productivity. Improved productivity could reduce the number of people who are frustrated with the decision-making process (like P13), and make them more inclined to participate in that process, which could increase its authority.

3.4.5 Theme 5: Data Representation Prevents Conceptual Correctness

In the OSM data model, only one value is allowed per key for any record; for example, a record representing a Dairy Queen shop cannot be tagged both "cuisine=burger" and "cuisine=ice_cream". A possible workaround is to concatenate multiple values with semicolons (e.g. "cuisine=burger;ice_cream"). However, this workaround is rarely used: it breaks most map rendering applications, and the OSM wiki recommends that contributors apply only the "primary" attribute of an entity [122].

However, the obvious problem is that many real-world entities have multiple valid values for certain attributes – a Dairy Queen does serve **both** ice cream and burgers. But OSM's data model forces contributors to choose one, meaning that the entity cannot be represented correctly. Participants (P10, P14) described examples of this problem:

...you have to map the hotel, and the restaurant, but sometimes it's just a restaurant with a hotel upstairs, so it isn't really two things, it's just one thing. So you have to say it like this is a restaurant with rooms available or something like that. (P10)

3.4.5.1 Discussion

To make sense of the problems noted here, it is useful to refer to Lessig's notion of *code* [61]. He observed that the code of a system constitutes an architecture that constrains human behavior by allowing, forbidding, encouraging, or discouraging certain actions. Specifically, the OSM data model – adopted for certain

technical reasons – constrains contributors' ability to represent the real world correctly. And thus, changes to the code – the data model – are needed to enable correct representation.

An existing OSM technical solution is to use the concept of "relations" to link two geometries (e.g. the restaurant and hotel geometries) into one entity. However, relations are rarely used in OSM, and do not accurately capture the real world semantics anyway. A more fundamental change to OSM's code, namely to allow multiple values per key, would solve the "multiple values per key" problem directly. Of course, it is likely that other parts of the OSM code base and applications that use OSM data rely on the one-to-one assumption, so such a change would have to be designed and implemented carefully.

3.4.6 Theme 6: Data Entry Tools May Harm Metadata Correctness and Privilege Certain Users

As we mentioned, OSM contributors use a number of tools to facilitate the data entry process. Despite their benefits, they also raise problems of metadata correctness.

Data import tools. The OSM community is wary of bulk import of data from external sources "because poor imports can have significant impacts on both existing data and [the] local mapping community." [136] Many participants (P1, P4, P5, P6, P8) expressed similar concerns that data imports can cause problems of metadata correctness.

> *They did a big data import from the New York State Department of Environmental Conservation and areas that are wilderness [they incorrectly tagged those areas]. "Landuse forest" [a tag], they're totally wrong… (P1)*
>
> *…any import is viewed extremely skeptically, and the only ones that happen as a result are the ones that nobody talks about and that are done stealthily and improperly. (P5)*
>
> *...if someone is importing a bogus tag...it comes up with a high number of occurrences, so just counting the number is not the correct method to say this is widely in use. So you would have to count a little bit more, you would have to count how many individual mappers have used it. (P8)*

3.4.6.1 Discussion

Searching for places of interest is a common use case for applications built on OSM data. Business places specifically have an incentive for accurate data representations in OSM, since they want to make it easy for potential customers to find them. Two participants (P4, P12) suggested that businesses update their own metadata in OSM – P12 specifically called for OSM to make it easier for businesses to do this. Since businesses maintain their own databases, this would require OSM to extend its sociotechnical code by developing data import tools and data correctness monitoring tools that enable it to trust bulk imported data.

However, there is another technical problem that would have to be solved to enable this. With current OSM tagging practices, it is not easy to identify accurately and reliably the OSM entities that correspond to

instances of a given business, e.g., Starbucks. Instead, user-assigned entity names (which often are inconsistent with official business names) and other entity attributes must be examined to infer that an OSM instance actually is (say) a Starbucks coffee shop. Modifying the OSM code to provide unique identifiers for real-world entities (say, a single ID for all Starbucks), for example, is a reasonable approach to solve this problem (which is somewhat similar to the suggestion by Girres et al. [24]). P5 explicitly noted that this approach could facilitate metadata import by businesses and other organizations.

Data Entry Tools. Widely used OSM editing tools such as JOSM and Potlatch facilitate the metadata entry process by suggesting 'preset' tags for entities. For example, a preset for tagging fast food restaurants suggests some relevant keys and likely values for the keys. From the perspective of metadata correctness, however, we note that (a) each tool implements different presets – P9 and P2 mentioned this; and (b) the source for the suggested presets is unclear; for example, P6 thought they were based on actual tagging practice. More fundamentally, our participants were unsure whether these presets were consistent with the metadata standard laid out in the OSM wiki (P1, P2, P4, P5, P7, P8, P9, P10, P12, P14). Some thought they were; others did not:

> *…there is not a single source for those presets…it's all manually done by the developers of the tool. (P8)*

> *[JOSM is] pretty much a direct implementation [of the wiki] (P9)*

Some of our participants noted another aspect of these tools: the *power* of their code [61]. While OSM has little central structure and the community's ethos champions contributor freedom, the data entry tools contributors use constrain their actions while elevating some contributors' positions:

> *If you really want to push for specific tagging…you have to get the tool [that uses the desired tagging] in [use in OSM]. So if something is a preset in JOSM, if something is a preset in iD, and the stand up map is rendering it, you have a very high chance of that being actively used (P8)*

3.4.6.2 Discussion

These observations strongly remind us of the power of code [61]. Of course, it is natural and useful for data entry tools to suggest plausible metadata for entities as they are entered. However, to the extent that the community has a standard for metadata correctness, the tools should base their suggestions on that, rather than on observed practice (which may or may not accurately follow the standard).

Further, P8's quote makes explicit the 'politics' of code, "the conditions of code and software in relation to power, capital, and control." [58] Those who create (or influence the creation of) OSM tools are exerting control over the community [58]. This perspective is consistent with that of Lin, who argued that "technical skills…are interlinked with the roles one holds" in OSM [63]. Another example of this in OSM can be seen in several participants (P1, P7, P15) mentioning that editing the wiki requires technical expertise, and that not all OSM contributors felt comfortable doing so. If editing the wiki were easier, more contributors might

be able and willing to participate creating and maintaining the standard for metadata correctness. Once again, this could be done by modifying the OSM sociotechnical code, for example, by training contributors in use of visual editing tools.

3.5 Reflecting on Correctness, Community, and Code

We now take the opportunity to summarize and reflect on how the concepts of **correctness**, **community**, and **code** let us make sense of our results.

Theme 1 emphasized how OSM contributor freedom ("Nobody is forced to obey") may result in incomplete (and thus incorrect) metadata. Contributors may settle for their own notion of "good enough", with no assurance that they supplied sufficient metadata for applications that use the data.

Theme 2 showed how an OSM project, Humanitarian OpenStreetMap (HOT), uses OSM's freedom to ignore global standards of metadata correctness, both completeness and consistency. They do this because they prioritize achieving their humanitarian goals; however, we also saw that their efforts help drive the global evolution of OSM. We also contrasted OSM's community management with Wikipedia by comparing the latitude given to HOT and WikiProjects.

Theme 3 examined a set of culture- and language-based issues that impact metadata correctness and related community practices. The Western origin of OSM has led to the OSM ontology embodying Western concepts, for example, of what constitutes a "highway" or "helicopter landing site". This reminds us of the power of code to constrain behavior. As contributors map in non-Western regions, they may have to appropriate and redefine the meaning of these concepts to make sense in their context, resulting in inconsistencies between "global" (more accurately, Western) and local standards. Moreover, different language versions of the OSM wiki – its metadata standard – may differ in their treatment of concepts, again putting a truly global correctness standard out of reach.

Theme 4 showed that the large amount of freedom OSM contributors enjoy and its weak community management can exacerbate common online community problems of inefficient communication and toxic behavior. Particularly important for our concerns is the effect of the problems on the development and propagation of metadata correctness standards. Some of our participants expressed concern that only a small, but very active group of people participates in the creation of these standards. In addition, toxic behavior, including sexism, can systematically exclude many contributors from participating in this process. This may result in standards that reflect the preferences of only a small group and not those of the entire OSM community.

Themes 5 and 6 emphasized the power of code in shaping OSM and OSM contributors' behavior. OSM's data model makes it impossible to represent certain real world entities – for example, a fast food place that serves both ice cream and burgers – in an intuitively correct way. OSM data entry tools facilitate metadata entry, but can propagate metadata practices inconsistent with the OSM standard. In effect, because they are

used by OSM contributors, they may define a second, unofficial standard of more power than that laid out in the wiki. We also saw that OSM community members with greater technical ability and access – for example, knowledge of how to edit the OSM wiki or ability to modify a data entry tool – had greater power to influence the nature of OSM data. *Code* is especially powerful in OSM because *community management* is (intentionally) weak.

4 Understanding the Effect of a Tension Between Freedom and Standardization in Peer-Produced Structured Content

4.1 Introduction

While the study in the previous chapter sought to understand the ways in which the tension between freedom and standardization manifests itself, the study in this chapter sought to understand the *extent* to which contributor freedom has been impacted by this tension.

In addition to core principles such as contributor freedom, we have noted that peer production communities also establish guidelines and rules to promote quality and consistency. For example, the pages in the OSM wiki serve as "informal standards" for OSM contributors. Given the prevailing ethos of contributor freedom, these guidelines and rules often require interpretation and generally do not *have* to be followed. As we have noted, there is good reason for such an attitude; when rules have proliferated and their enforcement has grown strict, productivity of peer production communities tends to decline [30]. Given the non-binding nature of the "informal standards", we posed a question:

(RQ) How do OpenStreetMap's 'informal standards' relate to actual contributor practice?

In this chapter, we use the term *standardization* to refer to the process by which OSM contributors orient their practice with the "informal standards" of OpenStreetMap.

We first investigated standardization by analyzing the extent to which OSM practice is consistent with the guidelines in the OSM wiki. We found that most applied metadata (or "tag" data) is in fact consistent with the wiki. However, this analysis revealed a second important observation: due to properties of the OSM data model, many of the guidelines cannot be complied with fully. Specifically, in a number of cases, the wiki accurately identifies multiple appropriate values for a given attribute: for example, a "Dairy Queen" serves both "ice cream" and "burgers". However, the OSM data model restricts each attribute to have a single value. This *ambiguity* is a problem for applications that use OSM data because entities are only partially described. Our analysis also led to a third observation about the OSM standardization process: the wiki guidelines reveal many unmet opportunities for applying metadata. For example, operating hours and phone number data are identified as relevant to many business entities, but are rarely applied. These data are commonly used when available by popular mapping services such as Yelp and Google Maps, which indicates their user demand.

Our work contributes by shedding light on the nature of standardization in OpenStreetMap as follows:

- Most applied metadata is consistent with the standard.
- The constraints of the OSM data model lead to a large amount of ambiguous metadata.
- The informal standard of the OSM wiki defines large unmet opportunities to apply useful metadata.

In the remainder of this chapter, we first discuss related work and provide additional relevant background about tagging in OSM. We then discuss our specific research context and methods. We next present our results. We conclude by discussing how our findings motivate changes to OSM and peer production more generally. Specifically, we discuss changes related to sociotechnical tools, data model structure, and community informal standards.

4.2 Related Work

In this study, we examined how well tagging practice and the OSM wiki align. Several previous tag standardization studies have considered the wiki, e.g. [15,69,80]. Our work differs from such studies by systematically analyzing a substantial portion of the wiki to extract tag application "guidelines" and determining the adherence of each tag to the guidelines over a large number of OSM records. For example, we consider whether the tags applied across thousands of McDonald's OSM records are each applied in accordance with wiki guidelines.

Further, an important part of the wiki standards are the prose-heavy descriptions that describe what entity characteristics can be represented through tags. Our robust, large-scale *qualitative and quantitative* approach involved analyzing and interpreting the wiki instructions, including this prose. This analysis aimed to follow the same process that OSM contributors can (if they choose) follow when mapping. This analysis approach is novel in the context of OSM research and led us to identify data standard and data model issues that were not discovered or analyzed in prior work.

We also build upon the work in the previous chapter and elsewhere [4] that has identified challenges in creating or following the OSM wiki. More specifically, the current study complements the previous chapter, which found that OSM struggles to craft the wiki to represent the views of all its contributors. This is due to problems such as cultural differences and toxic behavior by some contributors. We quantify the effect that such problems have on data standardization. Our prior work also identified the data model/data standard issue that results in what we call *ambiguous* data. We quantify this issue here.

4.3 Additional Relevant Information on OpenStreetMap Tags and Tagging Standards

As discussed in more detail in chapter 2, OSM's tag data model effectively limits contributors to applying one tag with a given key for any given mapped entity. This fact will be relevant to the study in this chapter.

In this chapter, we refer to the OSM wiki as the OSM tagging **standard**. As discussed, it provides informal tagging guidelines, offering guidance on how to apply tags via pages often consisting of significant amounts of unstructured text. Many such pages are specific to a given key or tag. For example, the wiki describes the tag "amenity=fast_food" as appropriate "for a place concentrating on very fast counter-only service and take-away food." [137]

As has been noted, contributors are not required to follow the wiki guidelines. However, the wiki represents common community tagging practices and consensus on how tags are intended to be used. As the community changes and grows, the wiki evolves. Such wiki modifications are often performed in accordance with a tag proposal and voting process [138]. This process determines what new tagging-related content should be added to the wiki.

In this chapter, we define tag **standardization** as the process of orienting contributor tagging practice with the informal standard of the OSM wiki. "Standardization" refers to the extent that tagging practice unambiguously adheres to the wiki guidelines.

4.4 Motivation for Analyzing Chain Business Standardization

Since the OSM wiki consists of significant amounts of prose tag descriptions and application instructions, comparing this informal standard to tagging practice is hard. It is not practical to manually compare every wiki page with each of the more than 500 million OSM records in our dataset. We therefore needed to find a tractable approach to measuring standardization.

We did this by identifying a large and interesting subset of entities with substantially similar structure, specifically *chain businesses* such as McDonald's, Starbucks, Safeway (a major U.S. supermarket chain), and Wal-Mart. Each individual McDonald's restaurant, Starbucks coffee shop, or Safeway supermarket is similar to every other one. This contrasts to boutique or "one-off" businesses, where each instance is potentially unique, and thus manual information lookup and analysis would be needed to determine whether an OSM representation follows a standard. Because of the standardized nature of chain businesses, all OSM records for say, U.S. McDonald's restaurants (likewise for other chain businesses) should be tagged substantially the same. These observations lead to a tractable process that yields a conservative estimate of standardization: group all OSM records for a chain business; identify the "substantially similar" metadata instances of a chain business should have; and determine whether individual OSM records have that metadata.

Focusing on chain businesses yields an approach that scales, but analyzing these businesses also has additional value. First, they are popular: McDonald's accounts for over 17% of the fast food market share in the United States [139], and over 40% of Americans visit a Wal-Mart each week [81]. Second, chain businesses have been under-studied by geographic HCI researchers and the broad social computing community, with most projects focusing on the discovery of boutique venues. Third, fast food restaurants and convenience stores (both of which we analyzed) are more prevalent in low SES areas [59]. Chain businesses are therefore important to the populations of those areas, who tend to be underserved in peer production [26,44]. Finally, and critically, if OSM contributors cannot apply tags in a standardized way to real-world entities that are in fact highly standardized, it is unlikely that less similar real-world entities will be standardized well in OSM either.

4.5 Methods

4.5.1 Clustering Algorithm

OSM does not provide a widely adopted formal means to link together different instances of the same business (or any other conceptual category). Therefore, to analyze standardization of chain businesses, we first needed to *extract* chain businesses from the OSM dataset, which we did by developing a clustering procedure. We handled inconsistencies in OSM representations of businesses through a hybrid clustering approach that combined automated algorithms with manual verification and coding. We detail our procedure next.

4.5.1.1 Selecting OSM Instances for Analysis

We used United States OSM data records from February 2014 that were available from [140]. Although our initial data contained records from outside the U.S., we used a U.S. census Tiger shapefile [11] to filter out these records. Our dataset contained the current state of all OSM data records (node and way objects) in the 50 U.S. states. This included roads, bodies of water, and other entities. We limited records to the U.S. because our manual coding process required familiarity with the business data, and all our coders are from the United States. We removed non-business records by filtering based on tags. For example, we identified non-business tags (e.g., "amenity=university") through manual inspection of the dataset and removed all records with these tags. This initial filtering step did not remove all irrelevant data; subsequent normalization and clustering steps were necessary.

4.5.1.2 Normalizing Instance Names

A naive approach to clustering would group all records with the same value for the "name=" tag. However, a "name" can appear inconsistently; for example, McDonald's locations have names ranging from the standard "McDonald's" to "McDonald's – East Liberty Station" to misspellings and variations in capitalization. We reduced these inconsistencies by 1) normalizing tag case, and 2) using Wikipedia redirects to help recognize naming variations of the same business. Wikipedia redirects link common name variations in Wikipedia searches to a central article. For example, Wikipedia redirects a search for "Chik-fil-A" (a fast food restaurant) to the article entitled "Chick-fil-A". Our method sought to capture naming variations similar to these by making the "normalized" name field available to the automated clustering algorithm discussed next.

Table 4.1: Chain Businesses Used in Standardization Analyses

Chain Business						
7-Eleven	Best Buy	CVS	Home Depot	Panda Express	Sam's Club	Taco Bell
Applebee's	Burger King	Dairy Queen	IHOP	Panera Bread	Sonic	Wal-Mart
Arby's	Chevron	Denny's	Jack in the Box	Pizza Hut	Staples	Walgreens
AutoZone	Chick-fil-A	Dollar Tree	KFC	RadioShack	Starbucks	Wells Fargo
Bank of America	Circle K	Dunkin' Donuts	McDonald's	Rite Aid	Stewart's	Wendy's
Barnes & Noble	Culver's	H-E-B	Olive Garden	Safeway	Subway	Whataburger

4.5.1.3 Automated Clustering + Post-Processing

We used a semi-supervised clustering algorithm [9] to further improve clustering results. The algorithm clustered instances based on their tags (including the "name=" tag and the redirect name created in the previous step). We manually selected the 50 largest business clusters, which represented some of the most common businesses in the United States.

We next performed a series of manual steps to ensure the precision of clusters. First, we combined several clusters that represented the same business (e.g. three McDonald's clusters, two CVS pharmacy clusters, etc…). Second, we only retained instances that had the 'standard' name for a business ("mcdonald's"), or small variations ("mcdonalds" or "mcdonald's - east liberty station"). This process resulted in the 42 distinct clusters shown in Table 4.1.

We explicitly highlight that our clustering process reflected a need for high cluster precision that was essential to the accuracy of our tagging standardization analysis. This is because the goal of the analysis was to compare tags applied to the instances of a business cluster – say McDonald's – to the wiki instructions that describe when those tags are appropriate. The comparison only made sense if all instances in the cluster did in fact represent McDonald's instances. If say 25 instances in the "McDonald's" cluster should belong to a "Safeway" cluster instead, we might falsely conclude that the tag "shop=supermarket" applied to those instances was applied in a way that was misaligned with wiki instructions. As noted previously, to avoid this problem we manually inspected "name" tags in our clusters to ensure instances were placed in appropriate clusters. Although we prioritized precision over recall, we note that our clustering approach identified some of the most common tagging practices for each business we analyzed.

After clustering was complete, our largest cluster was McDonald's, with 3424 instances. Our smallest was Sonic (a fast food restaurant), with 169. The mean number of instances across all clusters was 672 (s.d. = 674) and the median was 343. Across all clusters, there were 28,420 business instances total.

As mentioned, chain businesses are inherently standardized in the real world because instances of a given business share many characteristics (e.g. all Dairy Queen locations serve ice cream, all Starbucks have

operating hours, many McDonald's have a drive through). In our analysis, we focused on the tags corresponding to these inherent similarities. By focusing on the metadata that represents inherently standardized attributes of entities, our analyses should provide an upper bound of their standardization. Given these considerations, we removed, for example, tags related to the specific address of an instance (street address, city name, etc.) and miscellaneous notes pertaining to the instance (e.g. "created_by", "note", "attribution", etc.).

Further, to ensure manual coding was tractable, we selected the 10 most applied keys for each business and their associated values; this resulted in 41 distinct keys and 416 distinct business and key combinations, or "business-key pairs", collectively comprising over 94% of the remaining metadata in our clusters. Since we chose the most applied keys, this data also represented the most common tagging norms in terms of key applications in each respective business cluster.

4.6 Determining a Metadata Taxonomy

Different tags in the OSM wiki serve different descriptive roles. Certain tags are *appropriate for all instances of a given business*. Examples include tags like "amenity= fast_food" for McDonald's. Other tags contain a key that is *appropriate for all instances of a given business, but whose value is instance-specific*. This includes tags such as "opening_hours=<some operating hours value>" in the case of many businesses. Finally, other tags are *appropriate for some – but not all – instances of a given business*. This includes tags such as "drive_through=yes" to indicate the presence of a drive through at McDonald's. We developed a taxonomy to account for these different types of metadata. This taxonomy provides a foundation for evaluating the community's standardization process. We defined three classes of metadata:

- *Universal* metadata describe key-value pairs appropriate for all instances of a business. All U.S. Starbucks have the same brand, so all Starbucks instances can be tagged "brand=starbucks". The "brand" attribute has one value for all Starbucks instances.
- *Universal-Varying* metadata describe keys appropriate for all instances of a business, but whose values are instance-specific. All McDonald's locations have an operating hours attribute which can be denoted in OSM with the "opening_hours" key. The specific value appropriate for the key representing that attribute varies across instances of McDonald's.
- *Contingent* metadata describe real-world variation, i.e., keys that may or may not apply to any given instance of a chain business since the attribute they represent may or may not be present (we discuss metadata describing nonexistent attributes later). For example, some McDonald's locations have drive-through windows, and some do not, some are wheelchair accessible, and some are not, etc.

Table 4.2: Chain Business Metadata (Key) Role Classes

Universal	Universal-Varying	Contingent	
shop	ref:store_number	drive_through	delivery
amenity	building:levels	fax	smoking
contact:website	opening_hours	wifi	outdoor_seating
alt_name	contact:phone	area	contact:fax
cuisine	phone	operator	wheelchair
drive_in	building	fuel:diesel	atm
website		motorcar	dispensing
brand		fuel:octane_91	landuse
url		highway	internet_access
takeaway		entrance	

We categorized each key associated with a business as Universal, Universal-Varying, or Contingent. We did this categorization for keys (not tags), since a key can have only one value for a given OSM record, so, for example, the "amenity" key could not be both Universal and Universal-Varying.

