Examining Wikipedia With a Broader Lens: Quantifying the Value of Wikipedia's Relationships with Other Large-Scale Online Communities

Nicholas Vincent
PSA Research Group
Northwestern University
nickvincent@u.northwestern.edu

Isaac Johnson
PSA Research Group
Northwestern University
isaacj@u.northwestern.edu

Brent Hecht
PSA Research Group
Northwestern University
bhecht@northwestern.edu

ABSTRACT
The extensive Wikipedia literature has largely considered Wikipedia in isolation, outside of the context of its broader Internet ecosystem. Very recent research has demonstrated the significance of this limitation, identifying critical relationships between Google and Wikipedia that are highly relevant to many areas of Wikipedia-based research and practice. This paper extends this recent research beyond search engines to examine Wikipedia's relationships with large-scale online communities, Stack Overflow and Reddit in particular. We find evidence of consequential, albeit unidirectional relationships. Wikipedia provides substantial value to both communities, with Wikipedia content increasing visitation, engagement, and revenue, but we find little evidence that these websites contribute to Wikipedia in return. Overall, these findings highlight important connections between Wikipedia and its broader ecosystem that should be considered by researchers studying Wikipedia. Critically, our results also emphasize the key role that volunteer-created Wikipedia content plays in improving other websites, even contributing to revenue generation.

Author Keywords
Wikipedia, peer production, Stack Overflow, Reddit, online communities

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION
Over the past decade, Wikipedia has been a subject of tremendous interest in the computing literature. Researchers have used Wikipedia to understand online collaboration dynamics (e.g. [30,32,60,61]), to evaluate volunteer recruitment and retention strategies (e.g. [17,18,59,62]), and even to train many state-of-the-art artificial intelligence algorithms [12,13,25,26,36,58]. Indeed, Wikipedia has likely become one of the most important datasets and research environments in modern computing [25,28,36,38].

However, a major limitation of the vast majority of the Wikipedia literature is that it considers Wikipedia in isolation, outside the context of its broader online ecosystem. The importance of this limitation was made quite salient in recent research that suggested that Wikipedia's relationships with other websites are tremendously significant [15,35,40,57]. For instance, this work has shown that Wikipedia's relationships with other websites are important factors in the peer production process and, consequently, have impacts on key variables of interest such as content quality, reader demand, and contribution patterns [35,57].

Perhaps more importantly, however, this recent research has additionally suggested that the reverse is also true: Wikipedia content appears to play a substantially more important role in the Internet ecosystem than anticipated, with other websites having critical dependencies on Wikipedia content. In particular, McMahon et al. [35] showed that the click-through rates of Google SERPs (search engine results pages) drop dramatically when Wikipedia links are removed, suggesting that Google is quite reliant on Wikipedia to satisfy user information needs. Among other implications, this means that the Wikipedia peer production processes studied in the social computing literature likely have a substantial – and largely unstudied – impact on other websites. McMahon et al.'s results also raised important questions related to the revenue being generated by for-profit institutions using volunteer-created Wikipedia content, especially in light of Wikipedia's limited donation income.

Recognizing the importance of understanding Wikipedia's relationships with its broader online ecosystem, the Wikimedia Foundation (the operator of Wikipedia) has called for more research on these relationships as a central part of its "New Research Agenda" for Wikipedia [51]. The goal of the research presented in this paper is to help address this call. More specifically, we seek to extend McMahon et al.'s work on Wikipedia's relationship with Google into a

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ACM ISBN 978-1-4503-5620-6/18/04…$15.00
https://doi.org/10.1145/3173574.3174140

different and important area of the online ecosystem: other large-scale online communities.

In line with guidance in the social computing community to increase robustness through the use of multi-community analyses [48,53], we examine Wikipedia's relationship with two distinct large-scale online communities: Stack Overflow (SO) and Reddit. Users post links to Wikipedia on both SO and Reddit, facilitating an important, and potentially bidirectional, relationship with Wikipedia.

Following the high-level structure of McMahon et al [35], this paper asks two overarching research questions about Wikipedia's relationship with SO and Reddit*:*

RQ1: What value is Wikipedia providing to other large-scale online communities like Stack Overflow and Reddit? (i.e. Does Wikipedia content increase community engagement and/or company revenue?)

RQ2: What value do these large-scale online communities provide to Wikipedia? (i.e. Are they contributing page views? Editors?)

We additionally take an important step beyond McMahon et al. and investigate how the *quality* of Wikipedia articles affects the relationships examined in RQ1 and RQ2. In other words, we look at the association between the quality of articles on Wikipedia and the value that Wikipedia provides to external entities, and vice versa.

We address our RQs using a combined framework of associative and causal analyses that allows us to estimate Wikipedia's relationships with SO and Reddit under a range of conditions. For instance, at the upper bound of this range, our associative analyses allow us to ask, "How much value would be lost from SO and Reddit if posts containing Wikipedia links never appeared?" Similarly, to estimate a lower bound on the value Wikipedia could be providing to SO and Reddit, we use causal analysis to examine the counterfactual scenario, "What if the posts containing Wikipedia links remained unchanged content-wise, but instead had a link to a site other than Wikipedia?"

The results of our analyses indicate that Wikipedia creates a large amount of value for SO and Reddit, even in our lower-bound estimates. More specifically, we observe that posts containing Wikipedia links on both sites are exceptionally valuable posts, with engagement metrics like user-voted scores much higher than posts that do not contain Wikipedia links (often by a factor of at least 2x, and sometimes as much as 4x-5x). This results in an estimated increase in revenue on the order of $100K/year for both sites.

However, we find little evidence that posts with Wikipedia links provide direct value to the Wikipedia community. We were able to replicate work that showed that Wikipedia posts on the popular Reddit community "TIL" ("Today I Learned") were responsible for a large spike in viewership. However, our results suggest that this large effect does not generalize beyond the "TIL" community or beyond Reddit. Moreover, we see negligible increases in Wikipedia edits and editor signup despite the large volume of links posted on both sites. Through a smaller-scale qualitative analysis, we find evidence that suggests that the "paradox of re-use" [51] may be playing a role here: Wikipedia's permissive content licenses make it easy for Reddit and SO users to directly include the Wikipedia content in their posts, which could be mitigating the benefits of Wikipedia links in terms of traffic to Wikipedia and new Wikipedia edits/editors.

As we discuss below, these results have important implications for a number of constituencies. For companies that rely on Wikipedia content, our findings highlight the value (including both engagement and revenue) created by Wikipedia's free and volunteer-created content. For Wikipedia and its editors, RQ1's results further demonstrate the critical role the Wikipedia community plays in the broader online ecosystem outside of Wikipedia. However, our RQ2 results present more challenging implications for the Wikipedia community: these results highlight the need for further research on solutions to the paradox of re-use.

BACKGROUND

Prior to presenting related work, we first provide high-level context about Reddit, SO and how Wikipedia content appears on these sites. For both Reddit and SO, there is a class of posts that is highly-dependent (if not entirely reliant) on Wikipedia content (i.e. the posts could not exist without Wikipedia). Additionally, users can "upvote" and "downvote" content on Reddit and SO, which gives it a "score" (# of upvotes – # of downvotes).

Reddit and Wikipedia

Reddit is a large-scale online community that allows users to share links to external content (e.g. news articles, images, videos), post original text content (e.g. questions, opinions), and discuss this content through comments. As of July 2017, Reddit is the ninth-most-visited website globally [1], has 300M monthly visitors, and has a $1.8B valuation [64].

On Reddit, links to Wikipedia often appear in the well-known "TIL" (Today I Learned) "subreddit", among others. The term "subreddit" refers to the sub-communities that

Figure 1. This image shows a popular "Today I Learned" Reddit post that is entirely reliant on a Wikipedia article.

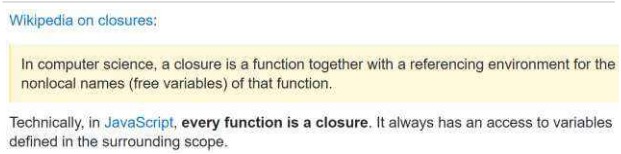

Figure 2. An answer on Stack Overflow that uses Wikipedia links and text.

together make up Reddit. Many posts in "TIL" are composed of only a Wikipedia link and a short summary or quote from the article. An example TIL post is shown in Figure 1.

Stack Overflow and Wikipedia

SO is a Q&A community for programmers and the 56th-most-visited website in the world [1]. Stack Exchange, the company that owns SO, raised $40M of venture capital funding in 2015, bringing it to a total of $70M raised [65].

On SO, Wikipedia supports answers in the form of links and quoted text. Answers often use technical terms or acronyms and include a Wikipedia link in lieu of defining these terms. An example post that uses a Wikipedia link and quote to support an answer is shown in Figure 2.

RELATED WORK

Wikipedia and Its Internet Ecosystem

This research project was directly motivated by two recent developments related to Wikipedia and its broader Internet ecosystem. First, the Wikimedia Foundation called on Wikipedia researchers to focus on these relationships as part of a "New Research Agenda" for Wikipedia [51]. Second, recent work further substantiated this call with new evidence showing how critical these relationships can be, both for Wikipedia and for the Web more generally [35].

With respect to the former, Dario Taraborelli – the Head of Research at Wikimedia – urged the large community of Wikipedia researchers to refocus their efforts on several new opportunities and challenges facing Wikipedia. One of those challenges is better understanding how Wikipedia (and Wikidata) content is used by external entities, the importance of this content to those entities, and the effect on Wikipedia (and Wikidata) of this external use. One concern of Taraborelli's is that the increasing re-use of Wikipedia content outside of Wikipedia will reduce traffic to Wikipedia. This would weaken Wikipedia's editor base over the long term, thus diminishing the re-use value of Wikipedia content. Taraborelli called this phenomenon the "paradox of re-use". We return to this concern below.

Recent work by McMahon et al. [35] provided clear evidence to substantiate Taraborelli's argument about the importance of understanding Wikipedia's relationships with its broader ecosystem. McMahon et al. found that click-through rates on Google SERPs drop dramatically when Wikipedia links are removed, and that Google is the mediator for the majority of observed Wikipedia traffic.

In addition to the direct implications for Google and Wikipedia, McMahon et al.'s study raised new questions for both the social computing community and the broader computing literature. For social computing, the amount of traffic to Wikipedia mediated by Google means that any changes to Google's presentation of Wikipedia content may have enormous effects on key variables of interest for those studying Wikipedia's peer production process, e.g., new editor recruitment and retention [7,17,18,62]. For the broader computing literature (especially information retrieval), McMahon et al. found that the effect size of the presence of Wikipedia content was much larger than many algorithmic search improvements. More generally, this finding demonstrated how important social computing phenomena like peer production can be for addressing core problems in other areas of computing like addressing user information needs in web search.

In this paper, we seek to expand on McMahon et al.'s work beyond search engines by considering Wikipedia's relationship with other large-scale online communities. We believed large-scale online communities would be an optimal domain to help follow-up McMahon et al.'s work for two reasons: (1) we anecdotally observed that this relationship may be particularly salient and (2) several research papers have provided early evidence that our anecdotal observations may generalize [15,40]. We also expand McMahon et al.'s lens to consider the quality of the associated Wikipedia content, i.e. does higher quality content create more value for external websites? This question provides further insight into how Wikipedia's internal quality metrics (and therefore investment of community effort) map to external value.

The early evidence that motivated us to examine large-scale online communities primarily came from two papers: Moyer et al. [40] and Gómez et al. [15]. Moyer et al. examined the causal effect of Reddit posts on Wikipedia page views and found that sharing Wikipedia content on Reddit's TIL community increased page views to the corresponding Wikipedia articles. Below, we replicate this result, but find that it does not generalize to other subreddits. While studying the role of links on SO with regards to innovation diffusion in the software engineering space, Gómez et al. [15] found an intriguing peripheral result: Wikipedia links were the second most common external link on SO. While we also observed that Wikipedia links are very common on SO, we found little *direct* evidence that SO is contributing substantial value to Wikipedia.

UGC and the Physical World

Through our research, we also seek estimate the economic impact of Wikipedia content on external entities. This lens on our results was motivated by very recent findings from Hinnosaar et al. [24], which showed that improvements to Wikipedia article content about places in Spain directly increased tourism revenue in those places [24]. Analogous results have also been identified with types of user-generated content (UGC) other than Wikipedia (e.g. [6,34]).

Factors Influencing Content Value

A key methodological challenge in our causal analyses below is identifying potential confounding factors for the substantial increases in value we see associated with Wikipedia-linked content on Reddit and SO. For example, users who post Wikipedia links may also write longer posts, and longer posts may be more popular. To search for potential confounds, we turned to the burgeoning literature

on Reddit, SO, and other large-scale online communities. We present the literature used to guide model decisions here, and summarize in greater detail how these factors were operationalized in the Methods section.

Specific factors influencing value on Reddit

From platform-specific work on Reddit, we identified three major factors that can influence user votes (score), a value metric that we use in this paper. Previous work identified content type [31] and title characteristics [23] as predictive of score on Reddit. Additionally, Gilbert showed that content posted multiple times (i.e. a particular image) that received a high score on later postings was often "missed" on early postings [14], implying that popularity is highly contextual, and not necessarily purely dependent on content itself.

Specific factors influencing value on SO

We also identified three factors that influence content value on SO. Anderson et al. found that contextual information, such as user reputation and time of arrival, were predictive of the long-term value of SO pages as measured by page views [2]. Ponzanelli et al. [43] showed that that adding readability metrics to simple textual metrics (e.g. percent uppercase, number of characters) improved low-quality post detection. Finally, Calefato et al. [5] identified promptness, presence of a hyperlink, presence of a code snippet, and answer sentiment as key predictors of an answer's likelihood to be selected as the best answer by the questioner.