To categorize keys, we systematically analyzed the OSM wiki page for each key, keeping in mind the context of each business it was applied to. Each key was placed into a single category (Universal, Universal-Varying, or Contingent) based on its role for the business. To ensure reliability of this qualitative process, I and a collaborator classified the keys independently and then resolved disagreements. See Table 4.2 for the results[6].

4.7 Results

4.7.1 Measuring Standardization

We next systematically compared tag data in each of our clusters to corresponding pages in the OSM wiki, thus assessing standardization. I and a collaborator carried out the coding procedures for this process. The procedure varied by metadata type.

4.7.1.1 Universal Metadata.

For each tag in each cluster, the coders analyzed corresponding wiki key and tag page descriptions. The coders performed this process to consider the tags' appropriateness for the business instances they were

[6] 5 keys were removed because both coders agreed they were not relevant (3 were not in the wiki, 1 was not business related, the final key "ref:arbys", was removed since it was for Arby's restaurants only).

applied to. For example, the wiki indicated that the tag "amenity=fast_food" would be appropriate for McDonald's cluster instances but not for Safeway cluster instances.

Although 1592 distinct key-values for Universal keys were applied in our dataset, we narrowed our focus by selecting the 10 most common values for each key for our coding process[7]. Remaining values were considered applications that did not align with wiki instructions. We believe selecting the 10 most common values was reasonable, since this included all key-values that appeared more than once within a business (with two exceptions: one 11th-most-popular value was applied twice, the other was applied thrice)[8]. This coding process identified 133 business-Universal key pairs with at least one appropriate (according to the wiki) value. We used metadata associated with these pairs for Universal metadata standardization analyses.

This process showed that some applied metadata did not align with the wiki. For example, "shop=supermarket" was applied to 8 instances of the pharmacy CVS. The wiki states that "shop=supermarket" is for "a full service grocery store" [141]. Given this, and given the coders' knowledge of CVS locations in the United States, it was clear that this tag was not appropriate for CVS instances. We classified such applications as **misaligned** since they did not align with wiki instructions. We consider *misaligned* metadata to be *unstandardized* metadata.

Many other tag applications were in alignment with the wiki instructions. For example, we observed both "amenity=fast_food" and "amenity=cafe" applied to different Panera Bread restaurants. Careful reading of the wiki suggested that both tags were appropriate. The wiki page for "amenity=fast_food" says that this tag should be used "for a place concentrating on very fast counter-only service and take-away food." and "They usually, but not always, have sit-down facilities ranging from two or three to many easy-to-clean chairs and tables." The wiki page for "amenity=cafe" describes a café as "a generally informal place with sit-down facilities selling beverages and light meals and/or snacks." Both tags provide accurate and useful descriptive information about Panera Bread instances and were applied consistently with the wiki instructions. However, due to OSM's one-key-one-value data model, only one of the values could be applied to a given Panera Bread instance. Hence, applications of either of these tags were considered **ambiguous**. More generally, whenever at least two distinct instances of the same business had different values for the same key and each value aligned with the wiki instructions, we considered those tag applications to be **ambiguous**. We consider *ambiguous* metadata to be *unstandardized* metadata.

4.7.1.2 Location-Specific Metadata: Universal-Varying and Contingent.

We found that very few **Universal-Varying keys** actually were applied to appropriate business instances. For example, "opening_hours" was Universal-Varying for Walgreens and other businesses, and thus was

[7] We coded *all* tags for website-related keys. Further details of website analyses are discussed in Detailed Results.

[8] Regardless, most data aligned with the wiki anyway.

appropriate for all of them. However, only 3% of Walgreens had this key applied, and this trend was consistent for other businesses, too. A similar scenario played out for phone number metadata. Across all Universal-Varying metadata, 88% of *potential* metadata was unapplied. Note that mapping applications such as Google Maps provide this data when it is available, indicating there is user demand for this location-specific content. OSM severely lacks this type of metadata, limiting its utility as a data providing source.

A likely reason for the lack of Universal-Varying metadata is that applying it requires more work from contributors than business-wide (Universal) metadata does; contributors have to look up information for each individual business location. This extra work may be more than contributors are willing to do; our prior research in Study 1 has shown that contributors limit their effort when tagging individual records.

Contingent metadata was even more rarely applied than Universal-Varying metadata. 94% of *potential* metadata was unapplied. Determining this was more complex than for Universal-Varying metadata since Contingent metadata only applies to some instances of a given business. (Although metadata is sometimes applied to indicate the lack of an attribute's existence, this was not common in our dataset.) Hence, the effort of determining if an attribute represented with Contingent metadata *is present* is possibly a reason why even less was applied. Given our need to look up location-specific information about Contingent metadata, we sampled an important and representative subset. For more details of this sampling process and of our rationale, see Appendix 9.1.

Given that Universal-Varying and Contingent metadata was so rarely applied when it was appropriate, we focused our remaining analysis on Universal metadata – 38,220 Universal business-key-values. We return to Universal-Varying and Contingent metadata when discussing important opportunities for the community to improve the number of tag applications.

4.7.2 Detailed Results

4.7.2.1 Universal Metadata Standardization.

Recall that Universal metadata were key-values (tags) that were universal to instances of a given business. Figure 1 illustrates the results for Universal metadata standardization. There were 38,220 applied Universal business-key-value triples. Only 3706 business-key-value triples did *not* align with wiki instructions. Thus, 90% of applied metadata aligned with wiki instructions.

However, out of the remaining 34,514 aligned triples, 76 of 133 Universal key-business pairs were ambiguous, leading to 18,841 ambiguous triples (49% of all triples). Thus, while most tag applications complied with the wiki, a significant amount of applied metadata was ambiguous. The result was that 15,673 triples were aligned and not ambiguous: that is, *only 41% of metadata did not have standardization issues*.

4.7.2.2 Universal Standardization Failures through Different Lenses.

We found that standardization of keys varied quite a bit, with a common pattern: keys whose OSM specifications are less clear are more likely to be misaligned. We discuss details next. We also observed that

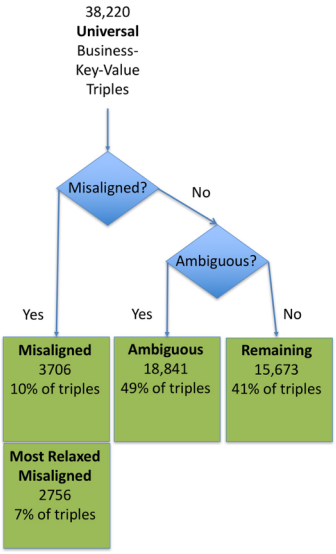

Figure 4.1: Universal Business-Key-Value Triples

standardization of businesses depends largely on the keys applied to them; if keys are problematic, the businesses will be, too. Thus, analyzing standardization by businesses provided little new insight, so we do not discuss that dimension further.

Misalignment by key. Business website information (represented in our dataset by the keys "contact:website" and "website") is prevalent in various mapping applications including those from Google, Bing, and Yelp – indicating its demand. Our initial analysis found that 63% of the website data was misaligned. As we coded the wiki, however, we observed that the wiki specifications for how to enter URLs were hard to interpret and contained very specific formatting instructions. Hence, many URLs were close to being aligned, with only small syntax problems, and most URL variations were infrequently applied. Although many URLs did not align precisely with the wiki, web browsers and other applications can handle a range of variations and still retrieve the appropriate web page. We applied some simple normalizations to the URLs in our dataset, which reduced the misalignment. We then took a further step: checking whether URLs navigated to a working website. This was the case for 93% of URLs in our dataset. Of the remaining 7%, most generated an HTTP 403 or 404 error, likely indicating that these URLs were not kept up-to-date as of the time we checked (April 2017). Figure 1 includes this most relaxed version of misalignment for website metadata. To sum up, from a strict syntactic perspective, most website data was misaligned, but from a practical perspective, nearly all website metadata was in fact *aligned*.

We also further examined non-website-related misaligned metadata to check if these tags were simply typos (i.e. slight and obvious misspellings of *aligned* metadata). 98% of misaligned metadata were not a result of

typos; *instead, the errors were due to substantive misalignments with wiki instructions. These tags were intentionally applied when the wiki does not indicate that they should be*.

Ambiguity by key. Four Universal keys were sometimes ambiguous: "amenity", "cuisine", "shop", and "website". Table 4.3 provides details about the businesses each key was ambiguous for. "amenity" was ambiguous for 9 of 28 businesses it was applied to. "cuisine" was ambiguous for almost all – 21/22 – the businesses it was applied to. "website" was ambiguous for 35 out of 41 businesses it was applied to.

There are a couple of interesting implications from these results. First, the one-key-one-value data model restriction appears to be particularly incompatible with attributes such as restaurant cuisine. Essentially all restaurants had multiple types of cuisine, *but a given instance could only have one of the types specified*. Thus, applications using the data may be unaware that a given restaurant has multiple types of cuisine. Second, it is possible that some ambiguity (especially in the case of website metadata) is due to the hard-to-interpret instructions in the wiki that were discussed previously. Thus, clarifying wiki instructions is important if the community would like to improve metadata consistency. Performing work focused on understanding the degree to which the OSM wiki is universally understood (and informed by sociolinguistic theory on achieving common ground) is an important first step in this direction. Improving consistency would also allow the various applications using OSM data to more easily process data.

4.7.2.3 Missed Opportunities to Apply Metadata.

In addition to studying standardization, we noticed significant missed opportunities to apply useful metadata. We discuss these next.

Universal Metadata. If all appropriate Universal metadata was applied to the businesses in our dataset (e.g., if all Dairy Queens had cuisine information or all Wal-Marts had website information), the amount of applied Universal metadata would increase from 38,220 to 95,926 Universal business-key-values, or by over 250%.

The amount of missed opportunities for Universal metadata applications varied widely across keys: mean = 76%, s.d. = 33%. Only the "amenity" key was applied with great consistency; just 4% of business instances where the wiki deemed this metadata appropriate (e.g. "amenity=fast_food" for McDonald's) did not have it. There are several possible reasons why "amenity" is applied consistently: 1) "amenity" is the "primary" point of interest (e.g. chain business) key according to Over et al. [74], 2) applications such as OsmAnd[9] appear to use "amenity=" tags to render icons, and 3) our method of filtering records (discussed in the Clustering Algorithm section of Methods) may have favored records with this key. Specifically, records chosen for further analysis either had "amenity=", "shop=", or "cuisine=" applied to them, or had the same "name=" tag as a record that did. We did this to remove the large amount of irrelevant records from our sample.

[9] http://osmand.net

Table 4.3: Business-Universal Pairs with 2 or More Aligned Values

amenity	Bank of America	Dairy Queen	Denny's	Dunkin' Donuts	IHOP	McDonald's
	Panera Bread	Starbucks	Wells Fargo			
cuisine	Applebee's	Arby's	Burger King	Chick-fil-A	Culver's	Dairy Queen
	Denny's	Dunkin' Donuts	IHOP	Jack in the Box	KFC	McDonald's
	Olive Garden	Panda Express	Panera Bread	Pizza Hut	Sonic	Starbucks
	Subway	Taco Bell	Wendy's			
shop	Chevron	CVS	Dairy Queen	Dollar Tree	Home Depot	Radio Shack
	Rite Aid	Sam's Club	Staples	Walgreens	Wal-Mart	
website	7-Eleven	Applebee's	Arby's	AutoZone	Bank of America	Barnes & Noble
	Burger King	Chick-fil-A	Circle-K	CVS	Dairy Queen	Denny's
	Dollar Tree	Dunkin' Donuts	Home Depot	IHOP	Jack in the Box	McDonald's
	Olive Garden	Panda Express	Panera Bread	Pizza Hut	Radio Shack	Rite Aid
	Safeway	Sam's Club	Staples	Starbucks	Subway	Taco Bell
	Walgreens	Wal-Mart	Wells Fargo	Wendy's	Whataburger	

As our research has shown in Study 1, OSM contributors have said that they just "basically" characterized objects with the "minimum" information; "it's too much work to add everything". Our results align with this observation: "amenity" is precisely the type of "minimum" "basic" information likely to be provided for an entity. The other Universal keys all had substantial missed opportunities to apply metadata; for example, 94% of *potential* "website" key applications did not exist. In the Discussion section, we consider ways to improve metadata application while still respecting OSM contributor values and attitudes.

Universal-Varying Metadata. Recall that Universal-Varying metadata were keys that were universal to instances of a given business, along with values that were location-specific. If all Universal-Varying metadata was applied to every instance of the respective businesses they belonged to (e.g. if all Olive Garden restaurants or Walgreens had operating hours information), the amount of applied Universal-Varying metadata would increase from 9,319 to 75,591 business-key-value, an increase of over 810%.

There was less variation in metadata application between different Universal-Varying keys compared to different Universal keys: nearly all appropriate Universal-Varying metadata was left unapplied. For example, two Universal-Varying keys – for operating hours ("opening_hours") and for phone number ("phone") – were applied to fewer than 5% of businesses they could be applied to. The information this metadata provides is very useful for potential customers, as evidenced by its use when available in applications such as Google and Bing Maps and Yelp. The absence of this data reduces the usefulness of mapping applications that use this information.

It makes sense that less Universal-Varying data would be applied than Universal data: determining the proper value for a Universal-Varying key for a given instance is a non-trivial task since location-specific information is needed. To illustrate the effort required, determining the opening hours for a specific Starbucks location requires looking up the information on the web. Obtaining other location-specific information may even require physically visiting the actual location (which is not part of OSM's common remote "armchair mapping"[10]).

Contingent Metadata. Recall that Contingent metadata are key-values describing attributes that are *not* universal to instances of a given business. As mentioned previously, we sampled important and representative Contingent metadata. Specifically, we considered the "internet_access" key for McDonald's and Starbucks and the "drive_through" key for McDonald's, Starbucks, and Walgreens. Google Maps and Yelp also use these types of information when available – again, indicating this data is important and in demand. Based on the results of our sampling process, if all Contingent metadata was applied to every instance of the respective businesses they belonged to (e.g., if all McDonald's containing a drive-through or all Starbucks with internet access had corresponding metadata applied), the amount of applied Contingent metadata in our dataset would increase from 14 to 245 Contingent business-key-values, an increase of 1750%.

Each business-key sampled had missed opportunities to apply metadata at least 90% of the time that it was appropriate. As mentioned previously and as discussed in Appendix 9.1, our samples were likely among the most applied Contingent metadata. The key "internet_access" for McDonald's locations had the largest amount of missed opportunities, missing them 98% of the time. All but one of the McDonald's sampled had internet access in reality – but there was no metadata to show for it.

4.8 Discussion

We summarize our core findings here. We found:

- The OSM community does a good job of applying data that is aligned with the wiki instructions.

[10] https://wiki.openstreetmap.org/wiki/Armchair_mapping

- The one-key-one-value OSM data model restriction results in a very significant amount of ambiguous applied metadata.
- A significant number of opportunities to apply metadata are missed.

Based on these findings, we next provide implications for OSM and peer production more generally.

Increasing precision in data standards. Related to ambiguity and misalignment**,** the discrepancy in the level of "structure" between the OSM standard (wiki) and the data itself leaves room for interpretation. Similar to the observations of prior work [1,77], sometimes wiki descriptions are quite general and hard to interpret (as was seen for website metadata). This issue may be due to an effort to make the definitions globally applicable and relatable across languages; after all, contributors try to create one global tagging standard. However, this leaves room for contributors to tag the same thing in different ways. It may make sense for the community to consider alternative definitions and descriptions for metadata. Increased use of tagging examples (e.g. of business-specific examples) could leave less room for interpretation. Further, more pictures in wiki descriptions, as suggested in Study 1, may mitigate this problem. Here, it's worth stating that since Wikidata is also a data repository with a single language-independent version, it may have similar problems and could potentially benefit from similar solutions.

OSM data model. While the above data standard changes might help reduce metadata misalignment/ambiguity, ambiguity stemming from the data model is still problematic. As we noted previously in Study 1, the data model could change to account for entities in the real-world that have multiple values for different attributes. A Dairy Queen specializes in both burgers and ice cream for its cuisine, and a contributor should not need to choose just one. This data model change would improve end-user experience since applications would have access to all information on OSM entities. This proposed data model has been shown to work in similar peer production communities: both Wikipedia and Wikidata have avoided OSM's data model issue by opting for multiple values per key in their structured data.

Metadata that is harder to apply or requires frequent maintenance is less likely to be applied. Based on our data and analyses of missed opportunities to apply metadata, 60% of *potential* Universal metadata, 88% of *potential* Universal-Varying metadata, and 94% of sampled *potential* Contingent metadata was not applied. This suggests that as metadata becomes more variable -- and thus, requires more work to apply – it will be applied less. Additionally, location-specific metadata requires more frequent updates than Universal metadata. For example, it is more likely that one specific Subway shop's operating hours will change than it is for the cuisine of all Subways to change. Likewise, it is more likely that a Walgreen's location will add or remove a drive-through than it is for all Walgreen's to become something other than a pharmacy. Indeed, discussion with OSM contributors has indicated that the need to maintain metadata is a deterrent from applying it in the first place. Enabled by the core community value of contributor freedom, OSM contributors

limit their tagging effort (based on our findings in Study 1), and this shows in the form of lower percentages of potential location-specific (Universal-Varying and Contingent) metadata being applied.

Given these considerations, it might be worth pursuing new ways to use automation for tagging. A possible option that was also discussed in Study 1 would be to integrate data entry tools with businesses' databases. While prior research (e.g. [16,44,99]) has shown the negative effects data imports and remote or non-local work can have on data quality, businesses are naturally incentivized to input and maintain accurate and detailed metadata for their locations. Of course, creating the code to facilitate business data imports would put an initial added burden on OSM contributors; however, we believe that in the end, this approach would ease the burden of getting business metadata into OSM.

Interestingly, the idea of businesses updating their own data has been considered in other peer production contexts, including Wikipedia and Wikidata. In fact, Wikipedia has a "conflict of interest"[11] policy preventing businesses from doing this. However, this is not true for Wikidata, a community that, like OSM, focuses on producing structured data instead of prose. Discussion in Wikidata around creating a similar conflict of interest policy to Wikipedia indicated a feeling that "since Wikidata does not allow for natural language, a lot of nuance and opportunity for bias goes away." [142] Given the similarities between Wikidata and OSM and the views that the Wikidata community has, it might be reasonable for OSM to follow suit and allow businesses to update their own metadata.

4.9 Limitations and Future Work

Our analysis of OSM wiki pages focused on the key and tag pages corresponding to metadata applied to instances of our chain business clusters. We note that eight of the businesses in our study currently have their own OSM wiki pages with business-specific tagging instructions. However, these pages are not widely adapted across businesses. Second, none of these pages actually existed when we obtained our dataset – hence, contributors could not follow them. Third and most importantly, these business pages provide specific tagging instructions, typically providing only one appropriate value for a given key. Given the one-key-one-value data model, this makes sense to do. However, the advantage to our approach is that it helps provide an understanding of the extent to which this data model constraint results in incomplete representations of the attributes of businesses. Our approach also more conservatively estimates misalignment, only considering a tag application to be misaligned if it is not helping to describe the entity that it is applied to (based on information regarding that tag in the wiki). Hence, we gave the benefit of the doubt to contributors when calculating misalignment.

As mentioned, the OSM wiki represents an ever-changing and expanding community tagging standard. As new keys and tags are added, opportunities to apply metadata increase accordingly. Because of this evolution,

[11] https://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest

it's important to note that not all the identified missed tagging opportunities were necessarily considered missed opportunities at earlier points in OSM's history. For instance, a contributor might have applied much of the relevant metadata to a McDonald's record when mapping it in 2010. However, by 2015, many opportunities might be missed for that record if additional relevant tags were introduced in the wiki but were not applied to the record. Future work should explore how the level of missed tagging opportunities has changed as OSM has matured.

"Coverage", or the degree to which OSM provides data describing the real-world, is a commonly used lens for considering OSM data quality. While both our work and prior work has considered missed coverage opportunities, prior work (e.g. [26,66,100]) measured such missed opportunities by considering how often objects (restaurants, highways, etc.) from the real world are represented by objects in OpenStreetMap. We, however, quantified coverage by instead considering missed opportunities to apply metadata for objects that do exist. Some work in OSM has provided evidence that coverage biases exist along dimensions such as population density [66]. Future OpenStreetMap work might consider whether similar biases occur with our definition of coverage, particularly given the substantial impact of these chain businesses in low-SES and rural areas as noted above.

5 Bot Detection in Wikidata Using Behavioral and Other Informal Cues

5.1 Introduction

> *The week-long period from April 22 to April 29, 2016 seems to represent an historic peak of human productivity. At around noon GMT on April 22, someone began making edits to Wikidata, a structured data peer production community and sister project to Wikipedia. Rather than choosing to have a relaxing Friday afternoon and evening, this dedicated contributor got caught up in the editing spirit and pushed on – editing late into the night. They performed their edits with incredible consistency, on average making an edit every 30 seconds. Amazingly, this contributor did not stop at all that night – nor the next, nor the next.... In fact, they continued working for 7 straight days, finally deciding to take a break over 21,000 edits later. What's more, they made all their edits anonymously, without even logging in to a Wikidata contributor account.*

It's unlikely that a single human could contribute at this pace and scale. Indeed, this editing behavior looks suspiciously like a bot. As noted previously, bot editing is particularly widely used in Wikidata (88% of all edits [86]). This is in part due to the relative ease with which its structured (key-value pair) content can be imported from external data sources. Community rules require bot operations to occur via Wikidata accounts approved to run with a "bot flag". But obviously, the prolific anonymous contributor above was not using a bot flagged account – they were anonymous! Was this contributor really superhuman? Or was someone operating a bot against community policy[12]? Highly accurate bot detection in peer production is important for multiple reasons. We discuss these next.

First, inaccuracies in bot identification could invalidate analyses of peer production that attempt to distinguish the role of bot and human edits. The second research challenge identified in this thesis sought to compare the value that human and automated contributions provide and we wanted to be confident that we were accurately distinguishing between the two. If studies of human contributors inadvertently use datasets containing a large amount of bot activity, results supposedly about human behavior could instead largely reflect bot behavior. Since bot edit rates often are much higher than human edit rates, just a few bot activity sessions mistakenly labeled as human activity could dramatically skew the results of some analyses. For example, related to prior peer production work, if the dataset used for the seminal conclusion that "Wikipedians are born, not made" [76] contained bot edits (something that the authors noted explicitly that they wanted to avoid), it could turn out that bots are "born" and human contributors are "made". Further, for work that has studied bot

[12] https://www.wikidata.org/wiki/Wikidata:Bots

contributions (e.g., [21,78,92]), it is important to include as many bot contributions as possible to provide a complete picture of the role bots play in the community.