Research on other question and answer (Q&A) sites also helped to inform our model design. On Math Overflow, a math-focused version of SO, both online reputation (points) and offline reputation (number of publications) were found to influence content popularity [52]. Harper et al. studied a variety of sites including Yahoo! Answers and identified that the number of hyperlinks in an answer was an indicator of answer quality [19]. We emphasize this potential confound in our analysis to understand the value that Wikipedia links add *beyond the value added by the presence of links overall*.

METHODS

In this section, we first present the two aspects of our methodology that cut across our investigations of both research questions: (1) data collection and (2) handling of current events. We then describe our methodology specific to Study One (RQ1) and Study Two (RQ2).

Datasets

We downloaded database tables corresponding to every Reddit post in 2016 from pushshift.io as hosted on Google BigQuery (metadata, i.e. the latest score, last updated July 2017 [66]). We also leveraged the Reddit API to obtain user information [67]. We used BigQuery to download full database tables for all SO questions, users, and answers, starting July 31, 2008 and ending June 11, 2017 [66].

Following statistical best practice [44], we separated the analysis into two phases: an initial phase for developing the methods and a testing phase for generating results. As we describe below, many of our analyses required using multiple rate-limited APIs (e.g. for calculating features for our causal analyses), so we employed random sampling to minimize API endpoint load and to make query times tractable. For each platform and phase, we used a random sample composed of ~1M posts from the entire dataset (1.10% of all Reddit posts and 4.46% of all SO posts) and, in the testing phase, an additional ~40K posts from the subset of Wikipedia-linking posts (to ensure that we had adequate Wikipedia links to pages of each class of article quality).

Defining "Value"

The concept of "value" is central to both of our research questions, i.e. the value that Wikipedia is providing large-scale online communities and the value that these communities are contributing to Wikipedia. In this paper, we seek to increase robustness and detect potentially interesting nuances by operationalizing the notion of value through multiple metrics rather than using just one (e.g. page views).

Specifically, we measure post value in RQ1 through four metrics. The first three metrics are user engagement statistics: (1) *Score*, equal to upvotes minus downvotes, (2) *Comments*, the number of comments a post receives, (3) *Page Views*, the number of views a post receives. To contextualize these metrics, we also calculate (4) *Revenue*, or the financial gain generated by Wikipedia posts. *Revenue* is calculated directly from the engagement statistics using publicly-available financial information (described in detail in Study 1 – Results). In the case of Reddit (which does not release page view data), it is important to note that score controls post visibility and correlates with page views [50].

With respect to RQ2, we assess the value that Wikipedia receives from external communities as contributions to the editing community and increased readership. Specifically, we measure this value with four metrics that capture changes in edits, editors, and viewership in a given week: (1) *Edit Count* is the number of times an article was edited, (2) *Editors Gained* is the number of new editors who edited an article, (3) *Editors Retained* is the number of new editors who made another edit in the future (we measured at one month and six months later, following past research on editor retention [7,39]), and (4) *Article Page Views* is the number of views that each Wikipedia article received. To capture the effect of Reddit and SO on Wikipedia, we calculated the metrics for the week before and the week after each post containing a Wikipedia link.

Influence of Current Events

One potential confound of all our measurements is the impact of current events on our value metrics. For instance, if Reddit users happened to post Wikipedia links related to current events, then any subsequent increase in Wikipedia page views might be largely driven by current events and not by the Reddit post. We predicted, however, that very few posts with Wikipedia links on either platform are related to current events because SO is strictly for programming discussion and the Reddit TIL community does not allow current events posts [68].

To formally verify this assumption, we performed a qualitative coding exercise. Following standard practices [41], we used a small (10 posts per site) calibration procedure with two researchers, achieved a 90% agreement, and then one researcher classified an additional 100 posts per site. In this qualitative analysis, we identified that only 5% of Wikipedia-linked Reddit posts were related to current events, and no Wikipedia-influenced Stack Overflow posts were related to current events. This gave us confidence that our results were largely not driven by current events.

STUDY ONE

Our first study targets RQ1, or "What value is Wikipedia providing to other large-scale online communities like Stack Overflow and Reddit?" Here, we present the methodology specific to this question and then present our results.

Study One – Methods

Controls and Treatments

In Study 1, our goal is to estimate the value that Wikipedia provides to SO and Reddit. We study value through two separate analyses. The various estimates are summarized in Table 1. First, we estimate the effects that the *presence* of Wikipedia links has on value. We then estimate the effects that Wikipedia article *quality* has on value. For each analysis, following standard practice, we label the change in value in the treatment group – the group associated with Wikipedia content – the "treatment effect". We defined three groups:

- *Has Wikipedia Link*: posts with at least one valid link to Wikipedia (as described below, this amounts to 0.13% of all Reddit posts and 1.28% of all SO posts).
- *Has Other Link*: posts with at least one external link, but no links to Wikipedia (49.1% of all Reddit posts and 31.2% of all SO posts).
- *No External Link:* posts with no external links at all (50.8% of all Reddit posts and 67.5% of all SO posts).

To estimate the effects of Wikipedia article quality on value to Reddit and SO (the second half of RQ1), we further subdivide the *Has Wikipedia Link* group into high-quality and low-quality groups. While there are many definitions of quality on Wikipedia [55], we rely on revision-specific predictions of quality along English Wikipedia's internal assessment scale [69] as produced by Wikimedia's ORES API [70]. Following Johnson et al. [29], we use the "C"-class assessment as a minimum assessment for a high-quality article because "C"-class articles are the first that are "useful to a casual reader" [71]. Specifically, we define our high-quality and low-quality groups as follows:

- *C-class or Better*: All posts with *any* links to C-class or higher articles are in this group (79% of all Wikipedia-linked Reddit posts and 77% of all Wikipedia-linked SO posts).
- *Below C-class*: All posts in which all Wikipedia links are to articles below C-class are in this group.

Simulating a World without Wikipedia (Counterfactuals)

We cannot know how SO and Reddit would function in a world without Wikipedia. Therefore, we draw upon well-established causal inference methods (e.g. [27,46]) that *estimate* the loss in value that would occur to these communities if posts with Wikipedia links were replaced with a *range of alternatives* (i.e. counterfactuals). In other words, we consider how SO and Reddit would be affected if they were not "treated" by Wikipedia content under a series of different assumptions. This approach provides a reasonable upper-bound, middle-ground, and lower-bound estimate of the value contributed by the Wikipedia community to SO and Reddit. In our consideration of the effect of Wikipedia article quality, we take a similar approach and estimate the loss in value if links to high-quality Wikipedia articles were replaced with a range of alternatives. We emphasize that causal analyses, the statistical methods we employ, can estimate causal effects with confounding effects reduced, but not eliminated (as in a randomized controlled trial) [27,46].