Highly accurate bot detection is also important since the Wikidata community is concerned with detecting *all* bot activities. Even just one undetected bot that is producing thousands of edits has the potential to do large amounts of damage to the community. To avoid damaging scenarios, mandatory bot approval policies have been put in place in both Wikidata and Wikipedia [105,143].

Concerns about unidentified bot activity are not merely theoretical. While not a topic of extensive analysis, previous work in Wikipedia and OpenStreetMap has identified behavioral anomalies as resulting from unidentified bots [31]. Further, I have personally come across unidentified bots during data explorations[13]. Thus, while unidentified bots occur with certainty, the prevalence of such edits remains to be seen.

To facilitate both peer production research and Wikidata community policy enforcement, our work offers the following contributions.

5.1.1 Contributions

- We describe an effective bot detection strategy using machine classification and implicit behavioral and other informal editing characteristics and show that it is effective in identifying unflagged bot activity in Wikidata.
- Using our model, we show that, for the most part, unflagged bot activities are rare. Hence, *the activities of unflagged bots appear unlikely to significantly change results in many studies of human and bot behavior*. However, analyses that are sensitive to outliers (e.g. max session duration) or that span relatively short subsets of Wikidata history should first extract previously unidentified bots from "human" contributions to ensure accurate analyses.
- We show that there is a meaningful amount of non-compliance with the *bot policy* in Wikidata. According to our model, 3% of registered "human" user edits overall and 2% of anonymous "human" user edits overall are from bots. These percentages are important since all bot activity that does not align with community policy matters to the Wikidata community.
- We make our datasets and bot detection code available under an open license: https://github.com/hall1467/wikidata_bot_prediction_model. This will facilitate future Wikidata contributor behavior research as well as better allow the Wikidata community to enforce its bot policies.

[13] For example, we have identified anonymous edits coming from the Wikimedia internal network (Wikimedia operates Wikipedia). The IP addresses associated with these edits had a prefix of "10.68." Such internal servers are not equipped to support a GUI/browser and thus the edits almost certainly came from bots.

5.2 Background and Related Work

5.2.1 Bot Detection

Current bot detection methods used in peer production rely on explicit signals (e.g. *flagged* bot accounts). Explicit signaling is common in communities such as Wikidata and Wikipedia since these communities are highly regulated and so tend to have relevant policies. Despite a long history of bot research, the methods used in many influential Wikipedia studies (e.g. [19,32,50,76]) are not always straightforward to apply and still miss some bots. Geiger and Halfaker [21] noted that "...getting a list of bot accounts has long been a challenge for Wikipedia researchers." As mentioned, Wikipedia and Wikidata policies require users to gain approval to run bots. This occurs through a separate "bot flagged" account. The communities maintain tables of both current and former approved bot accounts. Geiger and Halfaker state that a widely-used method of identifying bots in Wikipedia is to consider user accounts that are *currently* flagged as bots (some accounts will be de-flagged when they are inactive). One other technique mentioned is to find usernames with "bot" in them. These techniques have been applied in Wikidata research [72,78,86] and may have missed bots/bot edits for several specific reasons. First, contributors may forget to switch to their "bot account" to run bots, or even forget to log in at all, instead editing anonymously. Second, if a user wishes to avoid going through the bot approval process, they may secretly run their bot anyway, as an anonymous contributor or through their personal account. Third, Wikidata did not have an effective bot policy for the first 14 months of its existence[14]. And that policy undoubtedly was not consistently followed by all bot maintainers immediately after it was instituted.

In some contexts outside peer production – such as malware detection, video games, Twitter, etc. – it is not possible to recognize bots using explicit techniques. These contexts are often less regulated, and bots deployed in them will not identify themselves since their goal is precisely to avoid detection. Thus, implicit bot detection signals must be used in these contexts. Fortunately, bot behavior patterns often are quite distinct from human patterns, and bot detection techniques (e.g. [45,46,89,91]) can take advantage of implicit behavioral signals such as repetition.

5.2.2 Activity Session Behavioral Patterns

Because of the limitations of peer production bot recognition approaches using explicit signals/labels, we took insight from bot recognition techniques developed in contexts where only implicit behavioral signals are available. Similar behavioral patterns to those identified in the previous section have occurred in peer production. Specifically, implicit characteristics of human edit "sessions" have been shown to follow predictable patterns [19,31]. In the context of such work, an edit session is a contiguous series of edits performed by a user without a substantial break. This work showed that people follow approximately

[14] https://www.wikidata.org/w/index.php?title=Wikidata:Bots&diff=96055780&oldid=66572890

consistent distributions of inter-edit and between-session times, and that bot activity deviates from these human patterns.

Analyzing user contributions at the session-level makes sense for two additional reasons as well. First, some registered users will make most of their edits manually but occasionally decide to run a bot though their own account for a short time. Detecting bots at the account level would not detect this behavior since (presumably) most of such users' activity would follow human-like patterns. Second, bot detection at the account-level cannot be done for *anonymous* user contributions. Anonymous contributions are also known as "logged-out" or "*IP address*" contributions, since only the IP address is logged rather than a username/user id. IP addresses are often dynamically allocated. Hence, an IP address could correspond with a certain user on one day and a different user on the next. While anonymous contributions compose a small percentage of contributions in peer production [86] -- less than 1% in our Wikidata dataset -- work has indicated that anonymous contributions in Wikipedia are disproportionately valuable to the community [124,125]. Thus, important questions remain to be answered in Wikidata regarding the roles of anonymous contributors and the value they add.

5.3 Methods

We next discuss the steps we followed to build our bot prediction model. In-depth details of the generation of training, testing, and other necessary datasets used can be found in Appendix 9.2. Further, while not required to understand the discussions in this chapter, the reader may wish to review information on Wikidata terminology at [144].

We built our model using scikit-learn[15]. We tried two different models: a random forest classifier and a gradient boosting classifier. We expected these ensemble models to be effective at making predictions since they have proved effective in similar contexts such as Wikipedia vandalism detection [82]. We applied hyper-parameter optimization for both models, optimizing initially for ROC-AUC and PR-AUC. Three-fold cross validation was used, which is scikit-learn's default. ROC-AUC has been a standard machine learning metric [79]. However, our training dataset was heavily skewed towards human sessions. In such cases, PR-AUC, a metric commonly used in information retrieval when skewness is a concern, is preferred [79]. We chose the gradient boosting classifier because it had a higher PR-AUC on test data (0.528 versus 0.486).

All features used in our model were based on "informal" characteristics of edits and editing *sessions*. Such "informal" characteristics were not based on the formal, explicit signals (e.g. bot flags) used in prior work to identify bots. Details of the necessary step of "session-izing" our data can be found in Appendix 9.2. As will be discussed, we iterated twice in our model feature selection.

[15] http://scikit-learn.org/stable/

We first offer insights into our initial choice of model features. As stated, previous work showed that bots and humans have different inter-edit time patterns in peer production [31]. Further, bots have been shown to have longer periods of activity in (for example) online games [45]. We used these two implicit behavioral characteristics in our model and also considered the standard deviation of the time difference between edits – a similar behavioral signal as used in web robot detection [89]. We also conjectured that bots would focus mostly on actual *items* while people would likely also edit elsewhere in the community, for example performing policy development or engaging in other discussion. Such distinct editing activities occur in distinct "namespaces" and we derived features to measure the number of edits to these different areas of the community. Essentially, we felt bot and human edit sessions would have different distributions of edits across namespaces. Appendix 9.3 provides details about our initial features under Activity Pattern Features.

After developing and formally testing our initial model's fitness, we looked for ways to improve it. We applied the model to random samples from our anonymous user contribution sessions dataset (described in detail in Appendix 9.2) and spot-checked sessions classified as "bot" and "non-bot". This helped us identify more features. One feature was a variant of one of our initial features, designed to better detect bots based on inter-edit time, specifically by counting the number of edits per session with an inter-edit time less than 2 seconds. The remaining features took advantage of the content in *revision comments* which we describe next.

Revision comments provide information about the nature of a Wikidata *revision* (used interchangeably in this chapter with the word *edit*). Revision comments are composed of both a structured and an unstructured (free-form) component. Consider the comment "/* wbcreateclaim-create:2| */ [[Property:P107]], [[Q618123]]". The structured part is between "/*" and "*/" (e.g. "/* wbcreateclaim-create:2| */") and is automatically generated by Wikidata's software when a revision occurs. In this case, the structured part indicates that a claim was created. The rest of the comment (e.g. " [[Property:P107]], [[Q618123]]") contains information that a contributor is free to modify and replace. Bots sometimes provide additional information in this unstructured space about the nature of the revision. In our example, default information was left in place indicating the claim modified was mapping property 107 to item 618123. An example of a bot leaving its own information is "/* wbsetlabel-set:1|de */ Bot: change label for de after page move: Stena Nordica -> Malo Seaways"[16].

Most of our comment-based features distinguished different types of editing behavior by examining structured content. Among other things, such features measured the number of edits to different aspects of an item, for example, to descriptions, aliases, and labels. One feature based on the *unstructured* part of comments identified bots simply by looking for the word "bot" in the user-provided unstructured comment text. Similarly, another feature also relied on a (pseudo-)explicit signal, namely the presence of a "generic" boilerplate comment containing only a structured part of a certain syntax, for example: "/* wbeditentity-

[16] https://www.wikidata.org/w/index.php?title=Q3355946&oldid=209807505

Table 5.1: Bot Prediction Model Fitness on Registered User Contributions

Features	PR-AUC	ROC-AUC
Initial	0.528	0.888
Initial + Iteration 2	0.845	0.985

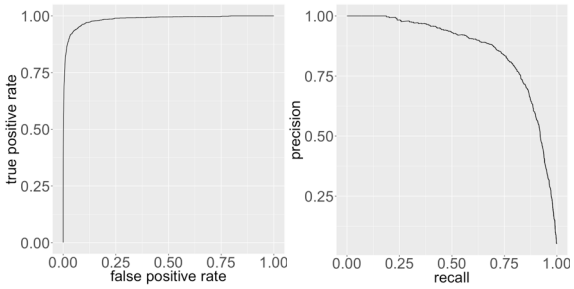

Figure 5.1: Precision-Recall and Receiver Operating Characteristic Curves

update:0| */". This particular comment syntax only comes from API edits carried out by bots. Although the latter two features looked for specific and explicit features, they still differed from bot *flags*. Specifically, these two features are based on informal usage rather than mandated community policies and have not been considered in any previous bot detection methods. More details on the second iteration of model features are listed in Appendix 9.3 under Revision Comment-Based Features.

Finally, we re-tuned the gradient boosting model using the additional features and a new distinct test set of registered user contributions to avoid overfitting. Using a grid-search based hyper-parameter optimization, we optimized for PR-AUC by adjusting the number of estimators, max depth, and learning rate. The resulting parameters were: number of estimators = 1100, max depth = 3, and learning rate = 0.1.

5.4 Results

5.4.1 Formal Model Evaluation on Registered User Edits

Table 5.1 provides fitness information for our gradient boosting bot prediction model, both before and after we added the "Revision Comment-Based" features. These statistics result from applying the models to their respective registered user testing datasets: 1) the dataset generated for testing the initial "Activity Pattern" features and 2) the dataset generated for testing the "Activity Pattern" plus "Revision Comment-Based" ("iteration 2") features. As can be seen, revision comment-based features considerably improved model performance. Figure 5.1 provides PR-AUC and ROC-AUC graphs from the final model (which contained both feature sets). Using testing data, with default parameters, this model had a precision of 0.88 and recall of 0.69.

5.4.2 Qualitative Model Evaluation on Anonymous User Edits

As stated earlier, anonymous edits are understudied and potentially quite valuable to peer production [124,125]. Hence, to better understand such edits, we wanted to ensure our model was effective at detecting anonymous bot contributions. When building our model, we hypothesized that the explicitly flagged bots used for model training would have behavioral characteristics similar enough to those of anonymous bots

and thus the model trained on one set would be effective on the other. However, we wanted to be sure that this was the case. In this section, we describe how we went about ensuring our model performed well on anonymous contributions.

For unregistered ("anonymous") user contributions, we did not have labeled testing data so we needed to evaluate model fitness in a different way. We first applied our model to our anonymous contributions dataset (described in detail in Appendix 9.2) and generated likelihood estimates for each session. These likelihood estimates give the probability that a given session was produced by a bot. Setting a model's confidence threshold at a given likelihood estimate will correspond to a given level of model precision and recall when applied to the test dataset. If a user wants precision in the model to be higher than 0.95, the confidence threshold in the model can be set to a corresponding likelihood estimate. The model then would return predictions with precision of 0.95, assuming it is applied to data from the same population it was tested on. However, we cannot assume that the registered user data used to train/test the model was similar to the anonymous user data we applied it to. Instead, we had to **test** whether this was true. Therefore, to get a sense for how the model performed across meaningful likelihood ranges, we sampled at likelihoods with matching test set recall ranges of 0.0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4...0.9-1.0. This strategy allowed us to check how precision (specificity) changed as we increased recall (sensitivity) when applying the model to anonymous contributions. If our model was effective, we should expect to see two results from this analysis. First, the model should find bots in the anonymous contributor pool *at roughly the same rate as among registered accounts*. Second, as one would expect in any prediction model, precision should be high when the recall associated with likelihood values is small and precision should decline as the recall associated with likelihood values increases. Eventually, precision should approach zero when all bots have been removed. If anonymous user sessions have significantly different behavioral characteristics than registered user sessions, we might find that precision stays roughly the same as recall increases. This would mean the model is ineffective at telling bot and human sessions apart.

To determine whether our model was successful for anonymous user contributions in these two regards, we needed to generate ground truth data. Specifically, we needed to know what sessions were from bots and what sessions were from humans at likelihood/recall strata. We refer to these strata as "recall strata" from this point forward for ease of discussion. We generated 20 random session samples per recall strata[17]. We next developed a code book, seen in Table 5.2, of which the purpose of each code was to identify *bot-like session characteristics.* Additionally, a final code in our codebook explicitly indicated overall coder judgement as to whether a session was from a bot or human. I and a collaborator had developed this code book. I then systematically went through each sampled edit session within each strata -- examining revision comments, timestamps, namespaces, and page titles. He analyzed enough revisions per session to feel

[17] In one case, there were only 18 sessions in the stratum, in which case we analyzed them all.

confident with codes he applied. Depending upon the size of the session, this involved looking at as many as 50-100 or more revisions per session.

5.4.2.1 Summary of Coding Results

Appendix 9.4 provides a detailed breakdown of coding results, including a table of results broken down by strata. As stated, the purpose of this coding was to check whether our model identifies bots in the anonymous contribution pool at roughly the same rate as for registered users and that it does so with high precision when recall is low and with declining precision as recall increases. Both characteristics occurred with our model.

Between 0.0 and ~0.3 recall, precision is very high. Between ~0.3 and ~0.7, precision drops noticeably as more human sessions appear. Finally, when recall is greater than ~0.7, most sessions are from humans.

Based on our results, we can make application-specific recommendations about how our model can be used. For applications requiring high recall bot prediction (e.g. filtering out bots for research of anonymous human contributors), we recommend optimizing a confidence threshold for recall of at least 0.7-0.8. For applications requiring high precision bot prediction (e.g. filtering out humans for research of anonymous bot contributions), we recommend optimizing a confidence threshold for recall of 0.2-0.3 or lower. Per our informal coding, when recall is between 0.0. and 0.1, precision is 1.0. When recall is between 0.1 and 0.2, precision is 0.9.

Our coding provides interesting insights into session characteristics in low recall strata where the prediction model was the most confident. Boilerplate bot comments (BBC), explicit bot comments (EBC), and fast edits (FE) codes were frequently applied when recall was below 0.4. We had derived model features to capture session characteristics represented by each of these three codes and our coding process identified the effectiveness of these features.

5.4.3 Summary of Model Evaluations

We ensured our bot prediction model worked effectively for both registered contributions (verified via a formal quantitative analysis) and anonymous contributions (verified via a qualitative thematic analysis [106]).

Table 5.2: Qualitative Analysis Codes

Code	Description
Fast edits (FE)	~10+ edits in under ~20 seconds
Consistent revision frequency (CRV)	~3+ edits occurring over ~5+ equally-spaced minutes
Boilerplate bot comment (BBC)	Comment(s) contain only a structured part and their syntax is likely indicative of bot edits coming from the Wikidata editing API. E.g. "/* wbeditentity-update:0| */"
Similar operations occur to different pages at a high frequency (SIM)	>= ~40 edits/hour, over 3+ hours
Explicit bot comment (EBC)	Comment(s) appear to explicitly (insensitive to case) indicate bot edits. E.g. "...Bot: change sitelink"
Short session with rapid revisions (SHORT)	1+ edits/second, over ~4+ seconds
Bot/human judgement	Given codes and intuition, is user a bot?

5.5 Applying the Model To Registered and Anonymous Users

Given the scenario described and the questions posed in the Introduction, is there compelling evidence to be concerned about impacts from unidentified bots? We applied our model to all registered "human" and anonymous user contributions from November 2012 to April 2017 for items and properties in Wikidata (the two main "content" namespaces) to find out[18]. Per the recommendations mentioned in the last section, we set a confidence threshold that would correspond to 0.3 recall on our test data.

We found that 3% of registered "human" contributions were predicted to have come from bots. We broke this percentage down by month and found that 8 months in Wikidata's history were predicted to have had more than 5% of such edits coming from bots (mean = 0.02, s.d. = 0.04). April 2017 experienced the largest percentage of predicted unflagged bot edits at 18%.

We found that 2% of anonymous contributions were predicted to have come from bots. We also broke this percentage down by month and found that 6 months in Wikidata's history were predicted to have had more than 5% of such edits coming from bots (mean = 0.02, s.d. = 0.04). May 2016 experienced the largest percentage of predicted unflagged bot edits at 16%.

5.5.1 Implications on Behavioral Research in Peer Production

Application of our model indicates that most human or bot behavioral studies in Wikidata do not likely need to consider the effects of unidentified bots. When studying behavior across periods of time of several months or greater, the chances of unidentified bot activity skewing results appear minimal if considering the

[18] One of our model features required edit sessions to have 3 or more edits in order to compute the standard deviation of inter-edit times within sessions. As a result, sessions smaller than this were considered human. We believe that this is justified since bots commonly edit at scale, and a primary goal when developing our model was to identify these *large* bot editing sessions. Such large sessions have the potential for a significant impact on the Wikidata community and on behavioral analyses if left unidentified. For more information on our model features, see Appendix 9.3.

behavioral characteristics of the "typical" contributor (e.g. taking means, medians, and similar outlier robust statistics).

However, for short-term behavioral studies, it's possible that anomalous spikes in undetected bot activity can affect results. For example, consider the scenario in April 2017 when 18% of "human" edits were predicted to be unflagged bots. Application of our model would be necessary to avoid false understanding of human behavior. Further, analysis of "outlier" contributor behavior (e.g. what human contributors produced the most edits in each month?) can be affected significantly by undetected bots. Consider Figure 5.2. The "All Sessions" plot represents the longest edit sessions per month in our anonymous contributor session dataset. The "Non-bot Only Sessions" plot represents the same analysis after we applied our bot prediction model to filter out bots. To ensure that nearly all bots would be removed, we used a confidence threshold that would correspond to 0.8 recall per the recommendations in the previous section.

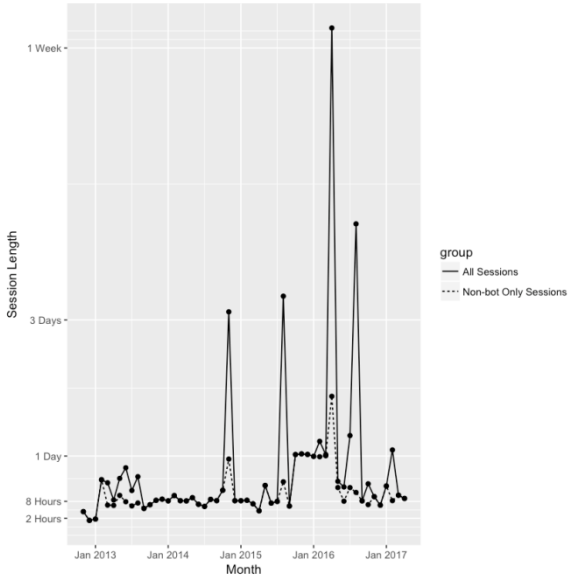

Figure 5.2: Maximum Monthly Anonymous User Session Lengths

5.5.2 Implications for the Wikidata Community and Applications using Wikidata

The 2-3% of contributions in human contribution datasets that are unidentified bots are of serious concern to the Wikidata community and applications using Wikidata. This equates to over 1 million such edits. If even a relatively small portion of these are vandalism, the effect on data quality in the community can be large. Further, such damaging edits could propagate downstream to applications that use Wikidata such as Wikipedia and Google Knowledge Graph. Hence, we hope that our prediction model in the future can help the community identify and quickly mitigate any potentially damaging behavior. Further, work [82] has used contributor type (bot or human) as a Wikidata vandalism detection model feature. Our work could help improve model performance.

5.6 Discussion

5.6.1 Model and Feature Implications

5.6.1.1 Future Studies

Our model opens the door for a number of exciting research directions and socio-technical contributor tools. For example, future important questions in Wikidata require accurate bot detection in order to be answered. Much more work is needed to understand how human, bot, and various editing tool contributions relate to content quality, and also how such contributions relate to interest and consumption patterns. Considering the initial findings in Wikidata that a higher percentage of bot edits correlates with higher quality content [78], are bots putting their effort towards content that gets used in applications like Google Knowledge Graph and Wikipedia? Further, work in Wikipedia has indicated that the supply of good content often does not align with reader interest (demand) [95]. Are bots creating "good" content in Wikidata that is of little interest or use? Additionally, how much – and what type – of work is being done by Wikidata bots that are not explicitly flagged as bots?

5.6.1.2 Feature Implications

Revision comment-based features appear to be highly effective at predicting bot edits. Our initial model did not use these features and had a PR-AUC of 0.528 and ROC-AUC of 0.888. The addition of 13 features based on characteristics of revision comments helped move PR-AUC to 0.645 and ROC-AUC to 0.985. Wikidata's structured comments denote the types of edits occurring. Capturing when these edit types occur appears to be a highly-predictive way of determining whether edits are from bots or humans. Further, based on our qualitative analysis, features derived from edit comments that are exclusive to bots (e.g. "generic" comments and comments containing the word "bot") appeared to be effective at identifying bots. For example, "generic" comments in Wikidata are indicative of bot edits coming from Wikidata's editing API. However, to ensure that other features also provided useful bot prediction signal, we removed the features *'Word ending in "bot" (case insensitive) in revision comment'* and *'# of bot generated generic comments'* and found that model performance changed minimally (PR-AUC: 0.849, ROC-AUC: 0.982). This indicates that the cumulative predictive power of other features in this set is quite useful.

Bot prediction models in other peer production communities such as OpenStreetMap and Wikipedia almost certainly would benefit from similar features to those employed in Wikidata. As shown, inter-edit time is an effective differentiator of bots and humans in Wikipedia [19,31] and could be incorporated into a future model. Additionally, Wikipedia edit tags [145] can perhaps provide explicit and implicit prediction signals similar to those identified in Wikidata revision comments. Further, although there is no "structured" component of Wikipedia revision comments to use to identify distinct editing behaviors, work [97] has been successful at "deriving that structure" with a prediction model to identify editing behaviors from unstructured Wikipedia comments. Further, features such as inter-edit time might also be relevant for predicting whether *OpenStreetMap changesets* (similar to Wikidata and Wikipedia editing sessions) are human or bot-produced.

Such changesets also contain tags which can provide signal regarding whether edits originate from bots [146]. For example, the tag "bot=yes" can be applied to explicitly indicate bot changesets [146].

5.6.2 Model Improvements

We inform future work to improve the prediction model in several ways. First, when recall is between ~0.3-~0.7 (and especially between ~0.4-~0.5), we coded a large number of sessions as "unknown" (it was unclear whether a session was from a bot or a human). From a feature engineering standpoint, close attention should be paid to characteristics of these sessions as additional, high-value bot prediction signal could be found there. Further, we uncovered an additional likely source of bot prediction signal in a comment: "/* *wbsetreference* */". In the future, this (and other similar comments, if they exist) could be incorporated into the feature that we called *"# of bot generated generic comments"*. Additionally, we noticed that bots often increment through items in order of their "item id". This fact could also be useful as future prediction signal. Further, there was a single session in the 0.8-0.9 recall stratum that was coded as a bot (see Appendix 9.4 for details). Outliers such as this may be indicative of humans running a bot in a "semi-automated" tool-like fashion. Given the lack of outliers such as this, it appears our model did well differentiating bots from semi-automated tools operated by humans. Future work may wish to differentiate human edits based on their level of automation (e.g. semi-automated tool use versus completely manual editing). Finally, the edit tags mentioned in the previous section are also used in Wikidata and could potentially be a good source of prediction signal.

6 Exploring the Effects of Manual and Automated Contributions on Content Value in Wikidata

6.1 Introduction

As we have discussed, there are several different "strategies" for contributing to Wikidata. At the most basic level, contributors can *manually* enter data via Wikidata's web interface. To edit content faster, contributors can create software either in the form of *semi-automated tools* or in the form of *fully-automated bots*. Now that we had improved confidence in the ability to distinguish bot contributions from other contributions, the final work in my thesis sought to answer a fundamental question about these different contribution strategies: *how does the value provided by each strategy differ?* For instance, do tools add more value than manual contributors? Do bots add more value than tools?

But, what precisely do we mean by *value*? Structured data value in peer production can be considered through different lenses, and we use two intuitive lenses to shape our analyses. Peer production communities including Wikipedia and Wikidata have created quality scales [109,147] to define precisely what constitutes content quality within their communities. While syntactic details of these scales vary between communities, they all include notions of completeness and well-sourced information as essential. Higher-quality content clearly provides more value to anyone using it. However, whether content is useful to a *typical consumer* also depends on the *demand* for that content. Content is of limited value to consumers if either it is of *high quality* but *low demand* or *low quality* but *high demand*. Hence, we define our first lens of content value by incorporating both item *quality* and *demand*. We base much of our exploration into this lens upon past Wikipedia research [95] which argues that content quality should *align* (*positively correlate*) completely with demand in order to provide the most consumer benefit. We refer to this lens as *consumer-level value*.

Understanding the alignment between quality and consumer demand is a good way to obtain a sense of the value that a peer production community provides to the average consumer of structured data. However, it does not provide a complete picture of the value that content provides to society more generally. The second lens through which we consider content value is *societal-level value*. A societal-level definition of content value should consider whether data have biases that can cause harmful broader effects. For example, if content about protected or minority populations is under-represented or if content about privileged populations is over-represented, then consumers of such content may form harmfully biased perceptions of the world. Prior work has indeed indicated the ability of technology to contribute to potentially harmful biased perceptions of the world through, for example, Google Image Search [48]. *Societal-level value* is *maximized* when such harmful content biases are *minimized*.

The fundamental peer production ethos of *contributor freedom* allows individuals to decide for themselves what content to produce [114,115]. Prior work studying Wikipedia and OpenStreetMap has extensively explored the effects of this ethos on the different lenses of content value that we have discussed above. For

example, work has shown that contributors will generally edit Wikipedia in a way that optimizes for *consumer-level value* [39]. Research on Wikipedia and OpenStreetMap (e.g. [26,42,95]) has also shown that contributor freedom may result in *societal-level* value problems, by underrepresenting (for example) female, rural, low-SES, and LGBTQ-related content. To date, neither of the two lenses of content value that we define has been studied extensively in Wikidata. This lack of research is concerning since *automated* editing (from bots and tools) is the dominant contribution mechanism in Wikidata [86]. And automation has great potential to shift community-wide production dynamics and affect content value at a large-scale. For example, significant harm may occur if automation were to shift production dynamics away from prioritizing societal-level value.

We *compare* the *value* that contribution strategies with differing levels of automation (bots, tools, and manual contributions) provide to peer production communities. We consider value in a robust manner by analyzing it through our two lenses: 1) consumer-level and 2) societal-level. We also investigate the underlying *reasons* for the behaviors we observe. Further, we perform our analyses *longitudinally* to see how community dynamics have changed. The results of this study are important for Wikidata and applications that use it as well as for other peer production communities as they increasingly integrate automation.

6.1.1 Contributions

This work provides several important contributions.

1. We find that manual contributors tend to work on higher-quality and higher-demand content than bots. Further, while the work performed by manual contributors tends to increase the alignment between content quality and demand, the work performed by bots has resulted in a misalignment between content quality and demand, which has increased over Wikidata's history. We also argue that complete alignment of content quality and demand may not actually be a realistic or even desirable scenario in a community like Wikidata where automated contributions dominate and contributor freedom is highly-valued.
2. We provide evidence that distinct contribution strategies play different roles in affecting the representation of minority or protected content. For example, bots tend to disproportionately focus on Global South and rural content versus Global North and urban content. However, they also disproportionately focus on male-related content versus female-related content. On the other hand, manual contributions disproportionately focus on female-related content, but also disproportionately focus on urban content. Further, we find that unregistered manual contributions can have distinct behavior from registered manual contributions. For example, unregistered manual contributions tend to disproportionately edit Global South content while registered manual contributions tend to disproportionately edit Global North content.
3. Our work provides implications for targeting contribution strategies to optimize for content value. For instance, our findings motivate initiatives and tools to help manual contribution strategies focus on high-demand content.

6.2 Related Work

We next discuss relevant prior work. The two subsections each correspond to one of the two lenses through which we consider content value.

6.2.1 Consumer-Level Value Analyses

As discussed in the introduction to this chapter, the relationship between quality and demand is important to content *consumers*. Prior peer production work (e.g. [55,73]) has considered the motivations for the type of content that people work on. As summarized by Warnacke-Wang et al. [95], past research "has found that consumer…demand is generally not a large consideration in how contributors decide to allocate their work." Hence, it is not surprising that past Wikipedia work has found a mixed relationship concerning the alignment between contributor effort and content demand (e.g. [25,39,60]). For example, while work by Hill and Shaw [39] has indicated that higher-demand content tends to be edited more, other work (e.g. [25,60]) has provided evidence that content demand does not necessarily have a positive relationship with efforts to improve content quality.

The analysis of consumer-level value in our study is largely motivated by the work of Warnacke-Wang et al. [95], which defined the *Perfect Alignment Hypothesis (PAH)*. The PAH indicates the *optimal relationship between consumer interest and content quality*. PAH alignment is maximized when content demand perfectly correlates with content quality. In other words, the PAH states that the Nth most highly-viewed content item should also be the Nth highest-quality content item. When such total alignment occurs, content provides maximum consumer value since the highest possible number of consumers benefit from the best content. The closer content is to total alignment, the more *efficient* contributors are at providing value to consumers.

Warnacke-Wang et al. tested the PAH in Wikipedia and found most article content was aligned. However, they also found that high-demand articles that were cumulatively viewed 2 billion times per month were substantially worse quality than the PAH indicates they should be. We expand upon the work by exploring the PAH in the automated contribution-dominated context of Wikidata. Further, while Warnacke-Wang et al. applied the PAH to a snapshot of Wikipedia's history, we apply the PAH longitudinally to Wikidata to see how PAH adherence has changed over time. We are also particularly interested in the ways in which different contribution strategies affect PAH alignment. When exploring how contribution strategies affect PAH alignment, we consider the two possible ways that *deviations* from alignment occur. We define terms for both. First, a *gap* indicates *lower* quality content than demand warrants. Second, a *surplus* indicates *higher* quality content than demand warrants. Coverage gaps point to content areas where the impact of future contributor editing maximizes added value, while coverage surpluses indicate editing that could have provided more value to consumers if directed elsewhere.

6.2.2 Societal-Level Value Analyses

Numerous prior studies have found problematic societal biases both in peer production communities (e.g. [26,38,42,57,87]) and other online contexts (e.g. [2,48]). For example, Kay et al. [48] found that Google

Images search results under-represent female-related content and also reinforce gender stereotypes. Work in the context of Wikipedia [57,68] and OpenStreetMap (our first study and [87]) found better representation of *male*-related content compared to *female*-related content. Further, peer-produced information quality has been shown to be associated with the socio-economic development status of a region [26,85]. Finally, work in the context of OpenStreetMap and Wikipedia [42,66,100] found that rural areas tend to have poorer quality content then urban areas.

To our knowledge, no work has studied problematic societal biases in *Wikidata* content. Given the disparities identified in other peer production communities, we considered it important to analyze biases in Wikidata along three dimensions: 1) male versus female, 2) Global North versus Global South, and 3) urban versus rural.

While prior work has shown that many problematic biases exist within peer production, work has also shown that it is possible for communities to reduce disparities once they are identified. Work by Halfaker [29] has indicated that concerted community efforts have significantly reduced gender disparities related to Wikipedia content on notable scientists. Therefore, an outcome of our explorations of problematic biases is to enable the Wikidata community to address any disparities that we find.

6.3 Methods and Data

6.3.1 Background on Wikidata Revision Data Used

Basic contribution activities in Wikidata revolve around the creation and modification of *items*. All creation and modification actions are represented as *revisions* of item *pages*. Thus, analyzing revisions yields a holistic picture of Wikidata content production. We needed Wikidata revision data to perform most of our analyses. We used "stub-meta-history" XML dump files from May 1, 2017 from the Wikimedia Foundation[19] to obtain Wikidata revision information. Wikidata's XML dumps provide a complete history of all revisions to all pages in Wikidata. These dump files provide information such as the item ids, user ids, and user names associated with each revision/edit (an "edit" is synonymous with a "revision" for the purposes of this chapter). The files also provided a revision comment for each edit, which typically provides insight about the type of edit performed. For additional information about Wikidata terminology and basic characteristics of Wikidata content, please refer to [144].

6.3.2 Determining Contribution Strategy Types of Revisions

To understand the value provided by different contribution strategies, we needed to partition our revision dataset based on strategy type. Appendix 9.5 provides a detailed description of our procedure for doing this. This process categorized a revision into one of four different contribution strategies: 1) bots, 2) tools, 3) registered manual contributions, and 4) anonymous manual contributions We decided to break down non-

[19] https://dumps.wikimedia.org

automated contributions into two groups (3 and 4) because prior peer production work [124,125] found that anonymous contributions can provide significant value to communities even though they only represent a small proportion of edits. Therefore, we thought it important to differentiate the value added by anonymous users from that added by registered users.

6.3.3 Analyzing Wikidata Contribution Strategy Behavior Longitudinally

The production dynamics of Wikidata have seen large changes over time as different automated and non-automated contributors have edited. Therefore, many of our analyses examined how contributor behavior has changed over time. To investigate this, we divided Wikidata's revision history into 4 distinct, yearlong periods:

1. Period 1: May 1, 2013 to April 30, 2014
2. Period 2: May 1, 2014 to April 30, 2015
3. Period 3: May 1, 2015 to April 30, 2016
4. Period 4: May 1, 2016 to April 30, 2017

We split the time periods between April and May because our revision data ended in May 2017 and year-long snapshots allowed us to account for yearly seasonality in interest.

6.3.4 Measuring Item Quality

The Wikidata community has developed a content quality scale [147] which measures the value of a given item based on the degree to which it includes *complete and well-referenced* information. We used this scale to obtain item quality information for our analyses. According to this scale, a Wikidata item may be assigned a quality score ranging from E (lowest quality) to A (highest quality). An E quality score means that an item does not "provide enough [structured] information to easily identify the item" and further that the item does not contain -- among other things -- a significant amount of references and multilingual metadata translations. An A quality score means that an item has -- among other things -- a very complete set of metadata, including multilingual translations and extensive references from a diverse set of external sources.

To measure item quality in accordance with this scale, we used the Objective Revision Evaluation Service (ORES) provided by the Wikimedia Foundation [129]. ORES is based on a machine learning model that provides high-accuracy[20] predictions of the quality of an item. ORES takes as input a Wikidata item's revision id from any point in Wikidata's history and predicts the quality of the item as of the time of the revision. ORES returns a "weighted sum" score giving the probability of the item belonging to each quality class. The weighted sum score formula (originally defined in prior work [29]) is computed as follows:

[20] https://ores.wikimedia.org/v3/scores/wikidatawiki/?model=itemquality&model_info

$$Weighted\ Sum = P(item\ is\ of\ quality\ class\ E) * 0 + P(item\ is\ of\ quality\ class\ D) * 1 + P(item\ is\ of\ quality\ class\ C) * 2 + P(item\ is\ of\ quality\ class\ B) * 3 + P(item\ is\ of\ quality\ class\ A) * 4$$

Given this formula, a weighted sum score of close to 0 would indicate item quality is likely of class E while a score close to 4 would indicate item quality is likely of class A.

6.3.5 Measuring Item Demand

Our consumer-level value analysis required us to quantify the consumer demand for each Wikidata item. Prior work [95] has used Wikipedia article *page views* as a proxy for a global sense of concept demand. In such a context, a "concept" corresponds with an article. We built off this approach. Given a Wikidata item, we calculated the total page views of all Wikipedia articles corresponding to the concept represented by the item. For example, for the Wikidata item pertaining to Barack Obama, we aggregated all page views across all language versions of Wikipedia that had an article pertaining to Barack Obama. This gave us a global sense of the consumer demand for information on the former U.S. president. To further improve the sense of demand for the concepts represented by Wikidata items, we also incorporated page views from any content that used the item's information on *any* Wikimedia project entities (e.g. Wikiquote entries, Wikinews entries, etc.).

Wikimedia maintains a manually curated link between Wikidata items and corresponding content on other Wikimedia projects via *sitelinks*. We used a sitelink dataset from May 1, 2017 that was available in "wbc_entity_usage" sql files from [148]. The files also indicated if any Wikimedia project pages were using the information found in Wikidata items. To obtain a measurement of the overall page view rate, we aggregated view information from Wikimedia logs[21] for a one-year period from June 2016 through June 2017. This was the most recent available data at the time. Using a yearlong period allowed us to account for seasonality[22].

6.3.6 Data Generation for Consumer-Level Value Analyses

6.3.6.1 Selecting Revisions

To understand how editing behavior related to content demand and other dynamics in Wikidata, we sometimes needed to know the quality of items (as predicted through ORES) at the time they were being edited. One challenge to obtaining item quality was the fact that it would take months to obtain ORES quality scores for all Wikidata item revisions. To get around this technical limitation, we used stratified samples in

[21] https://analytics.wikimedia.org/datasets/one-off/pageview_rate/20170607/

[22] Our work assumes that item demand is generally constant, aside from seasonal differences. Other theories of content demand dynamics do exist. For example, supply-side information economics would argue that item demand would increase as item quality increases.

our revision analyses. Specifically, we generated datasets of 100,000 random item edits of each contribution strategy (100,000 bot, 100,000 tool, 100,000 registered manual, and 100,000 anonymous manual edits) within *each* period in our dataset. For each revision in these samples, we then applied ORES and obtained the quality score for the corresponding item directly prior to the edit occurring. This was possible because our revision dataset contained the "parent id" -- the revision id of the previous edit to the same item. This approach meant that, for the analyses requiring a quality score, we could not consider revisions that created items since they would not have a parent id. However, we did not consider this an issue since the vast majority of edits do not create items and because individual edits typically provide a limited amount of quality. Hence, the edits that we did consider in such analyses would generate most of the quality found within items.

Filtering sampled revisions to obtain *content edits*. Some Wikidata item revisions do not affect the quality of item content. *We wanted to only analyze edits that were adding quality to item content*. We used revision comments to help ensure this was the case. Revision comments are composed of a structured and unstructured part. The structured part contains information automatically generated by Wikidata's software that indicates the type of edit that occurred. We filtered out edits that contained the text "sitelink", "client", or "merge" within the structured part of their comments. We refer to all remaining edits as *content edits*. According to Wikidata's quality scale [147], these edits add quality. Further, our results also tend to confirm that these edits add quality because our analyses show longitudinal quality improvements to items where content edits occur. We further believe that – adjusting for the edit-time quality of an item – the typical content edits across different contribution strategies tend to all provide the same amount of quality on average. This belief is bolstered by the fact that Wikidata's UI restricts the size of edits (unlike in Wikipedia). For additional details of the results of our revision filtering process, Table 1 in Appendix 9.6 shows the number of content edits from each contribution strategy and period after all filtering was complete.

6.3.7 Data Generation for Societal-Level Value Analyses

For each dimension of societal-level value that we considered (male versus female, Global North versus Global South, and urban versus rural), we wanted to generate as large a set of content edit samples as possible. For each dimension, we applied the same revision sampling and data processing approach that we described above, but we limited the populations that we sampled from to only relevant edits (e.g. edits to male and female content). In some cases, our populations were smaller than the desired sample size of 100,000 and we simply included the entire population in our sample. We next describe how we derived the relevant populations we sampled from.

6.3.7.1 Analyses Performed to Measure Male Versus Female Biases

As an initial exploration of gender biases in Wikidata item data, we decided to confine our analyses to the representation of content about *human males* and *human females*. Contributors can indicate that an item represents a human by applying the "instance of" [149] property with a value of "human" [150] and can indicate that an item is male or female by applying the "sex or gender" property [151] with a value of "male"

[152] or "female" [153]. We used the Wikidata Query Service [154] to identify all items matching that were human male[23] or human female[24]. For additional details, Table 2 in Appendix 9.6 shows the total number of human male and female item content edits sampled from each contribution strategy and period.

6.3.7.2 Analyses Performed to Measure Global North Versus Global South Biases

We wanted to see if geographic discrepancies exist in the representation of Wikidata item data between the Global North and Global South. Contributors can identify the geographical location of an item by applying the "coordinate location" property [155] along with a value that indicates a latitude and longitude pair. We used the Wikidata Query Service to identify such *geotagged* items[25]. Items were considered to be associated with the country they were physically closest to. We used a Python utility [90] to make this determination. We then flagged countries that were considered by the United Nations[26] to be "developed" as the Global North. All other countries were considered to be a part of the Global South. For additional details, Table 3 in Appendix 9.6 shows the total number of geotagged item content edits sampled from each contribution strategy and period.

6.3.7.3 Analyses Performed to Measure Urban versus Rural Biases

We also wanted to see if geographic discrepancies exist in the representation of Wikidata items between urban and rural areas. We considered such discrepancies at the county-level in the context of the continental United States. We first used the Shapely Python library [23] in conjunction with county-perimeter data used in prior work [43] to identify the county associated with items. We then merged this data with the county-level "Urban-Rural Classification Scheme" provided by the National Center for Health Statistics (NCHS) [156]. The above scheme is based on a 1-6 scale where 1 is the most urban counties and 6 is the most rural. We considered counties classified as 1-4 to be urban and counties classified as 5-6 to be rural. These distinctions were made based on what counties the NCHS consider to be "metropolitan" versus "nonmetropolitan" [156]. If the reader is interested, Table 4 in Appendix 9.6 shows the resulting number of urban and rural item content edits from each contribution strategy and period.

6.3.8 Additional Details of Our Analytic Methods

6.3.8.1 Generating and Using Item Expected Quality Distributions

Many of our analyses required us to compare the "expected quality" of items within content edits to their actual quality to see how well the PAH had been adhered to. According to the PAH, an item in a given percentile in the distribution of item demand should also be in the same percentile in the distribution of item

---

[23] http://tinyurl.com/ycdjuf53

[24] http://tinyurl.com/ya6jvhdy

[25] http://tinyurl.com/y73dg5c9

[26] https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/WESP2018_Annex.pdf

quality. Since quality changes over time, the expected quality distribution of items will also change. Given this, we wanted to ensure that we were making comparisons between actual and expected quality using distributions that contained the most accurate item expected quality possible. To help ensure this, we recalculated distributions of item expected quality on the first day of each month for all items in our revision dataset. When doing so, we considered all items in existence at that time. Once these distributions were calculated, we were then able to use them in our analyses. Specifically, we compared a given item's quality at edit time with its expected quality obtained from the item expected quality distribution generated for the month after the edit occurred.

Calculating all of the item expected quality distributions presented a technical challenge since it required obtaining many ORES scores. However, we generated monthly scores for each item in chronological order and noticed that many items were not edited from one month to the next. This scenario meant that the quality score obtained from the previous month could be used, making this a tractable approach.

6.4 Results

We organize discussion of our results around the two lenses through which we consider content value.

6.4.1 Value from a Consumer-Level Perspective

To understand content value from a *consumer* perspective, we first incorporated content *quality* and *demand* into our definition of value. The primary results of our exploration into consumer-level value address the following questions: 1) What is the relationship that content edits from different strategies have with demand 2) How have those relationships changed over time? and 3) How do the dynamics identified in the first two questions affect the alignment of item quality and demand?