Upper-bound Estimates: Our upper-bound estimate of the value created by Wikipedia on SO and Reddit assumes that without Wikipedia, *all posts containing a Wikipedia link would not exist (i.e. would not have been generated).* For our analysis of quality, we make an analogous assumption: we calculate the value that would be lost if all posts containing high-quality articles did not exist. This upper bound is simply equivalent to the value of all "treated" posts.

Middle-ground Estimates: The middle-ground estimate corresponds to the value that would be lost if the links to treated articles were *completely replaced with an identical post with no links*. In other words, this scenario assumes that the links to Wikipedia were removed, but the post still exists.

Lower-bound Estimates: A lower-bound for value contributed by Wikipedia can be obtained by estimating the value that would be lost if the links to Wikipedia article were

Estimate	Counterfactual	Calculation
UB	All content containing (1) WP links or (2) higher quality WP links was never generated	Mean values
MG	(1) WP links or (2) higher quality WP links were removed from all content (posts remain)	Propensity score stratified multivariate regressions
	(1) Non-WP links or (2) low quality WP links were removed from all content (posts remain)	Propensity score stratified multivariate regressions
LB	(1) WP links or (2) higher quality WP links were replaced with other external links	MG minus above estimate

Table 1. Summary of analyses used to obtain upper-bound (UB), middle-ground (MG), and lower-bound (LB) estimates of the effect of Wikipedia (WP) links. In the Counterfactual column, (1) refers to Presence of WP and (2) to Quality of WP.

replaced with links to alternative external domains. In other words, the lower-bound scenario assumes that the exact same post (with alternative links) could be written without Wikipedia, and so the added value is solely from Wikipedia's reputation or other factors associated with Wikipedia (in comparison to other websites). In the quality analysis, the lower bound is the estimate of value that would be lost if links to high-quality articles were replaced with links to low-quality articles (without other changes).

Causal Analysis

While the upper-bound estimate is relatively easy to calculate using descriptive statistics, our middle-ground and lower-bound estimates require the estimation of formal counterfactuals, which calls for causal analysis techniques.

One of the first steps in casual analysis is to consider other potential causal factors, i.e. other reasons that Wikipedia-treated posts may have increased in value besides the Wikipedia treatment. To address this challenge, we turned to the literature on factors associated with content value, discussed above in Related Work and summarized in Table 2. This review broadly shows that value may come from four sources other than Wikipedia itself: user characteristics (e.g. users with high reputation), stylistic and structural characteristics (e.g. long posts, posts with code snippets, posts with punctuation), post timing (e.g. posting on a certain day of the week), and the presence of any external link.

We operationalize these potential alternative causal factors through the calculation of features that capture them. Some of these features are numerical (e.g. post length) and others are dummy variables (e.g. whether the post includes code snippets). While most of these features were very straightforward to calculate (e.g. title length), readability and sentiment were more complex. With respect to readability, we use the Coleman-Liau index [8], which was used by Ponzanelli et al. [43]. For sentiment, we used the TextBlob library [33], which leverages sentiment analysis models (trained on customer review text) from the well-performing [45] library "Pattern" [49] to estimate objectivity and polarity. Additionally, we log-transform reputation and response time as they are otherwise highly-skewed variables. Further details about these implementations are available in our source code, which we have made publicly accessible for download[1].

In terms of our statistical approach to causal analyses, we performed four propensity-score-stratified regressions[2]. This method, originally described by Rosenbaum and Rubin [46] has been used in the context of HCI for many purposes and is an approach advocated for in the HCI community (e.g. [10,11,42,63]). We controlled for the potential alternative causal factors – user, style, structure, and timing – and then estimate the effect of including a Wikipedia link, a non-Wikipedia link, a high-quality Wikipedia link, and a low-quality Wikipedia link.

Research Project	Factors Addressed in Our Study
Leavitt and Clark [31]	Content Type
Lakkaraju et al. [23]	Text length, sentence type, sentiment
Gilbert [14]	Day of week, month, hour, poster reputation
Anderson et al. [2]	Answer length, poster reputation, response time
Ponzenelli et al. [43]	Post length, percent uppercase, percent spaces, percent punctuation, post starts capitalized, Coleman-Liau index (readability)
Tausczik and Pennebaker [52]	User reputation
Calefato et al. [5]	User reputation, response time, presence of hyperlink, presence of code snippet, sentiment
Harper et al. [19]	Answer length, number of hyperlinks

Table 2. Summary of related work that identified factors that may affect post value in large-scale online communities.

By finding the difference between corresponding estimates (*Has Wiki Link* vs *Has Other Link* and *C-class or Better* vs *Below C-class)*, we produce a robust estimate of the value that Wikipedia uniquely provides, and the value that high-quality articles uniquely provide. This estimate is minimally affected by bias from hidden variables or incorrect model assumptions because this bias should affect the corresponding regressions equally. For instance, if there was a hidden variable at play (perhaps users have a general unwillingness to visit external domains), that would affect posts with Wikipedia links and posts with external links equally and be removed by taking the difference.

The propensity score of a given post is an estimated probability that the post was "treated" based on all available covariates. In our case, the covariates are the computed features summarized in Table 2 (except for the number of hyperlinks, which relates directly to whether there is a Wikipedia link). The actual propensity score for each post is calculated by logistic regression and represents how likely a given post is to include a Wikipedia link (or a good Wikipedia link) based only on its features. Posts that have features that are commonly found in Wikipedia-linked posts will therefore have high propensity scores.

Using the open-source *causalinference* [72] library, we *stratify* [3] our datasets by propensity score (subdivide into many subsets) in order to emulate a randomized blocked trial using observational data. Each stratum contains posts with a small range of propensity scores, so posts have similar features with each stratum, which reduces the standardized bias (described by Rosenbaum and Rubin [47]). To

[1] https://github.com/nickmvincent/ugc-val-est

[2] To check robustness, we also performed propensity score matching [3] and propensity score stratification with covariate overlap adjustment [9] and confirmed that these methods led to the same conclusions.

determine the number of strata and propensity scores for each stratum, we use a bin-selection algorithm described by Imbens and Rubin [27] and implemented in [72].

We performed a multivariate linear regression separately for each stratum, extracted the treatment coefficient, and then computed a weighted average of the treatment coefficients based on the number of treated items in each stratum. In other words, a coefficient from a stratum with more Wikipedia links than another stratum will be given greater weight in determining the estimated effect. The weighted average is the Average Treatment Effect on the Treated (ATT), which is the main result we present. This value represents how much value would be lost if all the treatment posts were replaced with control posts with nearly the same covariates.