Manual and semi-automated strategies focus more on editing in-demand content. Recall from section 3.8.1 in Methods that we generated monthly distributions of item expected quality based on demand. Using these distributions, we computed the mean item demand/expected quality percentile across content edits and periods. Figure 6.1 shows our results. The y-axis shows the mean demand percentile and the x-axis shows each period in our study. We can see that *manual* contributors (especially anonymous manual contributors) tend to focus on the high demand items. Further, this trend has only increased over time. Across all periods, the typical anonymous manual contribution is to content that is roughly in the top 15% of demand, and the typical registered manual contribution is to content that is in the top 30% of demand. On the other hand, contributions with at least some degree of automation have less of a skew towards high-demand content. Semi-automated tools have a relatively minor skew towards high-demand content. Bots -- the most automated contribution strategy – appear to not edit in a manner that is positively associated with content demand at all and sometimes appear to even focus editing on lower-demand content. We explore bot editing in more detail next and find that the situation is more nuanced than Figure 6.1 indicates.

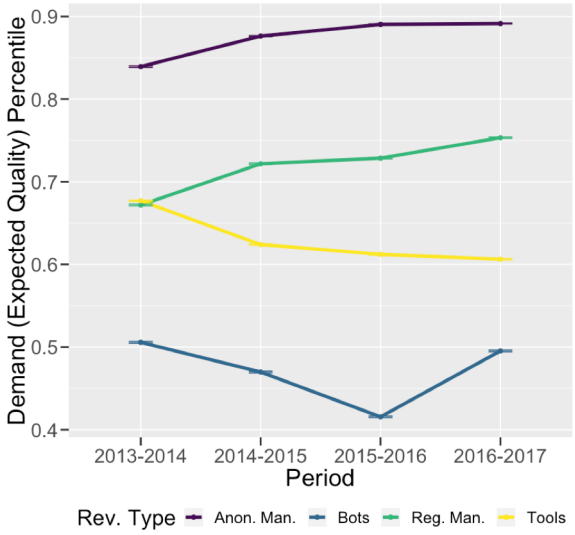

Figure 6.1: Mean Item Demand (Expected Quality) Percentile of Content Edits

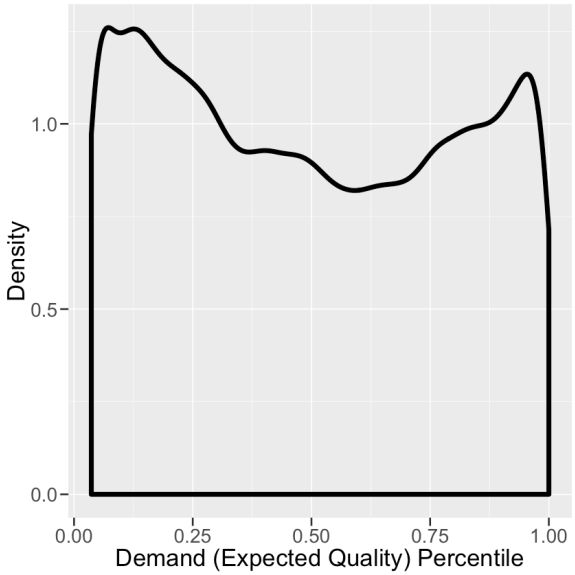

Figure 6.2: Density Function Breaking Down Demand (Expected Quality) Percentiles for Bot Content Edits in the 2016-2017 Period

Demand- and Initiative-Based Content Editing: The Twin Peaks of Bot Contributions. We further explored the distributions of expected quality percentiles for the different contribution strategies and noticed some intriguing patterns related to bot editing. Figure 6.2 shows the density function of expected quality percentiles for bot content edits for the 2016-2017 period in our study. The general trends were similar for other periods. The x-axis shows the expected quality quantile while the y-axis shows the density. Note the

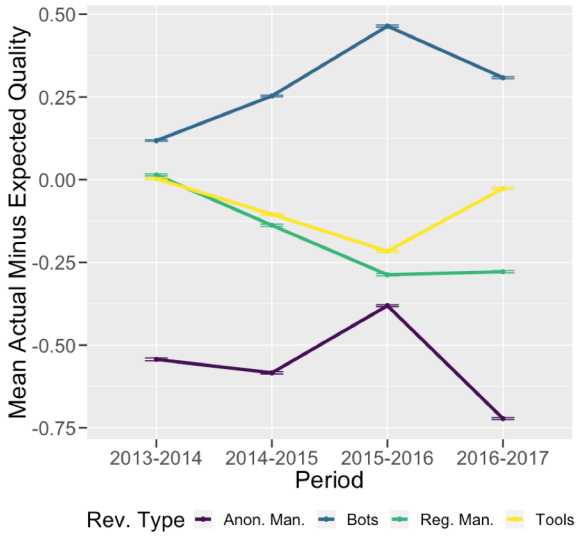

Figure 6.3: Mean Quality Difference (Actual Minus Expected) for Items within Content Edits

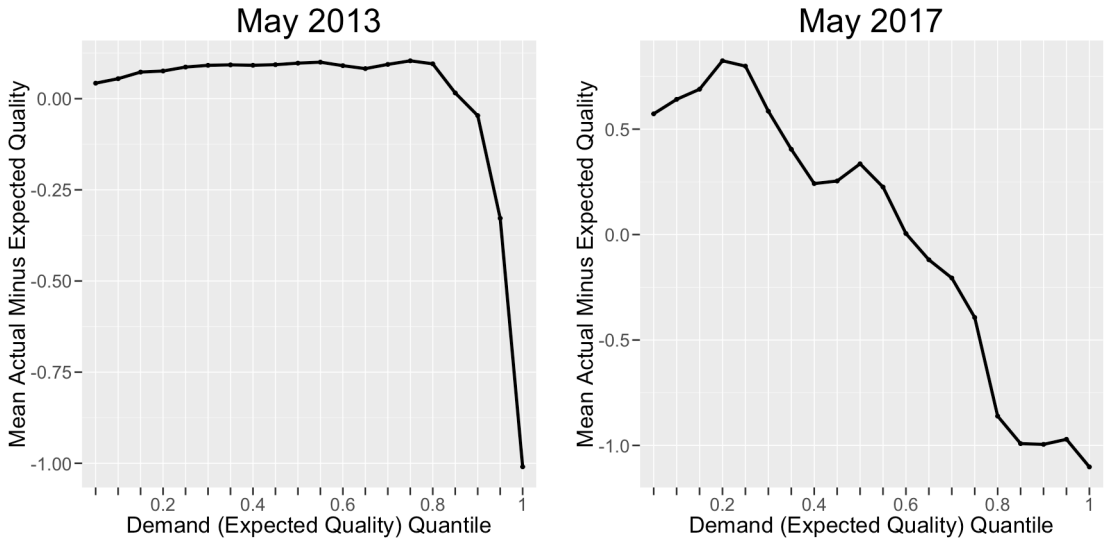

Figure 6.4: Item Mean Quality Difference (Actual Minus Expected), Broken Down by Demand (Expected Quality) Quantile. Plots represent the beginning and end of study.

two distinct rises (or peaks) in bot content edit counts. First, consider the rise on the right-hand-side of the distribution. Recall that Figure 6.1 made apparent that tool and manual contributions tend to focus on higher-demand content, but that such a trend was not clearly the case for bot contributions. The right-hand-side rise in Figure 6.2 indicates that, in fact, a sizable number of bot contributions are similar to tool and manual contributions with their focus on higher-demand content.

The rise on the left-hand-side of the plot in Figure 6.2 was also interesting to us since it was clearly responsible for the low mean item demand found in Figure 1. We wanted to understand why so much low-demand bot content editing was occurring, so we performed a qualitative coding process. Specifically, we first sampled 50 bot content edits that were below the 40th expected quality percentile. Each edit had an equal chance to have been from any of the 4 yearlong periods in our study. We then considered 1) what topics the items pertained to and 2) what aspects of the items were being edited. I carried out this process. We found that the majority of the content edits were to items that pertained to a) obscure living entities including insects, worms, plants, and crustaceans, or b) geographical locations including towns, bodies of water, and even asteroids, or c) biographical and other information on botanists, screenwriters, chess players, and films. The aspects of items that were edited varied but the majority of edits were either to descriptions or properties with quite a few label or reference edits as well (for the definitions of different aspects of Wikidata items, see [144]). Given our findings, it is evident that bots are heavily-used within initiatives that import various biological, geographical, biographical, and other large datasets into Wikidata. However, only a small amount of such information is currently in high-demand[27].

6.4.1.1 Identifying the Gaps and Surpluses Defined Through the Perfect Alignment Hypothesis (PAH).

Overall, bot contributions create content surpluses while tool and manual contributions fill content gaps. Next, we sought to understand how each contribution strategy has affected the alignment of content quality and demand as defined through the Perfect Alignment Hypothesis (PAH). Specifically, we wanted to see which strategies were creating quality *gaps* and which were creating quality *surpluses*. In Figure 6.3, we plotted out the mean difference between the actual[28] and expected quality of the items in content edits (y-axis), broken down across periods (x-axis). The quality difference was computed by subtracting expected item quality from actual item quality. In this figure, when the mean is *below* 0, this indicates editing behavior that is closing gaps while a mean *above* 0 indicates editing that is creating surpluses. The figure indicates that bots consistently create surpluses of quality while manual and tool contributors consistently fill quality gaps. Anonymous manual contributions, in particular, focus on content that has the largest quality gaps; in 3 out of the 4 time periods in our study, the typical anonymous manual contribution is to an item that is over *half* a quality class lower than the PAH suggests it should be.

Figure 6.4 provides insight into why the trends in Figure 6.3 exist. The figure shows the relationship between item alignment and demand and how this relationship has shifted over time. To derive the plots within the figure, we bucketed the items in our dataset based on the creation of 20-quantiles of item expected quality (demand). This process split an item dataset into 20 (approximately) equal-sized groups each spanning a 5-percentile range of demand/expected quality. For simplicity, we will refer to these ranges as *quantiles* and,

[27] Consider the over 400,000 Coleoptera beetles compared to the rare popularity of the Ladybug [113].

[28] The actual item quality in this figure (and other figures) representing actual item quality was calculated prior to the edits occurring.

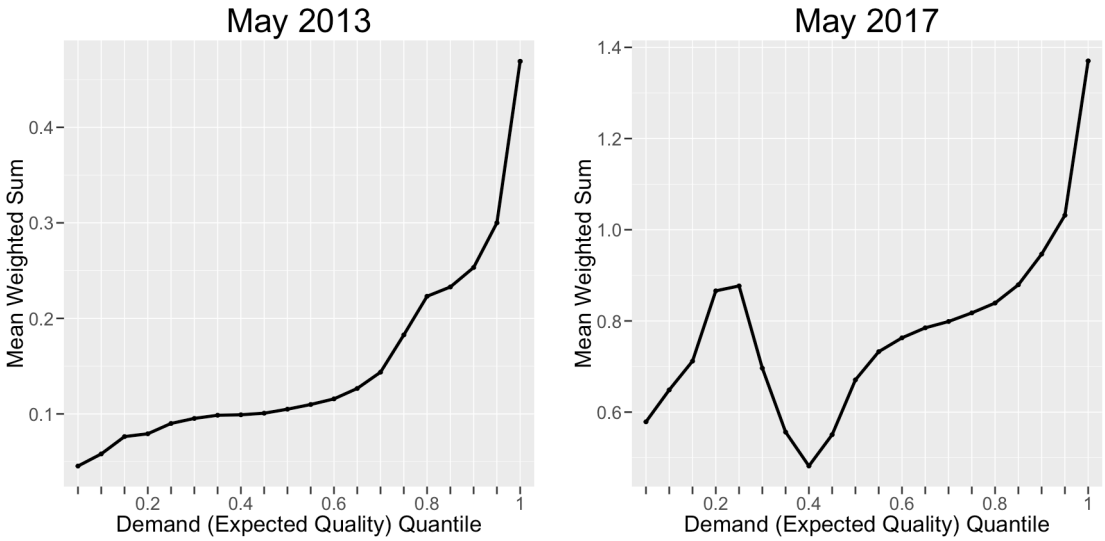

Figure 6.5: Item Mean Quality, Broken Down by Demand Quantile. Plots represent the beginning and end of study.

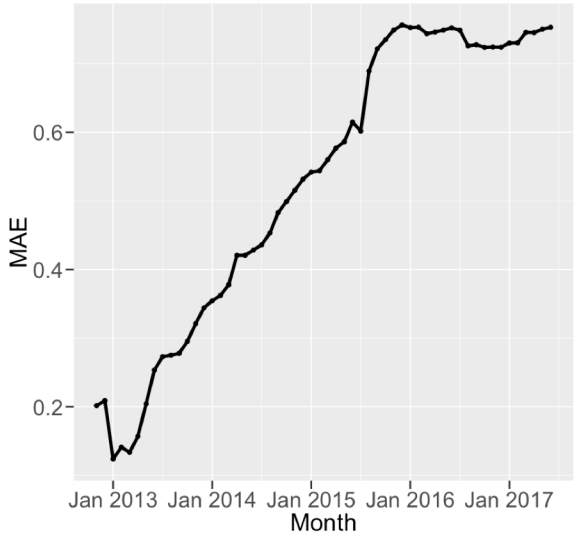

Figure 6.6: Mean Absolute Error (MAE) Based on Item Quality Difference (Between Actual and Expected Item Quality)

for our discussion, each quantile will be defined by the percentile that represents *its upper limit*. The plots show actual minus expected quality for the beginning and end of our study (similar trends existed for other points in the study). The y-axes represent the average quality difference (alignment) and the x-axes represent item demand quantiles.

The main takeaway from Figure 6.4 is that low- and midrange-demand items have a surplus of quality while high-demand items have a quality gap. Given this, Figures 6.1 through 6.4, combine to create an interesting picture of how contribution strategies affect PAH alignment. Bots tend to focus the most on low- and midrange-demand content (Figures 6.1 and 6.2) and this means the typical bot content edit adds a surplus of quality to items (Figure 6.3) since low- and midrange-demand content typically has a surplus of quality (Figure 6.4). Additionally, since tools and manual contributions tend to focus on relatively higher-demand content (Figure 6.1), this results in their typical edits "counteracting" the work of bots by filling in content gaps (Figure 6.3) since such high-demand content typically has a quality gap (Figure 6.4).

As a result of bot editing, increasing numbers of high- and midrange-demand quantiles need quality. Note the points in Figure 6.4 when each of the quality difference curves cross over 0 as demand increases. These points indicate where the mean item quality difference (actual minus expected quality) shifts from a surplus in quality to a gap in quality. In the first period within our study, we can see that this crossover occurred between the 85th and 90th percentiles. Over time, this crossover shifted downward to between the 60th and 65th percentiles. This shift indicates that more and more mid- and high-demand item quantiles have an overall quality gap. The bot-driven creation of a surplus of quality in low- and midrange-demand quantiles has created this shift. Figure 6.5 shows the mean quality (y-axis) in each demand quantile (x-axis), prior to the first and last periods in our study. We can see that, initially, the average quality of items increased monotonically with demand. However, over time, the focus of bot editing on low-to midrange-demand content pushed the quality of some of this content higher than that in some neighboring higher-demand quantiles. Such behavior also increases the expected quality of this higher-demand content. And, since the higher-demand content tends to have not improved in quality as much as the lower-demand content, the gap between this higher-demand content's actual and expected quality has increased. As stated earlier, bots provide most edits in Wikidata[29] and, because of this, they can drive such large-scale changes.

6.4.1.2 Considering Wikidata-Wide Alignment Longitudinally

Bot editing has perpetually decreased the alignment of the average item. We wanted to precisely measure the longitudinal effect of content editing behaviors on the alignment of the typical item. For each month in Wikidata's history, we applied a common error metric -- *mean absolute error (MAE)* -- to compare the absolute value of the difference between the actual and expected quality of all items in existence at that time. Figure 6.6 shows the results. The MAE is on the y-axis and each month is on the x-axis. Note the clear, steady increase in misalignment over time. Initially, the typical item had an actual minus expected quality difference of ~0.2 while this has now shifted to ~0.75. These results provide a clear view of the community impact that misaligning bot content edits to low- to midrange-demand items have had over time. We return to this figure in Discussion, Implications, and Future Work.

[29] Only a small proportion of bot edits were filtered out as non-content edits

6.4.2 Value from a Societal-Level Perspective

We consider societal-level value to be *maximized* when harmful biases in the representation of protected- or minority-related content are *minimized*. The primary results of our exploration into societal-level value center around the following questions: 1) How does the representation of content vary along important dimensions (i.e. male versus female, Global North versus Global South, and urban versus rural)? 2) How has representation changed over time? and 3) How do different contribution strategies affect such representation? Understanding how different strategies affect representation is important to aid content curators and other Wikidata community members in reducing problematic biases.

To consider representation of content, we performed some basic descriptive statistics and compared the results between groups (e.g. between female and male content) within each dimension of societal-level value that we analyzed. We computed all statistics at yearly intervals throughout our study (May 2013 through May 2017). As our first statistic, we simply determined the number of items in a group. We refer to this statistic as *item coverage*. When comparing item coverage between groups, we will show that our results uncover biases between the coverage of different groups (e.g. between male and female content). It is not clear what proportion of such biases are due to under-representations in the real-world (e.g. due to cultural biases or sexism) and what proportion are due to any biases in Wikidata's representations of the real-world. It is also not clear what content coverage the Wikidata community should strive for. For the purposes of our work, a non-controversial coverage goal for the Wikidata community would be for it to reflect trends in the real-world. However, even identifying precisely what such real-world coverage looks like can be difficult and was not a goal of this work. Our coverage results shed light on the shifts in differences in coverage between groups and, in some cases, our results also provide intuitive indications that Wikidata's representation of the real world is problematically biased away from protected or minority population content.

In addition to item coverage, we also considered the item quality of the typical item in a group. We refer to this statistic as *item quality* and it simply is mean item quality. Intuitively, the Wikidata community should seek to provide consumers of protected/minority population content and consumers of privileged population content the same quality content. If quality is lower for protected/minority group content than that of a privileged group, consumers of the former content will be more likely to create inaccurate or less complete perceptions of topics relating to such group.

6.4.2.1 Gender-Related Bias

Human male item coverage is/has been much greater than human female coverage. We found that in May 2013, Wikidata contained 5.43 human male items for every human female item. By May 2017, there were 5.01 human male items for every human female item. Even though we do not derive what the ratio between human male and human female content should be (based on real-world representation of male and female topics), having 5 male items for every female item is obviously problematic. Further, given the quite

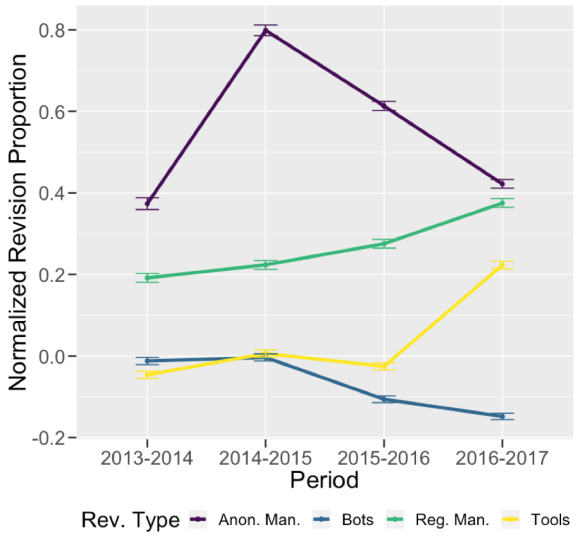

Figure 6.7: Normalized Human Female Item Content Edit Proportion

mild decline in the item coverage ratio, our results indicate that the problematic coverage bias in favor of male content is not fixing itself quickly.

Human male and female item quality is/has been roughly the same. While a large discrepancy exists between human female and human male content in terms of item coverage, we also found that human female item quality has consistently been approximately equal to that of human male items. As of May 2013, the item quality metric was 0.16 for human female items and 0.18 for human male items. As of May 2017, item quality was 1.28 for human female items and the same for human male items.

Manual contributors disproportionately focus on female items. Bots disproportionately focus on male items. To help explore how each contribution strategy has affected the representation of female and male content, we computed the proportion of content edits that occur to human female items relative to human male items. We performed this separately for each strategy and, to see how behaviors changed over time, for each period. Since, human male items significantly outnumber/have outnumbered human female items, it was unsurprising that contribution strategies tended to also devote significantly more editing to human male item content. However, we wanted to get a sense for how each contribution strategy was affecting representation shifts. This meant we needed to account for the fact that more human male items exist than human female items. To do so, we normalized each content edit proportion based on the proportion of female items to male items in existence at the beginning of a respective year-long period. Specifically, we divided the given content edit proportion by the item proportion. Finally, we subtracted 1 from the result. This last step led to a set of results that were intuitive to interpret. For example, a contribution strategy with a normalized human female content edit proportion of 0.2 would indicate that the strategy performs 20% more

content editing on human female content over the period than we would expect given disparities in item coverage at the beginning of that period. Positive scores indicate more than expected human female content editing and negative scores indicate more than expected human male content editing. Figure 7 shows our results. The y-axis is the normalized revision proportion and the x-axis is the time period.

The results in Figure 6.7 first indicate that manual (both registered and anonymous) contributors focus relatively the most on female-related items and bots focus the least. Second, given the normalizations performed, our results also indicate that manual contributors tend to have a consistent disproportionate focus on human female items. Anonymous manual contributions produce as much as ~80% more content edits towards female items than we would expect. Further, depending on the year, tools either disproportionately focus more or less on human female items than we would expect. Finally, bots have a consistent disproportionate focus on human male items, and this is becoming increasingly the case over time.

6.4.2.2 Global North versus Global South-Related Bias

A disparity in item coverage between Global South and Global North items still exists, but has significantly decreased. We found that in May 2013, Wikidata content contained 2.17 Global North items for every Global South item. By May 2017 the disparity in this ratio had declined substantially to 1.18. Much of this decline had occurred in the last two years of the study. The two-year period between May 2015 and May 2017 saw the creation of a particularly large number of items pertaining to the Global South (976,507 items). If the current trend continues, Wikidata will have the same number of Global South items as Global North items very soon.

Global North item quality is/has been higher than that of Global South items. While item coverage of Global South content has come close to matching that of the Global North in recent years, the average quality of Global North items has been consistently higher than that of Global South items, but only slightly. As of May 2013, the item quality metric was 0.09 for Global South items and 0.21 for Global North items. As of May 2017, the item quality metric was 0.62 for Global South items and 0.74 for Global North items. These differences are small enough such that they would result in a barely noticeable difference between the two groups.

Bots and anonymous manual contributors disproportionately focus on Global South items. Tools and registered manual contributors disproportionately focus on Global North items. Figure 6.8 shows the normalized content edit proportion between Global South and Global North content edits. This figure is analogous to the normalized revision proportion figure that was computed to study gender biases (Figure 6.7). In this figure, proportions greater than 0 indicate a greater than expected proportion of edits to Global South content while proportions less than 0 indicate a lesser than expected proportion. From the figure, we can see that bots and anonymous manual contributors focus the most on Global South content and registered manual contributions focus the least on such content. Further, while bots and anonymous manual contributors

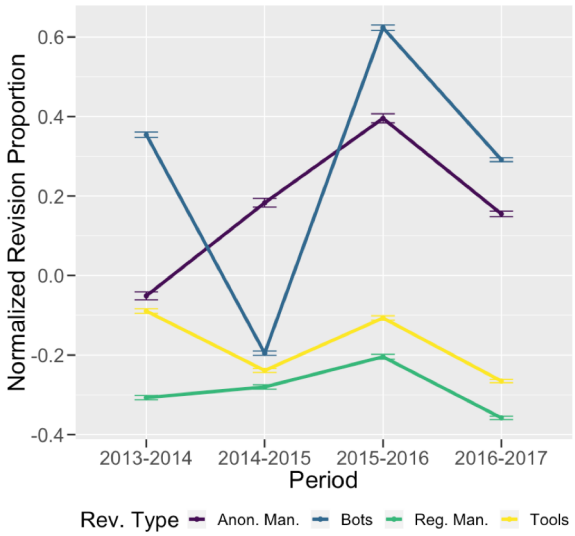

Figure 6.8: Normalized Global South Item Content Edit Proportion

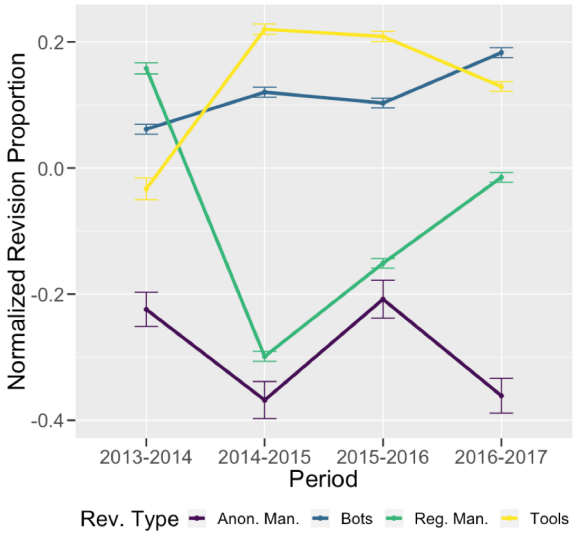

Figure 6.9: Normalized Rural Item Content Edit Proportion

(generally) have a consistent disproportionate focus on Global South content, tools and registered manual contributors have a consistent disproportionate focus on Global North content.

6.4.2.3 Urban versus Rural-Related Bias

Urban item coverage is/has been greater than rural item coverage. We found that in May 2013, Wikidata contained 1.93 (U.S.) urban (metropolitan) items for every (U.S.) rural (nonmetropolitan) item. By May 2017, this ratio had slightly declined down to 1.82.

Rural item quality is/has been higher than that of urban items. We found that rural item quality has been consistently higher than urban item quality. The quality difference has widened slightly but is still relatively small. As of May 2013, the item quality metric was 0.14 for rural items and 0.08 for urban items. As of May 2017, the item quality metric was 0.85 for rural items and 0.66 for urban items. Our results are promising in the sense that they indicate the community has succeeded in providing relatively high-quality representations of rural regions that have been shown to lack high-quality representations in prior peer production work (e.g. [42,66,100]).

Automated contributions disproportionately focus on rural items. Manual contributions disproportionately focus on urban items. Figure 6.9 (which is analogous to Figures 6.7 and 6.8) shows the normalized content edit proportion between rural and urban content edits. Proportions greater than 0 indicate a greater than expected proportion of edits to rural content while proportions less than 0 indicate a lesser than expected proportion. From the figure, we can see that (generally) bots and tools focus the most on rural content while anonymous contributors consistently focus the least on such content. Further, while bots and tools have a (generally) consistent disproportionate focus on rural content, manual contributors have a (generally) disproportionately focus on urban content. We revisit these findings in Discussion, Implications, and Future Work.

6.5 Discussion, Implications, and Future Work

6.5.1 Should Bots – and Wikidata – adjust editing behavior?

Bots should not be forced to align editing behavior with the PAH. Our results from applying the PAH indicate that – year after year -- the typical bot content edit is inefficient at provide quality content where there is demand, instead creating a surplus of quality in low-demand content. Given the dominance of bot editing, such misaligning behavior has a large effect on the overall lack of alignment in Wikidata. This result raises a question: *Should bots change their behavior to focus on closing content gaps?* Further, and more generally: *Is total PAH alignment realistic in Wikidata?*

To explore these questions, it is useful to first consider the PAH in the context where Warnacke Wang et al. [95] originally defined and applied it: *Wikipedia*. Warnacke Wang et. al. note that it's unclear if attaining total alignment in Wikipedia is an appropriate goal. They raise the concern that if the community decides to lessen fundamental contributor freedoms by insisting that editing be directed towards high-demand content, such initiatives might negatively impact community health and productivity. Given that Wikidata and Wikipedia share similar core contributor values -- in particular, related to contributor freedom – it is also unclear whether total alignment is an appropriate goal in Wikidata.

In fact, total alignment might be *even less* an appropriate goal in Wikidata. To understand why, recall that most Wikipedia edits are performed *manually*, but most Wikidata edits are *automated*. Since each manual edit relies on manual *effort*, it intuitively makes sense to optimize such editing towards high-demand content

to maximize the consumer benefit from the effort made. However, for automated edits, it not as clear that such a strategy makes sense. In fact, given our automated editing results, it actually may be inappropriate to expect that automated Wikidata editing should mostly focus on high-demand content. This is because of the following. Automated Wikidata edits often occur through bots that import large external datasets of information. A large amount of manual effort is devoted upfront by bot creators to create bots to process data. However, once created, a bot can often quickly perform large-scale editing over an entire dataset with little or no additional manual effort. Given this, it would not make sense in Wikidata to only import the high-demand data from datasets. This is because -- assuming adequate/available computing resources -- the lower-demand data that composes the majority of bot-imported datasets can be processed with essentially no additional community effort. While such lower-demand data does not provide as much consumer-value as higher-demand content does (i.e. it is not viewed as much), it still provides valuable content to consumers, and is therefore worth importing into Wikidata.

All contribution strategies – especially bots and tools -- should be more cognizant of perpetuating problematic biases. While we do not suggest that bots change their editing behavior to account for PAH alignment, this does not mean that they should continue any behaviors that perpetuate problematic biases. Our results point to bots consistently having a disproportionate focus on male- and Global North-related content. Further, our results also indicate that all other contribution strategies have a consistent disproportionate focus on content related to at least one privileged group, at the expense of content related to a protected or minority group. Moving forward, all contribution strategies should carefully consider the potential to perpetuate problematic societal biases when they are importing data. This is particularly important for bots and tools given their ability to edit rapidly and at-scale.

6.5.2 Potential Mechanisms to Support Aligned Content Editing

We argue above that total PAH alignment may not be an appropriate goal in Wikidata. However, if any contributor editing behaviors are already filling in content gaps -- and hence pushing content towards PAH alignment -- then it is still useful from a consumer value-perspective to support such behaviors. Our results indicate that manual and semi-automated tool content editing is and has been counteracting some of the misalignment created by bots by filling in content gaps in midrange- to high-demand content. Given this, community initiatives and socio-technical tools may be useful to help incentivize and facilitate such contributors. We next provide some specific suggestions.

Gamified Socio-Technical Tools. One possible means to facilitate and increase manual and tool content editing is through gamification. Outside of Wikidata, gamified applications such as Foldit [157] have been used to collect useful user-generated content [49]. Further, within Wikidata, some gamification has already occurred in the form of a tool called the Wikidata Game [158]. Additional development and promotion of gamified tools in Wikidata is a promising direction to help counteract the misalignment created by bots.

Socio-Technical Tools to Surface Content Demand. Socio-technical tools could potentially help identify in-demand content. There are indications that Wikidata contributors would find these tools useful. Specifically, one indication stems from our own findings. Our results in section 4.1 indicate that the more manual effort a contribution strategy requires in order to edit, the more the strategy focuses on in-demand items. Since manual effort comes from humans, this result provides some reason to believe that humans are *intentionally* focusing on Wikidata content that is in high-demand. Another indication stems from the socio-technical tools used in other peer production communities. SuggestBot [14] is a prominent tool used by the Wikipedia community and provides insight into article page view rates. Since Wikipedia and Wikidata are sister communities that share many community characteristics (and contributors!), tools similar to SuggestBot might be effective to facilitate high-demand content editing in Wikidata.

6.5.3 Using Automated Contribution Strategies to Reduce Rural Content Disparities

Prior work by Johnson et al. [42] found that there exists lower quality peer-produced geographic information in rural areas and that the potential pool of contributors in such regions is smaller than in urban areas. Further, the work notes that while automation has been looked at with skepticism in peer production (flawed automation has caused significant problems in Wikipedia and OpenStreetMap, e.g. [99,103]), it might be a good option for rural content since few alternatives exist. Our work backs up this recommendation with an example of the effectiveness of automation in representing rural content. Recall that we found that the item quality for rural items was higher than that of urban items. Bots and tools focused more on such content than manual contributors did. This fact, combined with the additional fact that the bot and tool populations we sampled our urban and rural edits from were much larger than those of manual contributions provides a compelling indication that automation played a large role in ensuring that rural item quality was relatively high.

6.5.4 Future Work

To shed additional light on our findings, future qualitative work should seek to better understand how Wikidata contributor motivations relate with the lenses of content value that we considered. Do contributors consider the potential to introduce or reduce problematic societal biases when importing data? Is content demand considered? If such factors are considered, how do they change behavior (if at all)? Such a study might include interviews or surveys with bot and tool operators and creators as well as with manual contributors.

Figure 6.6 provided a clear indication of the misaligning effect that bot editing has had on the overall alignment of Wikidata for most of its history. The trend of decreased alignment tends to plateau at the end of our study even though bot contributions still tend to add quality surpluses and manual and tool contributions still tend to fill quality gaps. We believe a likely explanation for the plateau behavior is an uptick in the creation of new content where the content's demand and quality tend to have only a relatively small misalignment. This would temporarily "mask" the misaligning behavior of bots in this figure. Moving

forward, we anticipate that alignment will continue to decrease. Future work should continue to monitor the overall alignment of Wikidata.

6.6 Limitations

While our analyses identified important societal disparities in content representation, further methodological improvements could occur to better account for the content demand that is specific to minority and protected populations. Our metric of content demand currently does not distinguish content demand coming from -- for example -- the Global North versus the Global South. However, our metric is also somewhat biased towards the needs of consumers in the Global North and those who are in urban areas. This is because more page views are occurring to Global North and urban content – a likely result of more consumers living in those areas. When considering the mean demand percentile of items each year from the beginning to the end of our study, Global North and urban item means were almost always higher than Global South and rural item means. This has grown to especially be the case for Global North and Global South content. As of the end of our study in May 2017, the mean demand percentile was 0.55 for Global North items and 0.35 for Global South items. Future work should hone metrics of content demand to specifically target the needs of minority and protected populations.

When considering how contribution strategies affect different dimensions of societal-level value, we assume that content edits to a certain group (e.g. human female items) average out to add the same amount of quality as the average quality of content edits to the group they are compared against (e.g. human male items). In reality, the amount of quality that a content edit provides is a function of the edit-time quality of an item. Content edits provide less quality to items that are higher quality at edit-time. However, we felt confident in our approach since the item quality of content within the various groups we were comparing (e.g. male versus female content) tended to be relatively quite close to each other on the Wikidata item quality scale.

7 Conclusion

7.1 Summary of Studies

In recent years, certain peer production communities have focused on producing structured data specifically designed for use by applications and algorithms. We identified two fundamental challenges associated with creating valuable content from a machine processing perspective. We summarize our explorations of those challenges next.

The first challenge focused on a tension unique to the context of peer-produced structured content. This tension is between the peer production ethos of contributor freedom and the need for highly-standardized structured data for effective machine readability. We studied this tension in two ways. First, we did an interview study focused on OpenStreetMap's knowledge production processes to investigate how – and how successfully – this community creates and applies its data standards. In this study, we extracted six themes that manifested the tension of freedom and standardization and three overarching concepts – correctness, community, and code – that help make sense of and synthesize the themes. Within the themes, we identified various groups within the OSM community (such as Humanitarian OpenStreetMap) that defined their own set of OSM tagging standards. We also identified cultural challenges to defining the global tagging standards found within OSM. Finally, we offered suggestions for improving OpenStreetMap's knowledge production processes, including new data models, sociotechnical tools, and community practices (e.g., stronger leadership).

Once we had identified challenges to standardization that resulted from contributor freedom, we wanted to measure the effect they had on data standardization. To do so, we carried out a second study in order to investigate adherence to OpenStreetMap's structured data recommendations/guidelines. We found that most applied structured data was consistent with the community's standards; however, we also found that the standards identified many opportunities for applying data that were not achieved. In addition, when we situated the standards in the context of OpenStreetMap's data model, we found a significant amount of ambiguity; the syntax allowed only one value, but everyday meaning – and the standards themselves – called for multiple values. These results suggested significant opportunities for OpenStreetMap to produce additional valuable content to power applications.

While initial contributions to peer production were largely manual work (e.g., manual Wikipedia article writing), automated contributions have played an ever-increasing role in these communities. Wikipedia, OpenStreetMap, and Wikidata all have taken advantage of automation to perform work at a rate and scale exceeding that of manual contributors. Given the large prevalence of automated contributions in creating structured content, it is important to understand how they *compare* to manual contributors in their effect on content value. The second challenge covered in my thesis focused on making this comparison. We first wanted to ensure that consistent differentiations between automated and manual contributions occurred. In

peer production, bot activities are not always explicitly flagged and could be mistaken for human contributions. Hence, in the third study in my thesis, we developed a machine classifier to detect previously unidentified bots using implicit behavioral and other informal editing characteristics. We showed that this method yields a high level of fitness under both formal evaluation (PR-AUC: 0.845, ROC-AUC: 0.985) and a qualitative analysis of "anonymous" contributor edit sessions. Our findings indicate that, most of the time, unidentified bots do not perform a significant portion of edits. However, we also showed that, in some cases, unflagged bot activities can significantly misrepresent manual behavior in analyses. Our model also has the potential to support future research and community patrolling activities.

For our final study, we explored how Wikidata's automated and non-automated contributions differ in the value they produce. In performing this exploration, we define content value through two important and intuitive lenses. These lenses consider 1) the relationship between content quality and consumer demand and 2) problematic societal-level biases. Our results indicate that automated contribution mechanisms are less effective than manual contributions at targeting work based on consumer demand. However, automated mechanisms also appear effective in improving the quality of underrepresented content (e.g., pertaining to rural areas and the Global South). Based on our findings, we provide actionable insights for Wikidata and other peer production communities.

7.2 Summary of Contributions

My work has provided the following set of contributions towards better understanding the value of peer-produced structured content from a machine perspective.

- We show *why* the large degree of contributor freedom affects the ability of peer production communities to be standardized. For example, some contributors – through greater technical skill or dedication to a cause – are able to influence standards. Cultural differences also cause standardization problems – for example, a "highway" can have different definitions in different regions.
- We also quantify standardization in OSM and find that 1) most applied metadata is consistent with the standard, 2) the constraints of the OSM data model lead to a large amount of ambiguous metadata, and 3) the informal standard of the OSM wiki defines large unmet opportunities to apply useful metadata.
- Based on our results from considering the tension of freedom and standardization, we offer several new sociotechnical strategies and tools to improve standardized data creation in peer production communities. For example, our work problematizes OSM's 1:1 tagging structure, motivates the need to be able to link similar entities, and informs the design of tools that can improve standardization without increasing the effort required to contribute.
- We describe an effective bot detection strategy using machine classification and implicit behavioral and other informal editing characteristics and show that it is effective in identifying unflagged bot activity in Wikidata.

- Using our bot detection model, we show that, for the most part, unflagged bot activities are rare. Hence, *the activities of unflagged bots appear unlikely to significantly change results in many studies of human and bot behavior*. However, analyses that are sensitive to outliers (e.g. max session duration) or that span relatively short subsets of Wikidata history should first extract previously unidentified bots from "human" contributions to ensure accurate analyses.
- We show that there is a meaningful amount of non-compliance with the *bot policy* in Wikidata. According to our model, 3% of registered "human" user edits overall and 2% of anonymous "human" user edits overall are from bots. These percentages are important since all bot activity that does not align with community policy matters to the Wikidata community.
- We make our bot detection datasets and bot detection code available under an open license: https://github.com/hall1467/wikidata_bot_prediction_model. This will facilitate future Wikidata contributor behavior research as well as better allow the Wikidata community to enforce its bot policies.
- In Wikidata, we find that manual contributors tend to work on higher-quality and higher-demand content than bots. Further, while the work performed by manual contributors tends to increase the alignment between content quality and demand, the work performed by bots has resulted in a misalignment between content quality and demand, which has increased over Wikidata's history. We also argue that complete alignment of content quality and demand may not actually be a realistic or even desirable scenario in a community like Wikidata where automated contributions dominate and contributor freedom is highly-valued.
- We provide evidence that distinct contribution strategies play different roles in affecting the representation of minority or protected content in Wikidata. For example, bots tend to disproportionately focus on Global South and rural content versus Global North and urban content. However, they also disproportionately focus on male-related content versus female-related content. On the other hand, manual contributions disproportionately focus on female-related content, but also disproportionately focus on urban content. Further, we find that unregistered manual contributions can have distinct behavior from registered manual contributions. For example, unregistered manual contributions tend to disproportionately edit Global South content while registered manual contributions tend to disproportionately edit Global North content.
- Our work provides implications for targeting contribution strategies to optimize for content value. For instance, our findings motivate initiatives and tools to help manual contribution strategies focus on high-demand content.

7.3 Implications and Future Work

The four previous chapters each discussed implications for the respective individual studies. I conclude this thesis by providing a brief, higher-level discussion of implications as well as some potential directions for future work.

Socio-Technical Tools. Our studies motivate many contributor *socio-technical tools* as mechanisms to improve the value of peer-produced structured content. For example, our results in Study 4 indicate that manual contributors tend to focus on relatively high-demand content compared to other contribution strategies. Tools could help facilitate such contributors in finding such content. Further, tools could be useful in both OpenStreetMap and Wikidata to increase community awareness about which contributors are increasing or decreasing problematic societal-level biases. Finally, our work has indicated that many opportunities exist to apply structured data within OpenStreetMap. Tools could help contributors efficiently add more tags. While OpenStreetMap and Wikidata are already employing some tools to help contributors apply structured data, our findings indicate that some of these tools also inconsistently enforce community standards in favor of creating their own. This results in data standardization issues that make the data less valuable to applications and algorithms. Given this, it is important for any new or existing tools to ensure that they accurately embody community standards.

Working Towards Fair Representation within Communities. Of course, even if tools accurately embody community standards, it is still problematic if such standards do not represent the views of the entire community. Our Study 1 results (and results in prior work, e.g. [87]) indicate that hostile and sexist behavior occurs in OpenStreetMap and that, further, certain initiatives such as Humanitarian OpenStreetMap are more influential than others in driving the evolution of community standards. This is problematic because, when not all are given a voice, community standards and practice will then be biased towards the views of a subset of the community. Further, the resulting effects on excluded individuals are problematic as well. While some excluded individuals have become highly-successful activists in communities such as Wikipedia [29,159], this scenario is rare. It is more likely that those whose voices are not heard will either 1) leave the community or 2) perform contributions in their own way. The latter outcome results in standardization issues. Future work should prototype and test various socio-technical means of monitoring community interactions for unhealthy behavior. Additionally, future work should focus on studying ways to balance both the needs of large-scale, important initiatives such as Humanitarian OpenStreetMap and the needs of other community members.

Resolving Misalignments between Practice and Standards. Communities like OpenStreetMap have a complex relationship between practice and standards. In OpenStreetMap, tagging practice tends to influence standards, but standards tend to also influence practice. Discrepancies inevitably occur between practice and standards as a result of contributor freedom and (the above mentioned) unequal power dynamics and biases. When such discrepancies occur, it is not always clear whether practice or data standards better represent community consensus. To facilitate the consistent representation of community consensus within community tagging standards and tagging practice, it would be useful to have tools that monitor practice and standards and then display where they align/differ. Further, tools could also show which contributors/contributions are responsible for the inconsistencies.

Importance of Addressing Standardization Issues when Creating Structured Data. As noted previously in this thesis, it is especially important for a peer production community to address standardization issues if it is producing *structured data*. While *unstructured* data (e.g., Wikipedia articles) is often still comprehensible by human consumers even if the data is not standardized well (e.g., even if Wikipedia article structures or formats vary), the applications and algorithms consuming *structured* data will not effectively process such data unless it is highly standardized. Thus, communities that produce large amounts of structured data such as OpenStreetMap and Wikidata should especially invest energy into sociotechnical or other mechanisms of improving contributor representation and consensus building. Such mechanisms should particularly focus on addressing community issues that we and others have identified (e.g., hostility) that can affect standardization.

Better Understanding the Relationship between Contributor Motivations and Value Metrics. The final study in this thesis identifies the effects that different contribution strategies have on different lenses of content value. A natural follow-up study would seek to understand the contributor motivations that resulted in the trends we observed. Such a study might include interviews or surveys targeted at manual contributors and the creators and operators of bots and tools. Study questions could inquire about how contributor motivations relate (if at all) with content quality, demand, and problematic societal biases.

Better Understanding Application Needs. Another potential follow-up study to the work in this thesis would be a qualitative exploration of content value from the perspective of application and algorithm designers and creators. Such a study could inquire into what content characteristics are the most valuable to applications and algorithms and what issues are dealt with when processing peer-produced data. The study could look for general themes and challenges across many different applications and algorithms to help guide communities in making content that is optimized for value and for widespread usage. When performing such a study and publishing its results, care would need to be taken to not create the appearance that peer production communities are beholden in any way to certain large corporations or proprietary applications (e.g. Google Knowledge Graph). After all, content produced in peer production is open for all to use and it would not be in line with peer production principles to create content that is designed specifically for use within only certain proprietary applications.

Being Cognizant of the Effect of Changes on Contributor Freedom. Finally, it is worth stating that while I have provided suggestions towards improving the value of structured content created in peer production communities, communities that consider making changes to improve content value should carefully analyze the potential effects those changes have on contributor freedom. Communities such as Wikipedia, Wikidata, and OpenStreetMap were founded on principles of contributor freedom. As has been shown to be the case in Wikipedia [30], making community changes to improve content value can reduce contributor motivation if contributor freedom is inhibited. Given this, it would perhaps be too idealistic to assume that data produced by these communities can always be optimized to provide the most possible value to applications. Hence, it

is evident that applications and algorithms using peer-produced structured data (e.g. Google Knowledge Graph, Mapbox, etc…) will necessarily have to learn to deal with some inconsistencies and other data problems when using peer-produced structured data.