The analysis of SO page views requires special handling because page views correspond to one question page, which could have many associated SO answers. This means we must take care to avoid overestimating page view effects. When computing mean page views for the upper-bound estimate, we only count each question once. In our causal analysis (i.e. middle-ground and lower-bound estimate), we make a conservative assumption that all page views are attributed to the top-scoring answer for a question, which means that only top answers are included in this analysis.

Study 1 – Results

In this section, we first present high-level descriptive statistics about the relationships between Wikipedia and SO and Reddit (e.g. # of Wikipedia-linked posts). We then present the results from our core analyses for RQ1.

Descriptive Results

Overall, we were surprised to find that Wikipedia links represent only 0.13% of posted content on Reddit. However, further examining this result, we found that Wikipedia links are substantially over-represented in high-value locations. For instance, Wikipedia is the third most-linked site (after YouTube and Imgur) on the ten most-popular subreddits. This relatively low-quantity/high-quality dynamic is one we see frequently in our formal analyses below.

On SO, Wikipedia links appear in 1.28% of posts, but this makes them the fourth-most-common type of external link (after github.com, msdn.microsoft.com, and the popular "code playground" jsfiddle.net). Notably, github and jsfiddle are used to share code, and msdn is Microsoft's code documentation library, meaning that Wikipedia is the most important conceptual reference for programming on SO.

Effects of Wikipedia-linked content on Post Value

The results from our full analysis to address RQ1 are presented in Table 3. Below, we unpack the main trends in Table 3, as well as discuss the implications of Table 3 with respect to Wikipedia's aggregate impact on SO and Reddit.

Effects on Reddit posts: Table 3a shows that Wikipedia-linked posts on Reddit are exceptionally valuable. To a post's score, Wikipedia adds between 108 points (ATT, **Δ** in Table 3a) and 151 points (Mean Values: Has WP Link, with a middle-ground estimate of 141 points (ATT: Has WP Link vs No External Link). Relative to the average post's score of 30 points, this is a 4x-5x increase. Aggregating these findings across all 120K Wikipedia-linked posts from 2016, this means that Wikipedia is responsible for an increase in

		Mean Values			ATT [99% CI]			
	Variable	Has WP Link (UB)	No WP Link	Δ	Has WP Link vs No External Link (MG)	Has Other Link vs No External Link	Δ (LB)	Ratio (LB-UB)
3a) Presence of WP								
R	**Score**	151	31	120	141 [119.4-163.1]*	34 [31.0-36.4]*	108 [85.6-130.0]*	4.5-4.9
	Comments	19	8	11	7 [2.4-11.0]*	-4.0 [-5.2- -2.7]*	11 [6.2-15.1]*	2.3-2.3
SO	**Score**	6.5	2.5	4.0	3.4 [0.9-6.0]*	0.5 [0.3-0.7]*	2.9 [0.4-5.4]*	2.1-2.6
	Comments	1.7	1.4	0.3	0.0 [-0.04-0.1]	0.0 [-0.1- 0.0]	0.0 [-0.02-0.1]	1-1.2
	Page views	8535	5062	3473†	1337 [727-1947]*	383 [252-515]*	954 [330-1578]*	1.2-1.7
3b) Quality of WP								
	Variable	**C-class or Better (UB)**	**Below C-class**	**Δ**	**C-class or Better vs No External Link (MG)**	**Below C-class vs No External Link**	**Δ (LB)**	**Ratio (LB-UB)**
R	**Score**	163	107	56	151 [126.5-175.3]*	90 [63.9-115.4]*	61 [25.7-96.7]*	1.5-1.5
	Comments	21	13	8	7 [2.6-12.2]*	-1 [-5.7-4.0]	8 [1.5-15.1]*	1.6-1.6
SO	**Score**	6.6	6.2	0.4	4.3 [0.3-8.4]*	2.5 [0.8-4.2]*	1.9 [-2.5-6.2]	1-1.1
	Comments	1.7	1.7	0.0	0 [-0.1-0.1]	0 [-0.1-0.1]	0 [-0.1-0.1]	1-1
	Page views	8834	7955	880†	1492 [762-2222]*	1309 [141-2478]*	183 [-1195-1560]	1-1.1

Table 3. Summary of upper-bound (UB), middle-ground (MG) and lower-bound (LB) estimated effects that WP links have on Reddit (R) and Stack Overflow (SO) for both the Presence of WP analysis (3a) and Quality of WP analysis (3b); * indicates p < 0.01; † indicates upper-bound estimate for SO page view analysis. "Ratio (LB-UB)" is the value ratio - how much more valuable is a treated post compared to average - for the LB and UB estimates.

user-voted score of between 13.1M and 18.4M points in 2016 (up to 0.7% of all points on the site).

Additionally, Table 3a also shows that Wikipedia adds 11-19 comments per post. This means Wikipedia-linked posts generate twice as much discussion as average posts. In total, this amounted to between 1.3M and 2.3M comments in 2016.

Effects on Stack Overflow posts: For SO, Table 3a displays a similar trend to what we saw with Reddit. For instance, Wikipedia-linked content on SO adds 2.9-6.5 points per post, with a middle ground of about 3.4 points. This means that Wikipedia-linked answers are roughly twice as valuable as other answers and, across the 280K Wikipedia-linked answers on SO, increased total score between 0.8M and 1.7M points. This score increase is accompanied by a page view increase of 954-3473 per question, with a middle ground of 1337. Estimating based on the 12K questions with Wikipedia-linked answers in our page view analysis, Wikipedia may have added 64M to 234M views (which we use for revenue estimation). However, we see no evidence of an effect on SO comments.

Overall, the presented estimates show that even assuming all authors could continue writing the same posts, except with non-Wikipedia alternative links, Wikipedia still adds significant value (i.e. through its brand or other factors).

Effect of Wikipedia Article Quality

Compared to posts that only have links to *Below C-class* Wikipedia articles, we find mixed evidence that links to *C-class or Better* articles contribute to value. Relative to *Below-C-class articles*, *C-class or Better* articles add 61-163 points on Reddit (1.5-1.5x increase). Similarly, *C-class or Better* articles also add 8-21 comments on Reddit (1.6-1.6x increase). However, we observe no effect on SO score, comments, or page views.

The above results suggest that article quality for the purposes of the SO community may mean something different than article quality for the Wikipedia community (and Reddit). For instance, SO members may not differentiate between the value of short or stub-like articles and longer, high-quality articles, as long as those articles contain a specific piece of desired technical information.

Back-of-the-Napkin Revenue Estimation

To better understand how the results of Study 1 translate into real-world revenue for SO and Reddit, we use the following "back-of-the-napkin" revenue estimations, incorporating as much actual public financial information as possible and making conservative assumptions (e.g. assuming all SO ads were sold for the lowest price listed). Both platforms sell ads at a fixed cost per thousand impressions and therefore revenue scales linearly with page views.