8 Bibliography

[1] Ahmed Loai Ali, Falko Schmid, Rami Al-Salman, and Tomi Kauppinen. 2014. Ambiguity and plausibility: managing classification quality in volunteered geographic information. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 143–152. Retrieved November 9, 2015 from http://dl.acm.org/citation.cfm?id=2666392

[2] Rachel Baker, Thomas Dee, Brent Evans, and June John. 2018. Bias in Online Classes: Evidence from a Field Experiment. CEPA Working Paper No. 18-03. *Stanford Center for Education Policy Analysis* (2018).

[3] Andrea Ballatore and Michela Bertolotto. 2011. Semantically Enriching VGI in Support of Implicit Feedback Analysis. In *Web and Wireless Geographical Information Systems*, Katsumi Tanaka, Peter Fröhlich and Kyoung-Sook Kim (eds.). Springer Berlin Heidelberg, 78–93. Retrieved May 12, 2015 from http://link.springer.com/chapter/10.1007/978-3-642-19173-2_8

[4] Andrea Ballatore and Peter Mooney. 2015. Conceptualising the geographic world: the dimensions of negotiation in crowdsourced cartography. *International Journal of Geographical Information Science* 29, 12 (2015), 2310–2327.

[5] Andrea Ballatore and Peter Mooney. 2015. Conceptualising the geographic world: the dimensions of negotiation in crowdsourced cartography. *International Journal of Geographical Information Science* 0, 0 (August 2015), 1–18. DOI:https://doi.org/10.1080/13658816.2015.1076825

[6] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1075–1084. Retrieved September 18, 2016 from http://dl.acm.org/citation.cfm?id=2208553

[7] Christopher Barron, Pascal Neis, and Alexander Zipf. 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS* 18, 6 (December 2014), 877–895. DOI:https://doi.org/10.1111/tgis.12073

[8] Yochai Benkler and others. 2016. Peer production and cooperation. *Handbook on the Economics of the Internet* 91, (2016). Retrieved May 10, 2017 from http://books.google.com/books?hl=en&lr=&id=jXwhDAAAQBAJ&oi=fnd&pg=PA91&dq=info:Vb-VOwGptmsJ:scholar.google.com&ots=bNxztCB_E2&sig=ybRGYAE85tzPC87U5jsPfd_CiDk

[9] Mikhail Yuryevich Bilenko. 2006. *Learnable similarity functions and their application to record linkage and clustering*. Retrieved September 24, 2015 from http://www.cs.utexas.edu/~ml/papers/marlin-proposal-03.pdf

[10] Nama R. Budhathoki and Caroline Haythornthwaite. 2013. Motivation for Open Collaboration Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist* 57, 5 (May 2013), 548–575. DOI:https://doi.org/10.1177/0002764212469364

[11] US Census Bureau. Census.gov. Retrieved June 4, 2018 from https://www.census.gov/en.html

[12] Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology* 62, 10 (2011), 1899–1915.

[13] Chris Ciaccia. 2019. Doris Day Wikipedia page defaced with graphic image after her death. *Fox News*. Retrieved May 21, 2019 from https://www.foxnews.com/tech/doris-day-wikipedia-graphic-image

[14] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*, 32–41. Retrieved September 16, 2016 from http://dl.acm.org/citation.cfm?id=1216309

[15] Nikola Davidovic, Peter Mooney, Leonid Stoimenov, and Marco Minghini. 2016. Tagging in Volunteered Geographic Information: An Analysis of Tagging Practices for Cities and Urban Regions in OpenStreetMap. *ISPRS International Journal of Geo-Information* 5, 12 (2016), 232.

[16] Melanie Eckle and João Porto de Albuquerque. 2015. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. In *ISCRAM*.

[17] Killian Fox. 2012. OpenStreetMap: “It’s the Wikipedia of maps.” *the Guardian*. Retrieved March 7, 2018 from http://www.theguardian.com/theobserver/2012/feb/18/openstreetmap-world-map-radicals

[18] R. Stuart Geiger. 2011. The lives of bots. (2011).

[19] R. Stuart Geiger and Aaron Halfaker. 2013. Using edit sessions to measure participation in Wikipedia. In *CSCW*, 861–870.

[20] R. Stuart Geiger and Aaron Halfaker. 2013. When the levee breaks: without bots, what happens to Wikipedia’s quality control processes? In *OpenSym*, 6.

[21] R. Stuart Geiger and Aaron Halfaker. 2017. Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of “Even Good Bots Fight.” (2017).

[22] Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature*. DOI:https://doi.org/10.1038/438900a

[23] Sean Gillies. *Shapely 1.6.4.post2 : Geometric objects, predicates, and operations*. Retrieved April 4, 2019 from https://github.com/Toblerity/Shapely

[24] Jean-François Girres and Guillaume Touya. 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14, 4 (August 2010), 435–459. DOI:https://doi.org/10.1111/j.1467-9671.2010.01203.x

[25] Andreea D. Gorbatai. 2011. Exploring Underproduction in Wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (WikiSym ’11), 205–206. DOI:https://doi.org/10.1145/2038558.2038595

[26] Mordechai Haklay. 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets , How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environ Plann B Plann Des* 37, 4 (August 2010), 682–703. DOI:https://doi.org/10.1068/b35097

[27] Muki Haklay and Nama Budhathoki. OpenStreetMap–Overview and Motivational Factors. *ResearchGate*. Retrieved September 18, 2016 from https://www.researchgate.net/publication/44295974_OpenStreetMap-Overview_and_Motivational_Factors

[28] Scott A. Hale. 2012. Net Increase? Cross-Lingual linking in the blogosphere. *Journal of Computer-Mediated Communication* 17, 2 (2012), 135–151.

[29] Aaron Halfaker. 2017. Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect. In *Proceedings of the 13th International Symposium on Open Collaboration* (OpenSym ’17), 19:1–19:9. DOI:https://doi.org/10.1145/3125433.3125475

[30] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2012. The rise and decline of an open collaboration system: How Wikipedia’s reaction to popularity is causing its decline. *American Behavioral Scientist* (2012), 0002764212469365.

[31] Aaron Halfaker, Oliver Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. 2015. User session identification based on strong regularities in inter-activity time. In *WWW*, 410–418.

[32] Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. 2009. A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia. In *WikiSym* (WikiSym ’09), 15:1–15:10. DOI:https://doi.org/10.1145/1641309.1641332

[33] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don’t bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *WikiSym*, 163–172. Retrieved January 5, 2016 from http://dl.acm.org/citation.cfm?id=2038585

[34] Aaron Halfaker and John Riedl. 2012. Bots and cyborgs: Wikipedia’s immune system. *Computer* 45, 3 (2012), 79–82.

[35] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2009. Automatic Quality Assessment of Content Created Collaboratively by Web Communities: A Case Study of Wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL ’09), 295–304. DOI:https://doi.org/10.1145/1555400.1555449

[36] Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on Communities and technologies*, 11–20. Retrieved August 4, 2016 from http://dl.acm.org/citation.cfm?id=1556463

[37] Brent Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 291–300. Retrieved September 21, 2016 from http://dl.acm.org/citation.cfm?id=1753370

[38] Brent Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. *ICWSM* 14, (2014), 197–205.

[39] Benjamin Mako Hill and Aaron Shaw. 2014. Consider the Redirect: A Missing Dimension of Wikipedia Research. In *Proceedings of The International Symposium on Open Collaboration - OpenSym '14*, 1–4. DOI:https://doi.org/10.1145/2641580.2641616

[40] Meiqun Hu, Ee-Peng Lim, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. 2007. Measuring Article Quality in Wikipedia: Models and Evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (CIKM '07), 243–252. DOI:https://doi.org/10.1145/1321440.1321476

[41] Musfira Jilani, Padraig Corcoran, and Michela Bertolotto. 2014. Automated highway tag assessment of OpenStreetMap road networks. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 449–452. Retrieved November 9, 2015 from http://dl.acm.org/citation.cfm?id=2666476

[42] Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at home on the range: Peer production and the urban/rural divide. In *CHI*, 13–25. Retrieved April 24, 2017 from http://dl.acm.org/citation.cfm?id=2858123

[43] Isaac L. Johnson, Subhasree Sengupta, Johannes Schöning, and Brent Hecht. 2016. The Geography and Importance of Localness in Geotagged Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 515–526. DOI:https://doi.org/10.1145/2858036.2858122

[44] Isaac Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at home on the range: Peer production and the urban/rural divide. (2016). Retrieved August 5, 2016 from https://uhdspace.uhasselt.be/dspace/handle/1942/20228

[45] Ah Reum Kang, Jiyoung Woo, Juyong Park, and Huy Kang Kim. 2013. Online game bot detection based on party-play log analysis. *Computers & Mathematics with Applications* 65, 9 (2013), 1384–1395.

[46] Hongwen Kang, Kuansan Wang, David Soukal, Fritz Behr, and Zijian Zheng. 2010. Large-scale bot detection for search engines. In *GROUP*, 501–510.

[47] Nikos Karagiannakis, Giorgos Giannopoulos, Dimitrios Skoutas, and Spiros Athanasiou. 2015. OSMRec Tool for Automatic Recommendation of Categories on Spatial Entities in OpenStreetMap. In *Proceedings of the 9th ACM Conference on Recommender Systems*, 337–338. Retrieved October 27, 2015 from http://dl.acm.org/citation.cfm?id=2796555

[48] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 3819–3828. DOI:https://doi.org/10.1145/2702123.2702520

[49] Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popovi, Mariusz Jaskolski, and David Baker. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat Struct Mol Biol* 18, 10 (September 2011), 1175–1177. DOI:https://doi.org/10.1038/nsmb.2119

[50] Aniket Kittur, Ed Chi, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web* 1, 2 (2007), 19.

[51] Aniket Kittur and Robert E. Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (CSCW '08), 37–46. DOI:https://doi.org/10.1145/1460563.1460572

[52] Marina Kogan, Jennings Anderson, Leysia Palen, Kenneth M. Anderson, and Robert Soden. 2016. Finding the Way to OSM Mapping Practices: Bounding Large Crisis Datasets for Qualitative Investigation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2783–2795. Retrieved May 26, 2016 from http://dl.acm.org/citation.cfm?id=2858371

[53] Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. Community, consensus, coercion, control: cs* w or how policy mediates mass participation. In *Proceedings of the 2007 international ACM conference on Supporting group work*, 167–176. Retrieved September 11, 2016 from http://dl.acm.org/citation.cfm?id=1316648

[54] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. 2015. Measuring Article Quality in Wikipedia Using the Collaboration Network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (ASONAM '15), 464–471. DOI:https://doi.org/10.1145/2808797.2808895

[55] Karim R. Lakhani and Robert G. Wolf. 2003. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. (2003).

[56] Shyong K. Lam, Jawed Karim, and John Riedl. 2010. The effects of group composition on decision quality in a social production community. In *Proceedings of the 16th ACM international conference on Supporting group work*, 55–64. Retrieved September 13, 2016 from http://dl.acm.org/citation.cfm?id=1880083

[57] Shyong Tony K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. 2011. WP: clubhouse?: an exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th international symposium on Wikis and open collaboration*, 1–10.

[58] Ganaele Langlois, Fenwick McKelvey, Greg Elmer, and Kenneth Werbin. 2009. Mapping commercial Web 2.0 worlds: Towards a new critical ontogenesis. *Fibreculture* 14, (2009), 1–14.

[59] Helen Lee. 2012. The role of local food availability in explaining obesity risk among young school-aged children. *Social Science & Medicine* 74, 8 (April 2012), 1193–1203. DOI:https://doi.org/10.1016/j.socscimed.2011.12.036

[60] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. 2014. Reader Preferences and Behavior on Wikipedia. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (HT '14), 88–97. DOI:https://doi.org/10.1145/2631775.2631805

[61] Lawrence Lessig. 1999. *Code and Other Laws of Cyberspace*.

[62] Yilun Lin, Bowen Yu, Andrew Hall, and Brent Hecht. 2017. Problematizing and Addressing the Article-as-Concept Assumption in Wikipedia. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work & Social Computing*. DOI:https://doi.org/10.1145/2998181.2998274

[63] Yu-Wei Lin. 2011. A qualitative enquiry into OpenStreetMap making. *New Review of Hypermedia and Multimedia* 17, 1 (2011), 53–71.

[64] Nedim Lipka and Benno Stein. 2010. Identifying Featured Articles in Wikipedia: Writing Style Matters. In *Proceedings of the 19th International Conference on World Wide Web* (WWW '10), 1147–1148. DOI:https://doi.org/10.1145/1772690.1772847

[65] Ina Ludwig, Angi Voss, and Maike Krause-Traudes. 2011. A Comparison of the Street Networks of Navteq and OSM in Germany. In *Advancing geoinformation science for a changing world*. Springer, 65–84.

[66] Afra Mashhadi, Giovanni Quattrone, and Licia Capra. 2013. Putting Ubiquitous Crowd-sourcing into Context. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (CSCW '13), 611–622. DOI:https://doi.org/10.1145/2441776.2441845

[67] Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies* 67, 9 (September 2009), 716–754. DOI:https://doi.org/10.1016/j.ijhcs.2009.05.004

[68] Amanda Menking, David W. McDonald, and Mark Zachry. 2017. Who wants to read this?: A method for measuring topical representativeness in user generated content systems. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2068–2081.

[69] Peter Mooney and Padraig Corcoran. 2012. The Annotation Process in OpenStreetMap. *Transactions in GIS* 16, 4 (August 2012), 561–579. DOI:https://doi.org/10.1111/j.1467-9671.2012.01306.x

[70] Peter Mooney, Padraig Corcoran, and Adam C. Winstanley. 2010. Towards Quality Metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (GIS '10), 514–517. DOI:https://doi.org/10.1145/1869790.1869875

[71] Michael Muller. 2014. Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (eds.). Springer New York, 25–48. DOI:https://doi.org/10.1007/978-1-4939-0378-8_2

[72] Claudia Müller-Birn, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. 2015. Peer-production system or collaborative ontology engineering effort: What is Wikidata? In *OpenSym*, 20. Retrieved June 27, 2016 from http://dl.acm.org/citation.cfm?id=2789836

[73] Oded Nov. 2007. What motivates wikipedians? *Communications of the ACM* 50, 11 (2007), 60–64.

[74] M. Over, A. Schilling, S. Neubauer, and A. Zipf. 2010. Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany. *Computers, Environment and Urban Systems* 34, 6 (November 2010), 496–507. DOI:https://doi.org/10.1016/j.compenvurbsys.2010.05.001

[75] Leysia Palen, Robert Soden, T. Jennings Anderson, and Mario Barrenechea. 2015. Success & scale in a data-producing organization: the socio-technical evolution of OpenStreetMap in response to humanitarian events. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 4113–4122. Retrieved September 15, 2016 from http://dl.acm.org/citation.cfm?id=2702294

[76] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *GROUP* (GROUP '09), 51–60. DOI:https://doi.org/10.1145/1531674.1531682

[77] S. Hassany Pazoky, F. Karimipour, and F. Hakimpour. 2015. YOU DESCRIBE IT, I WILL NAME IT: AN APPROACH TO ALLEVIATE THE EFFECT OF USERS'SEMANTICS IN ASSIGNING TAGS TO FEATURES IN VGI. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* 40, (2015). Retrieved October 26, 2015 from http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-3-W3/39/2015/isprsarchives-XL-3-W3-39-2015.pdf

[78] Alessandro Piscopo, Chris Phethean, and Elena Simperl. 2017. What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *Social Informatics* (Lecture Notes in Computer Science), 305–322. DOI:https://doi.org/10.1007/978-3-319-67217-5_19

[79] Martin Potthast, Benno Stein, and Teresa Holfeld. 2010. Overview of the 1st International Competition on Wikipedia Vandalism Detection. In *CLEF (Notebook Papers/LABs/Workshops)*.

[80] Giovanni Quattrone, Martin Dittus, and Licia Capra. 2017. Work Always in Progress: Analysing Maintenance Practices in Spatial Crowd-sourced Datasets. Retrieved March 13, 2017 from http://www0.cs.ucl.ac.uk/staff/l.capra/publications/cscw2017gio.pdf

[81] Sapna Maheshwari BuzzFeed News Reporter. 19 Facts That Show Just How Massive Walmart Really Is. *BuzzFeed*. Retrieved August 8, 2016 from http://www.buzzfeed.com/sapna/19-facts-that-show-just-how-massive-walmart-really-is

[82] Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. 2017. Building automated vandalism detection tools for Wikidata. In *WWW*, 1647–1654.

[83] C. Sarasua, A. Checco, G. Demartini, D. Difallah, M. Feldman, and L. Pintscher. The Evolution of Power and Standard Wikidata Editors - Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits.

[84] Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. 2006. Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 181–190. Retrieved September 19, 2016 from http://dl.acm.org/citation.cfm?id=1180904

[85] Shilad W. Sen, Heather Ford, David R. Musicant, Mark Graham, Oliver SB Keyes, and Brent Hecht. 2015. Barriers to the localness of volunteered geographic information. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 197–206. Retrieved August 4, 2016 from http://dl.acm.org/citation.cfm?id=2702170

[86] Thomas Steiner. 2014. Bots vs. wikipedians, anons vs. logged-ins (redux): A global study of edit activity on wikipedia and wikidata. In *OpenSym*, 25. Retrieved June 24, 2016 from http://dl.acm.org/citation.cfm?id=2641613

[87] Monica Stephens. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal* 78, 6 (2013), 981–996.

[88] Besiki Stvilia, Abdullah Al-Faraj, and Yong Jeong Yi. 2009. Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research* 31, 4 (December 2009), 232–239. DOI:https://doi.org/10.1016/j.lisr.2009.07.005

[89] Pang-Ning Tan and Vipin Kumar. 2004. Discovery of web robot sessions based on their navigational patterns. In *Intelligent Technologies for Information Analysis*. Springer, 193–222.

[90] Ajay Thampi. 2019. *A fast, offline reverse geocoder in Python. Contribute to thampiman/reverse-geocoder development by creating an account on GitHub*. Retrieved April 4, 2019 from https://github.com/thampiman/reverse-geocoder

[91] Ruck Thawonmas, Yoshitaka Kashifuji, and Kuan-Ta Chen. 2008. Detection of MMORPG bots based on behavior analysis. In *Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*, 91–94.

[92] Milena Tsvetkova, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri. 2016. Even Good Bots Fight. *arXiv preprint arXiv:1609.04285* (2016).

[93] Arnaud Vandecasteele and Rodolphe Devillers. 2015. Improving Volunteered Geographic Information Quality Using a Tag Recommender System: The Case of OpenStreetMap. In *OpenStreetMap in GIScience*, Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney and Marco Helbich (eds.). Springer International Publishing, 59–80. Retrieved March 23, 2015 from http://link.springer.com/chapter/10.1007/978-3-319-14280-7_4

[94] Maja van der Velden. 2013. Decentering Design: Wikipedia and Indigenous Knowledge. *International Journal of Human-Computer Interaction* 29, 4 (April 2013), 308–316. DOI:https://doi.org/10.1080/10447318.2013.765768

[95] Morten Warncke-Wang, Vivek Ranjan, Loren Terveen, and Brent Hecht. 2015. Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. In *ICWSM*. Retrieved September 16, 2016 from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10591

[96] Yanxiang Xu and Tiejian Luo. 2011. Measuring article quality in Wikipedia: Lexical clue model. In *2011 3rd Symposium on Web Society*, 141–146. DOI:https://doi.org/10.1109/SWS.2011.6101286

[97] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying Semantic Edit Intentions from Revisions in Wikipedia. In *EMNLP 2017*, 2000–2010. Retrieved July 7, 2018 from https://www.aclweb.org/anthology/D17-1213

[98] Hongyu Zhang and Jacek Malczewski. 2017. Quality Evaluation of Volunteered Geographic Information: The Case of OpenStreetMap. In *Volunteered Geographic Information and the Future of Geospatial Data*. IGI Global, 19–46.

[99] Dennis Zielstra, Hartwig H. Hochmair, and Pascal Neis. 2013. Assessing the effect of data imports on the completeness of OpenStreetMap–a United States case study. *Transactions in GIS* 17, 3 (2013), 315–334.