Estimating Reddit's 2016 ad revenue is relatively simple, as Reddit only makes money from ads and Reddit Gold (a subscription service). Using Reddit's $20M revenue projection [73] and an approximation of $1M revenue from "Reddit Gold" [63], we presume a total ad revenue of $19M. To obtain a rough figure for Stack Overflow's total revenue from 2008-2016, we use the following equation:

$$rev_{SO} = (views - votes) * \frac{ads}{views} * \frac{cost}{ad} * adblock$$

The $(views - votes)$ accounts for the fact that high reputation users see reduced ads [75] (we conservatively assume that every vote was made by a high reputation user, and therefore the corresponding view should not be included in ad revenue). We also multiply by an "adblock coefficient" of 0.75 to account for the 25% of desktop users who block advertisements [76]. We conservatively assumed all ads cost $0.00466 [77], the lowest price listed in August 2017 and that users see only two impressions per page (users who scroll down actually see three).

Finally, because Reddit page view data is not available, we estimate page views using score, based on research showing that the score of a Reddit post correlates with views [50].

Overall, under these conservative assumptions, we estimate that in 2016, Wikipedia was responsible for between $114,500 and $124,200 of Reddit's revenue, and from 2008 to 2017, Wikipedia annually was responsible for between $49,700 and $180,900 of SO's annual revenue.

In total, considering only the content analyzed from two communities and the limited time periods we studied (nine years of SO activity and only one year of Reddit activity), Wikipedia may have been (conservatively) responsible for about $1.7 million in revenue, entirely from volunteer work of the community.

STUDY 2

In this section, we present a study that addresses RQ2, or "What value do Reddit and SO provide to Wikipedia?" For the before-and-after windows for each post (one week each), we calculated the *Edit Count*, *Editors Gained*, *Editors Retained*, and *Article Page Views* (reported daily by the Wikipedia page view API), as discussed above.

Study 2 – Results

Table 4 shows the full results of our quantitative before-and-after investigation into RQ2.

The clearest trend in Table 4 is the near-absence of any significant results for both edit behavior and page views. With regard to edit behavior, on both platforms, we did not observe a significant increase in editors gained or editors retained during the week after a Wikipedia-linking post appeared. While we did observe an effect for the number of edits from Reddit posts, this effect is quite small (about 1 edit per 2 posts). If we consider all the Wikipedia-linking Reddit posts from 2016, this amounts to about 0.002 edits per second. Compared to Wikipedia's 10 edits per second [78], this effect is largely negligible. We did observe one minor significant effect that warrants discussion: Low-quality articles ("Stub" and "Start") were edited enough to achieve a statistically significant increase in edits relative to high-

	SO			Reddit		
	Before	*After*	*Increase (99% CI)*	*Before*	*After*	*Increase (99% CI)*
Edit count	1.440	1.426	-0.014 [-0.054 – 0.027]	4.584	4.990	0.405 [0.1 – 0.7]*
Editors gained	0.050	0.047	-0.003 [-0.009 – 0.004]	0.102	0.098	-0.004 [-0.02 – 0.01]
Editors retained (1 month later)	0.011	0.011	0.000 [-0.003 – 0.003]	0.016	0.017	0.002 [-0.004 – 0.007]
Article page views (daily)	10472	10697	225 [-243 – 694]	94323	96536	2213 [-2082 – 6508]
Article page views from TIL posts (daily)				47737	54026	6289 [2228 – 10350]*

Table 4. Summary of the effects of SO and Reddit on Wikipedia; * p < 0.01.

quality articles, even on SO where no other effects were observed. The effect was only 0.04 edits per post, but the presence of this small effect suggests that lower-quality articles received proportionally more contribution from Reddit and SO. We discuss the implications below.

During this analysis, we also analyzed the effect of Reddit posts exclusively from the TIL community on Wikipedia page views, a replication of work by Moyer et al. that used 2012 Reddit data [40]. In this specific case, we found the increase in page views corresponded with previous results. However, when including posts from outside TIL and when looking at SO, we found that the increase in page views was not statistically significant, indicating that while this result replicates, it may not generalize (we note that we used the Wikipedia pageview API, which returns daily statistics, whereas Moyer et al. used hourly page views). In other words, we saw no evidence that an arbitrary link to a Wikipedia page on Reddit and SO significantly increases traffic to the Wikipedia page in the week following.

DISCUSSION AND IMPLICATIONS

Paradox of Re-use

As noted above, a concern that has been raised within the Wikipedia community is the "paradox of re-use" – i.e. that the quality of Wikipedia's content and Wikipedia's editor base could decline over the long term through the external re-use of Wikipedia content (and that the re-using parties would then suffer as well) [51]. This concern leads to the question of whether Wikipedia needs to adapt its permissive licensing to survive, especially as its content is increasingly appearing on platforms more than a "click away" (e.g. voice assistants). Given that we observed that Wikipedia-linking posts were receiving a great deal of attention and engagement on SO and Reddit, we were surprised that so little of this attention and engagement was returned to Wikipedia in the form of page views, edits, and editors. We hypothesized that the paradox of re-use may be a factor.

To understand exactly how SO and Reddit use content from Wikipedia, we performed a small-scale qualitative coding exercise. Using the same approach to qualitative coding as the current events analysis, we classified whether each post matched the following definition of direct re-use: "Has text (quoted or summarized) from article." Two authors conducted a calibration coding and achieved reasonable inter-rater agreement (Reddit: Cohen's κ = 0.74, 90% agreement; SO: Cohen's κ = 0.62; 90% agreement), and then one author coded 100 posts with Wikipedia links per site.

Our coding analysis revealed that 79% of Reddit posts had quoted or summarized text from the linked Wikipedia article, whereas this was true of 33% of SO posts. This result shows that many posts can be characterized by "direct re-use".

These early results help to expand our understanding of the nature of the "paradox of re-use" from McMahon et al. [35]. McMahon et al. found causal evidence for the paradox of re-use on Google – i.e. Knowledge Card assets in Google search results were capturing views that would have otherwise gone to Wikipedia. This indicates a "strong" version of the paradox of re-use, in which Google was capturing traffic that would have gone to Wikipedia. Our (associative) results suggest that a "weak" form of the paradox may be occurring in Reddit and SO: the Wikipedia links on Reddit and SO posts might have generated new traffic for Wikipedia (rather than capturing existing traffic), but the re-use of content directly in Reddit and SO may be mitigating this effect. However, it is critical to note that our results merely allow this as a possibility: future work should test this hypothesis directly, ideally through experimental approaches.

Broadening Peer Production Research

At a high level, the results of Study 1 demonstrate that there are important relationships between different peer-production communities (with Reddit and SO themselves being peer production communities). We see that when content creators on Reddit and SO leverage existing peer-produced content (i.e. Wikipedia) to create new content in their communities, some of the "value" transfers and the magnitude of that value is reflective of the re-used content's quality (in this case, the quality of Wikipedia articles). In other words, Reddit and SO users can take advantage of the work already performed by the Wikipedia community to post higher value content in their communities with less effort.

These relationships between peer production communities have important implications for research. Studies of some important variables such as content quality may want to consider external dependencies and implications, e.g. Wikipedia article quality matters well outside of Wikipedia, adding importance to the body of work that studies Wikipedia quality and how to improve it [7,18,55,56,62,70].

Coverage Biases and External Entities

The work on Wikipedia content biases is another key research area in the Wikipedia literature that our results suggest should be examined with a broader lens. Wikipedia (and other UGC communities) have been shown to over-focus on some topics and under-focus on others, particularly

along the lines of gender (e.g. [22,37,54]), geography (e.g. [20,29]), and language (e.g. [21]). Importantly, our results suggest that these biases may be having ripple effects around the web, including on SO and Reddit. For instance, the findings from Study 1 indicate that these biases mean that the TIL community on Reddit is working from a dataset that disadvantages women, certain geographies (e.g. rural areas [29]), and topics outside of the cultural sphere of most English speakers. Any posts about these Wikipedia-disadvantaged topics that do appear on TIL are less likely to have a Wikipedia link (or high-quality link) available, and thus cannot get the benefits of that link.

There is a potential silver lining regarding coverage biases: while the number of edits contributed to Wikipedia from Reddit that we observed in Study 2 was quite small in the context of Wikipedia as a whole, it may be that this effect would be meaningful if applied to specific under-covered content areas. For instance, if more Wikipedia links were posted on a subreddit for rural issues, the throughput of edits that is quite small for all of Wikipedia could make non-trivial improvements to coverage of some rural topics, for which a small number of edits is a large relative increase in edits.

One could also imagine the Wikipedia community (e.g. WikiProject leaders) working with the moderators of subreddits in under-covered areas on intentional "quality improvement projects" [55]. This might involve rallying these subreddits to encourage edits to relevant Wikipedia articles and to encourage members to become long-term editors. It would be interesting to compare the results of such an effort to our findings, which would provide an excellent baseline to determine if the effort was effective.

Implications for Reddit and Stack Overflow

McMahon et al.'s work [35] indicated that financially supporting Wikipedia may be highly incentivized for search companies, and our work suggests that the same might be true for Reddit and SO (although to a lesser degree). In other words, our results suggest that by donating to Wikipedia, entities like SO and Reddit can not only garner goodwill, but also could feasibly see a return on investment. In general, donations to support staffing and infrastructure are an understudied aspect of the peer production process, especially relative to their importance. Our results indicate that one direction of research in this space that might be fruitful is identifying formal value propositions for external consumers of peer-produced content.

Our results also suggest that Reddit and SO design interventions could help increase the mutual value of Wikipedia-SO and Wikipedia-Reddit relationships. For instance, Reddit and SO could implement a feature that detects when linked peer produced content is low quality or under-covered and directly encourage users to contribute. By improving Wikipedia content, this would in turn add value to the Reddit and SO and could even be gamified, e.g. giving extra Reddit/SO "karma" when users contribute. Similarly, to facilitate higher-quality posts on Reddit and SO, topic modeling could also be used to suggest related Wikipedia articles to improve post quality.

FUTURE WORK AND LIMITATIONS

An important direct extension of this work would be to attempt to replicate our causal analysis on observational data with a randomized experiment. However, experiment-driven posting behavior could be seen as deviant by the SO and Reddit communities, so ethical and ecological validity challenges should be carefully considered.

Secondly, while the "back of the napkin" estimations of revenue above provide a minor contribution to understanding the economic implications of volunteer peer-produced content, much more work is needed in this area. This paper, along with prior research [35], suggests that volunteers' content creation efforts generate important economic value well outside the corresponding communities. Understanding this value could provide intrinsic motivation for volunteers and incentivize donations from beneficiaries. Some economists even believe that better understanding this value could produce improved national GDP estimates [4]. However, moving forward in this space will be difficult because many of the corporate beneficiaries of volunteer-created content do not release the necessary information for rigorous estimates. As such, new methods that go beyond our "back of the napkin" approach will likely be necessary, such as using targeted ad buys to estimate actual revenue, advancing techniques similar to those of Cabañas et al. [16], and attempting to measure the value of volunteer-created content to profitable machine learning systems.

Finally, given that most of our results were not community-specific but rather generalized across Reddit and SO, similar relationships likely exist between Wikipedia and other discussion-based communities (e.g. akin to Reddit: Hackernews, Facebook groups; akin to SO: Quora, Yahoo! Answers). Future work should seek to examine these relationships, as well as those between these communities and peer-produced data repositories like OpenStreetMap.

CONCLUSION

In this paper, we presented results that identify and quantify relationships between Wikipedia and the large-scale communities Reddit and Stack Overflow. In general, we observe a one-way relationship in which Wikipedia-influenced content adds value to the external communities, but no evidence of substantial contributions in the reverse direction is observed. This research highlights the value of examining online communities using a broad lens, as cross-community relationships can have large effects.

ACKNOWLEDGEMENTS

We would like to thank our colleagues at GroupLens (University of Minnesota) and HCI Group Bremen (University of Bremen) for their feedback on this work. This project was funded in part by the U.S. National Science Foundation (CAREER IIS-1707296 and IIS-1707319).

REFERENCES

1. Alexa.com. 2018. Alexa Top 500 Global Sites. Retrieved July 17, 2017 from http://www.alexa.com/topsites
2. Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '12), 850–858. https://doi.org/10.1145/2339530.2339665
3. Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46, 3: 399–424.
4. Erik Brynjolfsson and Andrew McAfee. 2014. *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.
5. Fabio Calefato, Filippo Lanubile, Maria Concetta Marasciulo, and Nicole Novielli. 2015. Mining successful answers in stack overflow. In *Mining Software Repositories (MSR), 2015 IEEE/ACM 12th Working Conference on*, 430–433.
6. Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43, 3: 345–354.
7. Giovanni Luca Ciampaglia and Dario Taraborelli. 2015. MoodBar: Increasing new user retention in Wikipedia through lightweight socialization. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 734–742.
8. Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60, 2: 283.
9. Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96, 1: 187–199.
10. Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2098–2110.
11. Dean Eckles and Eytan Bakshy. 2017. Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint arXiv:1706.04692*.
12. Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 1625–1628.
13. Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, 1606–1611.
14. Eric Gilbert. 2013. Widespread Underprovision on Reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (CSCW '13), 803–808. https://doi.org/10.1145/2441776.2441866
15. Carlos Gómez, Brendan Cleary, and Leif Singer. 2013. A Study of Innovation Diffusion Through Link Sharing on Stack Overflow. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (MSR '13), 81–84. Retrieved from http://dl.acm.org/citation.cfm?id=2487085.2487105
16. José González Cabañas, Angel Cuevas, and Rubén Cuevas. 2017. FDVT: Data Valuation Tool for Facebook Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3799–3809.
17. Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. 2013. Making peripheral participation legitimate: reader engagement experiments in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 849–860.
18. Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*, 163–172.
19. F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. 2008. Predictors of Answer Quality in Online Q&A Sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), 865–874. https://doi.org/10.1145/1357054.1357191
20. 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.
21. 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.
22. Benjamin Mako Hill and Aaron Shaw. 2013. The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. *PloS one* 8, 6: e65782.
23. Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. *International AAAI Conference on Web*

and Social Media; Seventh International AAAI Conference on Weblogs and Social Media. Retrieved January 1, 2013 from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6085