[100] Dennis Zielstra and Alexander Zipf. 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. In *13th AGILE international conference on geographic information science*. Retrieved January 5, 2016 from http://agile2010.dsi.uminho.pt/pen/shortpapers_pdf/142_doc.pdf

[101] 2015. 20 Interesting flickr Stats and Facts | By the Numbers. *DMR*. Retrieved May 20, 2019 from https://expandedramblings.com/index.php/flickr-stats/

[102] 2017. Wikipedia:Bot Approvals Group. *Wikipedia*. Retrieved January 20, 2018 from https://en.wikipedia.org/w/index.php?title=Wikipedia:Bot_Approvals_Group&oldid=807843217

[103] 2017. Wikipedia:History of Wikipedia bots. *Wikipedia*. Retrieved January 20, 2018 from https://en.wikipedia.org/w/index.php?title=Wikipedia:History_of_Wikipedia_bots&oldid=812914046

[104] 2017. Wikipedia:Five pillars. *Wikipedia*. Retrieved February 28, 2018 from https://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=815197819

[105] 2018. Wikipedia:Bot policy. *Wikipedia*. Retrieved January 20, 2018 from https://en.wikipedia.org/w/index.php?title=Wikipedia:Bot_policy&oldid=820435660
[106] 2018. Coding (social sciences). *Wikipedia*. Retrieved July 7, 2018 from https://en.wikipedia.org/w/index.php?title=Coding_(social_sciences)&oldid=834193623
[107] 2018. Wikipedia:AutoWikiBrowser. *Wikipedia*. Retrieved July 8, 2018 from https://en.wikipedia.org/w/index.php?title=Wikipedia:AutoWikiBrowser&oldid=840931199
[108] 2019. Wikipedia:The Free Encyclopedia. *Wikipedia*. Retrieved May 21, 2019 from https://en.wikipedia.org/w/index.php?title=Wikipedia:The_Free_Encyclopedia&oldid=878091811
[109] 2019. Wikipedia:Content assessment. *Wikipedia*. Retrieved April 4, 2019 from https://en.wikipedia.org/w/index.php?title=Wikipedia:Content_assessment&oldid=886342369
[110] 2019. Facebook, Axios And NBC Paid This Guy To Whitewash Wikipedia Pages. *HuffPost Canada*. Retrieved May 21, 2019 from https://www.huffpost.com/entry/wikipedia-paid-editing-pr-facebook-nbc-axios_n_5c63321be4b03de942967225
[111] 2019. History of free and open-source software. *Wikipedia*. Retrieved May 20, 2019 from https://en.wikipedia.org/w/index.php?title=History_of_free_and_open-source_software&oldid=893525939
[112] 2019. Wisdom of the crowd. *Wikipedia*. Retrieved May 21, 2019 from https://en.wikipedia.org/w/index.php?title=Wisdom_of_the_crowd&oldid=897664237
[113] Alexa Top 500 Global Sites. Retrieved January 24, 2018 from https://www.alexa.com/topsites
[114] Wikipedia:Be bold - Wikipedia, the free encyclopedia. Retrieved August 7, 2016 from https://en.wikipedia.org/wiki/Wikipedia:Be_bold
[115] Good practice - OpenStreetMap Wiki. Retrieved August 7, 2016 from http://wiki.openstreetmap.org/wiki/Good_practice
[116] Official Google Blog: Introducing the Knowledge Graph: things, not strings. Retrieved August 8, 2016 from https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
[117] Humanitarian OpenStreetMap Team. Retrieved August 7, 2016 from https://hotosm.org/
[118] Reasonator. Retrieved August 7, 2016 from https://tools.wmflabs.org/reasonator/
[119] Category:Templates using data from Wikidata - Wikipedia, the free encyclopedia. Retrieved August 8, 2016 from https://en.wikipedia.org/wiki/Category:Templates_using_data_from_Wikidata
[120] About Jimmy | Jimmy Wales. Retrieved May 21, 2019 from http://jimmywales.com/about-jimmy/
[121] Wikipedia Gets 9 Out of 10 Health Ailments Wrong. *Time*. Retrieved May 21, 2019 from http://time.com/118904/study-dont-trust-wikipedia-when-it-comes-to-your-health/
[122] Semi-colon value separator - OpenStreetMap Wiki. Retrieved December 5, 2016 from http://wiki.openstreetmap.org/wiki/Semi-colon_value_separator
[123] Wikidata. Retrieved August 7, 2016 from https://www.wikidata.org/wiki/Wikidata:Main_Page
[124] Who Writes Wikipedia? (Aaron Swartz's Raw Thought). Retrieved January 18, 2018 from http://www.aaronsw.com/weblog/whowriteswikipedia
[125] Research:Measuring edit productivity - Meta. Retrieved April 17, 2018 from https://meta.wikimedia.org/wiki/Research:Measuring_edit_productivity
[126] TIGER - OpenStreetMap Wiki. Retrieved January 20, 2018 from https://wiki.openstreetmap.org/wiki/TIGER
[127] TIGER fixup - OpenStreetMap Wiki. Retrieved January 20, 2018 from https://wiki.openstreetmap.org/wiki/TIGER_fixup
[128] Import/Guidelines - OpenStreetMap Wiki. Retrieved January 20, 2018 from https://wiki.openstreetmap.org/wiki/Import/Guidelines
[129] ORES - MediaWiki. Retrieved July 5, 2018 from https://www.mediawiki.org/wiki/ORES
[130] How did you contribute to OpenStreetMap ? Retrieved September 18, 2016 from http://hdyc.neis-one.org/
[131] About | Humanitarian OpenStreetMap Team. Retrieved August 7, 2016 from https://hotosm.org/about
[132] HOT Tasking Manager -. Retrieved August 7, 2016 from http://tasks.hotosm.org/
[133] LearnOSM. Retrieved September 15, 2016 from http://learnosm.org/en/coordination/remote/

[134] Wikipedia:WikiProject Council/Guide/WikiProject - Wikipedia. Retrieved December 31, 2016 from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Guide/WikiProject#Violating_policies
[135] Episode 708: Bitcoin Divided. *NPR.org*. Retrieved September 13, 2016 from http://www.npr.org/sections/money/2016/06/29/484029238/episode-708-bitcoin-divided
[136] Import/Guidelines - OpenStreetMap Wiki. Retrieved January 6, 2017 from http://wiki.openstreetmap.org/wiki/Import/Guidelines
[137] Tag:amenity=fast_food - OpenStreetMap Wiki. Retrieved May 28, 2018 from https://wiki.openstreetmap.org/wiki/Tag:amenity=fast_food
[138] Proposal process - OpenStreetMap Wiki. Retrieved June 10, 2018 from https://wiki.openstreetmap.org/wiki/Proposal_process
[139] Fast food industry market share in the U.S. 2015 | Statistic. Retrieved August 7, 2016 from http://www.statista.com/statistics/196611/market-share-of-fast-food-restaurant-corporations-in-the-us/
[140] Index of /full-history-extracts/latest/continents. Retrieved June 2, 2018 from http://osm.personalwerk.de/full-history-extracts/latest/continents/
[141] Tag:shop=supermarket - OpenStreetMap Wiki. Retrieved June 3, 2018 from https://wiki.openstreetmap.org/wiki/Tag:shop%3Dsupermarket
[142] Wikidata:Requests for comment/Conflict of Interest - Wikidata. Retrieved March 6, 2018 from https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Conflict_of_Interest
[143] Wikidata:Bots - Wikidata. Retrieved July 4, 2018 from https://www.wikidata.org/wiki/Wikidata:Bots
[144] Wikidata:Glossary - Wikidata. Retrieved July 1, 2018 from https://www.wikidata.org/wiki/Wikidata:Glossary
[145] Manual:Tags - MediaWiki. Retrieved July 9, 2018 from https://www.mediawiki.org/wiki/Manual:Tags
[146] Proposed features/changeset tags - OpenStreetMap Wiki. Retrieved September 3, 2018 from https://wiki.openstreetmap.org/wiki/Proposed_features/changeset_tags
[147] Wikidata:Item quality - Wikidata. Retrieved April 4, 2019 from https://www.wikidata.org/wiki/Wikidata:Item_quality
[148] Wikimedia Downloads. Retrieved April 4, 2019 from https://dumps.wikimedia.org/
[149] instance of. Retrieved April 4, 2019 from https://www.wikidata.org/wiki/Property:P31
[150] human. Retrieved April 4, 2019 from https://www.wikidata.org/wiki/Q5
[151] sex or gender. Retrieved April 4, 2019 from https://www.wikidata.org/wiki/Property:P21
[152] male. Retrieved April 4, 2019 from https://www.wikidata.org/wiki/Q6581097
[153] female. Retrieved April 4, 2019 from https://www.wikidata.org/wiki/Q6581072
[154] Wikidata Query Service. Retrieved April 4, 2019 from https://query.wikidata.org/
[155] coordinate location. Retrieved April 4, 2019 from https://www.wikidata.org/wiki/Property:P625
[156] Data Access - Urban Rural Classification Scheme for Counties. Retrieved August 7, 2016 from http://www.cdc.gov/nchs/data_access/urban_rural.htm
[157] Solve Puzzles for Science | Foldit. Retrieved April 7, 2019 from https://fold.it/portal/
[158] Wikidata - The Game. Retrieved April 7, 2019 from https://tools.wmflabs.org/wikidata-game/
[159] The new alchemy: turning online harassment into Wikipedia articles on women scientists – Wikimedia Blog. Retrieved May 31, 2019 from https://blog.wikimedia.org/2016/03/08/alchemy-turning-harassment-into-women-scientists/
[160] QuickStatements - Meta. Retrieved April 4, 2019 from https://meta.wikimedia.org/wiki/QuickStatements
[161] PetScan - Meta. Retrieved April 4, 2019 from https://meta.wikimedia.org/wiki/PetScan

9 Appendix

9.1 Study 2 Sampling Contingent Metadata

We sampled Contingent keys (specific samples discussed below), and consulted each business's website store locator to determine whether the attribute represented by the key was *present* for a specific business location. If the attribute was present, we checked if metadata indicating that attribute existed. For example, we checked if a particular McDonalds had a drive-through and -- if it did -- if it then had metadata to represent that attribute.

We checked for attribute presence since tags may sometimes indicate a non-existent attribute. For example, the OSM wiki states that it is valid to explicitly indicate that a fast food restaurant does not have a drive through: "drive_through=no". It's important to note that in our analysis of "missed tagging opportunities", we did not consider it to be a "missed opportunity" if a tag for a non-existent attribute was unapplied. This is because this situation appears fairly rare in practice and tag application in this scenario is not necessarily described in the wiki. Thus, we chose a conservative interpretation of OSM's ontology when defining "missed opportunities".

To sample Contingent metadata, we manually selected the key "internet_access" for McDonald's and Starbucks and the key "drive_through" for McDonald's, Starbucks, and Walgreens. We chose these businesses and keys because the businesses were among the United States' leaders in their market categories and because the two keys are important attributes for potential customers of these businesses. Chain businesses are important fixtures in low-income areas, and internet penetration also suffers as well. Whether a given McDonald's has internet access can thus, be very important. When available, these types of metadata are also used by applications such as Google Maps and Yelp – indicating their importance and demand. Given that we looked at prominent businesses and broadly important attributes, we felt the sampled metadata *should* be among the Contingent metadata that is applied the most. We analyzed 60 instances per business for 180 total instances and 300 *potential* business-key-value triples. Out of the 300 triples, 245 of the attributes represented by those triples actually existed in the real-world.

Given our choice of Contingent metadata sampled, the fact that Contingent metadata is laborious to apply, and the fact that even less Contingent metadata was applied in scenarios where it was appropriate relative to Universal-Varying metadata, we had confidence that our sampling approach provided a reasonable "best case" proxy for the degree to which all Contingent metadata is applied when it is applicable. And further, we were confident that that degree of application was very low.

9.2 Study 3 Details of Data Preparation

9.2.1 Extracting Revision Data and Sessionization

Building our prediction model required historical Wikidata *revision* data. When a bot or human edits Wikipedia, they produce a revision (we use the word "edit" interchangeably) which contains the changes made as well as metadata such as namespace, page title, revision id, user id, username (in the case of anonymous user edits, IP address), timestamp, and comment. A given namespace corresponds to a distinct type of Wikidata content. Namespaces exist for items, properties (which represent attributes of items), general community discussion, policy development, and more. To access revision data for all pages throughout Wikidata's history, we used the "stub-meta-history" xml dump files from May 1, 2017 from https://dumps.wikimedia.org. There were over 476 million revisions, predominantly from namespace 0 (item entities). We extracted this dump data using the *mwxml* utility[30].

We next "session"-ized all revision data. As mentioned previously, prior work [19,31] has defined a human editing "session" as a period in which a human makes edits without taking large breaks. This work identified implicit behavioral signals in editing sessions that have predictive power to distinguish bots from human contributors. Specifically, people generally edit with predictable inter-edit times. Bots typically edit with significantly smaller inter-edit times. We incorporated these and other implicit session characteristics in our model.

To session-ize all the revisions from logged-in users for model training and testing (ground truth labels are discussed in section A.2), we applied the *mwsessions* utility[31] with default parameters. As recommended by prior research [19,31], the default cutoff of greater than one hour of inactivity was used to determine the end of one session and the beginning of another by the same user. This produced 4,990,652 logged-in user sessions corresponding to 177,134 unique user ids. The median number of sessions per user id was 1, while the mean was 28.17.

We then randomly sampled 100,000 sessions for model training and 100,000 for testing. We filtered out sessions from the test set that were also in the training set to ensure no overlap. We also filtered out sessions that did not contain at least one edit to a *Property*- or an *Item-page* – the main content of Wikidata. Further, since we are concerned mainly with large sequences of potential bot activity, we filtered out sessions with less than three edits. After these filtering steps, we had 36,252 training and 35,191 testing sessions.

To test our classifier model's performance on *anonymous* user bot activities, we gathered a similar sample of anonymous activities. We session-ized all of Wikidata's 2,114,900 anonymous revisions into 677,050

[30] http://pythonhosted.org/mwxml/

[31] http://pythonhosted.org/mwsessions/

sessions, session-izing based on IP address. After following the same filtering steps that we used for training and testing datasets, we had 110,288 sessions for testing the detection of anonymous bot editing.

Recall that anonymous contributor edits are tied to an IP address; therefore, one cannot assume that a given IP address will correspond to only one user over long periods of time. However, an IP address is unlikely to change within a continuous activity session, so it serves as a durable-enough identifier for our purposes [19].

9.2.2 Ground Truth Bot Account Data

To obtain labeled training and testing data for registered user edits, we compiled a list of all currently and formerly approved bot accounts. We obtained this data by looking for "bot flags" in two publicly accessible tables called "user_groups" and "user_former_groups"[32]. We then merged this bot account data with the registered user revision training and testing data based on the user id fields in each dataset[33]. All accounts not labelled as bots were considered human accounts unless their username ended in "bot" (insensitive to case), in which case they were omitted from the training and testing datasets. This was a conservative approach in that we wanted to train/test our model on the least noisy data possible in order to get the best predictions. In our training set, 34,797 were labeled as human sessions and 1455 as bot. In our testing set, 33,775 were labeled as human sessions and 1416 as bot. Since there are more bot edits than human edits in Wikidata, one might wonder why our dataset contains so many more human sessions than bot sessions. This is because bot edit sessions tend to produce more edits than human sessions.

[32] https://quarry.wmflabs.org/query/19668

[33] one "bot flagged" user account was not considered to be a bot account when merging because we had reason to believe that it had not been used exclusively for bot editing.

9.3 Study 3 Bot Prediction Model Features

Activity Pattern Features	
Feature	**Description (if needed)**
Inter-edit time mean	
Inter-edit time standard deviation	
# of edits	
(Session) length in seconds	
# of edits with inter-edit time < 5 seconds	
# of edits with inter-edit time >= 5 seconds and <= 20 seconds	
# of edits with inter-edit time > 20 seconds	
# of Namespace 0 edits	The main "item" namespace.
# of Namespace 1 edits	The "talk" namespace of Namespace 0.
# of Namespace 2 edits	Namespace for contributors to provide information about themselves.
# of Namespace 3 edits	The "talk" namespace of Namespace 2.
# of Namespace 4 edits	Namespace for specification of community rules/policies.
# of Namespace 5 edits	The "talk" namespace of Namespace 4.
# of Namespace 120 edits	Where properties are defined.
# of Namespace 121 edits	The "talk" namespace of Namespace 120.
Revision Comment-Based Features	
Feature	**Description (if needed)**
# of claim creations	
# of unique claims changed	This was measured by determining the number of properties changed.
# of distinct pages edited (across all namespaces)	
# of distinct edit types	We parsed the "structured" part of comments to find different revision types. For example, "wbsetdescription" indicates a description is being set. "wbcreateclaim-create" indicates the creation of a new claim.
# of bot generated generic comments	Certain boilerplate revision comments can only be created via the Wikidata editing API (e.g. "/* wbeditentity-update:0| */"). These comments likely come from bots.
Word ending in "bot" (case insensitive) in revision comment	
# of sitelinks changed	
# of aliases changed	
# of labels changed	
# of descriptions changed	
Same exact edit occurred more than once (possible edit war)	
# of revisions removing content	
# of revisions modifying content	
# of edits with inter-edit time < 2 seconds	

Table 5.3: Bot Prediction Model Features

9.4 Study 3 Detailed Anonymous Contributor Qualitative Coding Results

We break down the high-level results of our anonymous contributor qualitative coding in each stratum. Results are summarized in Table 5.3.

Recall: 0.0-0.1 stratum. As seen in Table 5.3, all 15 sessions sampled (this stratum had 15 sessions total) were clearly from bots. Most sessions (11) in this stratum contained revisions with boilerplate bot comments (coded as BBC) that made bot recognition straightforward. Some also contained quickly occurring edits (that were coded as FE).

Table 5.3: Overview of the Results of the Anonymous Contributor Qualitative Coding Summary.

This table organizes results of our coding by "recall stratum" (defined in the text). The fundamental pattern is that low-recall strata include all or nearly all bot sessions, high-recall strata include all or nearly all human sessions, and that the intermediate strata contain a mix of the two. The coding results also provide explanations of the makeup of the sessions in each stratum, specified by the codes defined in Table 5.2 of the chapter. This analysis confirms that the model is effective at identifying edit sessions made by bots.

Recall Stratum	# Bots Sessions	# Human Sessions	# Unknown Sessions	Notes (See Table 5.2 for code descriptions)
0.0-0.1	15	0	0	11/15 sessions were coded BBC. Lots of FE sessions
0.1-0.2	18	0	2	9/20 sessions were EBC or FE. Quite a few BBC sessions as well
0.2-0.3	16	4	0	Quite a few EBC or FE sessions again. A few obvious human sessions including edit war(s)
0.3-0.4	11	7	2	Most bot sessions were BBC or FE sessions
0.4-0.5	6	4	10	Half of the sessions were coded as unknown
0.5-0.6	7	6	7	A few bot sessions were SIM. One session simply had blank comments
0.6-0.7	6	8	6	More edits appeared human than bot. A number were still hard to identify as either and coded as unknown
0.7-0.8	2	16	2	Nearly all sessions were obviously human. A couple sessions appeared to be possible vandalism
0.8-0.9	1	18	1	19/20 sessions were obviously human in behavior. One contained "(BOT)" in beginning of the unstructured part of its comment
0.9+	0	20	0	All sessions were clearly human

Recall: 0.1-0.2 stratum. 18 sessions were clearly from bots. 2 were hard to identify and coded as unknown. Of the bot sessions, half were coded as EBC (they contained "bot" in their revision comments) or FE. As with the previous stratum, a number of sessions were coded as BBC as well.

Recall: 0.2-0.3 stratum. 16 sessions were clearly from bots and 4 were clearly from humans. As with previous sessions, several EBC or FE sessions occurred. One of the human sessions appeared to be part of

edit war(s) -- the contributor was restoring content on multiple pages. For example, related to the item for Kim Kardashian, the contributor stated in the comment of a restoring edit: "...I definitely don't understand why [you] removed Kendall and Kylie, they're her sisters".

Recall: 0.3-0.4 stratum. As seen in Table 5.3, at this point, a fair number of human sessions began appearing (7 human, 11 bot, 2 unknown). Bot sessions were primarily BBC or FE.

Recall: 0.4-0.5 stratum. Often no clear indicators existed to code sessions as either bot or human in this stratum, and the result was a large increase in sessions coded as unknown (10). There were 4 human and 6 bot sessions.

Recall: 0.5-0.6 and 0.6-0.7 strata. At this point, model precision had decreased significantly. The number of unknown sessions had decreased from the previous stratum while human sessions increased. 13 sessions (cumulatively across the two stratum) were from bots, 14 from humans, and 13 were coded as unknown.

Recall: 0.7-0.8 stratum. This represented a clear turn towards human sessions. As seen in Table 5.3, 16 sessions were clearly from humans, 2 from bots, and 2 were coded as unknown. One human editor appeared to be vandalizing pages, providing comments such as "cgyjcfgmfhuk".

Recall: 0.8-0.9 and 0.9+ strata. In the remaining two strata, all sessions were clearly from human editors, except for one in the 0.8-0.9 stratum which was coded as EBC.

9.5 Study 4 Additional Details of Determining Contribution Strategy Types of Revisions

We detail how we portioned our dataset based on contribution strategy type.

9.5.1 Identifying Revisions from Bots

To identify revisions coming from *bots*, we obtained the user ids of all current and former community-approved bots[34][35] and flagged any revisions in our dataset that had those ids. Additionally, we identified unapproved bots running on internal Wikimedia servers based on their IP address and flagged the corresponding revisions in our dataset as well.

9.5.2 Identifying Revisions from Semi-Automated Tools

Some semi-automated tools such as QuickStatements [160] and PetScan [161] often leave a clear trace in revision comments, but there is no complete list of comment traces that we could use to identify all semi-automated tools. To build such a corpus, we first parsed the revision comments of all item revisions[36] and computed a list of 100 of the most common words found within them. Semi-automated tools use an exact substring to identify their activity and such extreme consistency should show up at the top of a simple word frequency measurement. I, along with a collaborator, then compared this list with known Wikidata editing tools. We then used these words to create regular expressions to flag semi-automated tool edits from registered users based on their revision comments. Further, informal exploration let us identify a few more regular expressions that also match semi-automated tools. To further ensure the validity of our approach, we posted all our regular expressions on a Wikimedia Meta page[37] and asked Wikidata contributors to review our list via the Wikidata mailing list. We made changes based on contributor feedback and spot checked to ensure that our regular expressions correctly flagged tool edits. Finally, we also flagged registered user edits based on a list of revision comment regular expressions that had been used in previous work to denote semi-automated tool edits[38] [83] and a list of "change tags" indicative of tool edits.

9.5.3 Identifying Revisions from Manual Effort

All remaining edits not previously flagged were considered to come via direct manipulation of Wikidata's user interface. There were two types: 1) registered manual contributions and 2) anonymous manual

[34] https://quarry.wmflabs.org/query/19668

[35] We removed one user id from this list after the owner of the id indicated the account was also being used for tool edits.

[36] We also included property revisions. Properties define the structured data that can be applied to items. The vast majority of edits are to items.

[37] https://lists.wikimedia.org/pipermail/wikidata/2017-September/011197.html

[38] https://meta.wikimedia.org/wiki/Research:Understanding_Wikidata%27s_Value/final_semi-automated_tool_edit_indicators

contributions. Anonymous edits were those associated with an IP address rather than a user id of a registered contributor.

9.6 Study 4 Additional Tables and Figures

Table 6.1: Content Edits Sampled from Each Contribution Strategy and Period

	Bots	Reg. Manual Contributors	Anon. Manual Contributors	Tools
May 2013- April 2014	98979	74524	48935	99767
May 2014- April 2015	98865	66843	68548	97366
May 2015- April 2016	98686	70903	83208	97175
May 2016- April 2017	99180	74752	79121	98624

Table 6.2: Content Edits Sampled for Male and Female Items

	Bots	Reg. Manual Contributors	Anon. Manual Contributors	Tools
May 2013- April 2014	99548	80112	47022	99962
May 2014- April 2015	99614	78274	77259	99458
May 2015- April 2016	99706	81885	95343	99545
May 2016- April 2017	99851	84343	89806	99286

Table 6.3: Content Edits Sampled for Global North and Global South Items

	Bots	Reg. Manual Contributors	Anon. Manual Contributors	Tools
May 2013- April 2014	98561	77760	27216	86294
May 2014- April 2015	96894	68125	22873	97519
98198	98198	67963	24821	96462
May 2016- April 2017	99028	76829	36662	99553

Table 6.4: Content Edits Sampled for Urban and Rural Items

	Bots	Reg. Manual Contributors	Anon. Manual Contributors	Tools
May 2013- April 2014	97459	76201	2740	9515
May 2014- April 2015	95774	45659	1914	99500
May 2015- April 2016	98978	67292	2187	97238
May 2016- April 2017	94798	77306	2181	99705