24. Marit Hinnosaar, Toomas Hinnosaar, Michael Kummer, and Olga Slivko. 2017. *Wikipedia Matters*.
25. Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194: 28–61.
26. Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. 2008. Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 179–186.
27. Guido W Imbens and Donald B Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge University Press.
28. Jack Clark. 2015. Google Turning Its Lucrative Web Search Over to AI Machines - Bloomberg. *Bloomberg L.P.* Retrieved from https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines
29. 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 *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 13–25.
30. 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*, 37–46.
31. Alex Leavitt and Joshua A. Clark. 2014. Upvoting Hurricane Sandy: Event-based News Production Processes on a Social News Site. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems* (CHI '14), 1495–1504. https://doi.org/10.1145/2556288.2557140
32. Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS)* 2, 2: 11.
33. Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, and others. 2014. TextBlob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*.
34. Michael Luca. 2011. Reviews, Reputation, and Revenue: The Case of Yelp.com. Retrieved June 21, 2017 from http://www.hbs.edu/faculty/Pages/item.aspx?num=41233
35. Connor McMahon, Isaac L Johnson, and Brent J Hecht. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *ICWSM*, 142–151.
36. Olena Medelyan, David Milne, Catherine Legg, and Ian H Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies* 67, 9: 716–754.
37. Amanda Menking and Ingrid Erickson. 2015. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 207–210.
38. David Milne and Ian H Witten. 2013. An open-source toolkit for mining Wikipedia. *Artificial Intelligence* 194: 222–239.
39. Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 839–848.
40. Daniel Moyer, Samuel L Carson, Thayne Keegan Dye, Richard T Carson, and David Goldbaum. 2015. Determining the influence of Reddit posts on Wikipedia pageviews. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*.
41. Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey research in HCI. In *Ways of Knowing in HCI*. Springer, 229–266.
42. Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. In *CSCW*, 370–386.
43. Luca Ponzanelli, Andrea Mocci, Alberto Bacchelli, Michele Lanza, and David Fullerton. 2014. Improving low quality stack overflow post detection. In *Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on*, 541–544.
44. Yuqing Ren and Robert E Kraut. 2014. Agent based modeling to inform the design of multiuser systems. In *Ways of Knowing in HCI*. Springer, 395–419.
45. Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1: 1–29.
46. Paul R Rosenbaum and Donald B Rubin. 1984. Reducing bias in observational studies using

subclassification on the propensity score. *Journal of the American statistical Association* 79, 387: 516–524.

47. Paul R Rosenbaum and Donald B Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 1: 33–38.
48. Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213: 1063–1064.
49. Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research* 13, Jun: 2063–2067.
50. Greg Stoddard. 2015. Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News. In *ICWSM*, 416–425.
51. Dario Taraborelli. 2015. The Sum of All Human Knowledge in the Age of Machines: A New Research Agenda for Wikimedia.
52. Yla R. Tausczik and James W. Pennebaker. 2011. Predicting the Perceived Quality of Online Mathematics Contributions from Users' Reputations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 1885–1888. https://doi.org/10.1145/1978942.1979215
53. Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM* 14: 505–514.
54. Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *ICWSM*, 454–463.
55. Morten Warncke-Wang, Vladislav R Ayukaev, Brent Hecht, and Loren G Terveen. 2015. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 743–756.
56. Morten Warncke-Wang, Vivek Ranjan, Loren G Terveen, and Brent J Hecht. 2015. Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. In *ICWSM*, 493–502.
57. Robert West, Ingmar Weber, and Carlos Castillo. 2012. Drawing a data-driven portrait of Wikipedia editors. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, 3.
58. Eric Yeh, Daniel Ramage, Christopher D Manning, Eneko Agirre, and Aitor Soroa. 2009. WikiWalk: random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, 41–49.
59. Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. 2014. The Impact of Membership Overlap on the Survival of Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 281–290. https://doi.org/10.1145/2556288.2557213
60. Haiyi Zhu, Robert E Kraut, Yi-Chia Wang, and Aniket Kittur. 2011. Identifying shared leadership in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3431–3434.
61. Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Effectiveness of shared leadership in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 407–416.
62. Haiyi Zhu, Amy Zhang, Jiping He, Robert E Kraut, and Aniket Kittur. 2013. Effects of peer feedback on contribution: a field experiment in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2253–2262.
63. 2017. OSSM17: Observational Studies Through Social Media.
64. Reddit worth $1.8 billion. Retrieved from https://www.cnbc.com/2017/07/31/reddit-worth-1-point-8-billion.html
65. Stack Exchange, a site for software developers, raises $40 million | Fortune.com. Retrieved from http://fortune.com/2015/01/20/stack-exchange-40-million/
66. Google BigQuery. Retrieved from https://bigquery.cloud.google.com/dataset/bigquery-public-data:stackoverflow
67. reddit.com: api documentation. Retrieved from https://www.reddit.com/dev/api/
68. Today I Learned (TIL). Retrieved from https://www.reddit.com/r/todayilearned/
69. Wikipedia:WikiProject Wikipedia/Assessment - Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikipedia/Assessment
70. Artificial intelligence service "ORES" gives Wikipedians X-ray specs to see through bad edits – Wikimedia Blog. Retrieved July 17, 2017 from https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/
71. Template:Grading scheme - Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Template:Grading_scheme
72. Causal Inference in Python — Causalinference 0.1.2 documentation. Retrieved from http://causalinferenceinpython.org/

73. Reddit's plan to become a real business could fall apart pretty easily - Recode. Retrieved from https://www.recode.net/2016/4/28/11586522/reddit-advertising-sales-plans
74. Reddit Gold Counter - How much is reddit making from reddit gold. Retrieved from http://gold.reddit-stream.com/gold/table
75. Ad Banners - Developer Advertising Solutions | Advertise on Stack Overflow. Retrieved from https://www.stackoverflowbusiness.com/advertise/solutions/ad-banners
76. IAB Study Says 26% of Desktop Users Turn On Ad Blockers – Adweek. Retrieved from http://www.adweek.com/digital/iab-study-says-26-desktop-users-turn-ad-blockers-172665/
77. Book a campaign | Stack Exchange Self-Serve. Retrieved from https://www.selfserve-stackexchange.com/
78. Wikipedia:Statistics - Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Statistics

Note: in the previously published version, references 1-63 had an off-by-one error. This has been fixed in this version.