Why We Read Wikipedia

Philipp Singer[*1], Florian Lemmerich[*1], Robert West[†2],
Leila Zia[3], Ellery Wulczyn[3], Markus Strohmaier[1], Jure Leskovec[4]

[1]GESIS & University of Koblenz-Landau, [2]EPFL, [3]Wikimedia Foundation, [4]Stanford University
[1]firstname.lastname@gesis.org, [2]robert.west@epfl.ch, [3]firstname@wikimedia.org, [4]jure@cs.stanford.edu

ABSTRACT

Wikipedia is one of the most popular sites on the Web, with millions of users relying on it to satisfy a broad range of information needs every day. Although it is crucial to understand what exactly these needs are in order to be able to meet them, little is currently known about why users visit Wikipedia. The goal of this paper is to fill this gap by combining a survey of Wikipedia readers with a log-based analysis of user activity. Based on an initial series of user surveys, we build a taxonomy of Wikipedia use cases along several dimensions, capturing users' motivations to visit Wikipedia, the depth of knowledge they are seeking, and their knowledge of the topic of interest prior to visiting Wikipedia. Then, we quantify the prevalence of these use cases via a large-scale user survey conducted on live Wikipedia with almost 30,000 responses. Our analyses highlight the variety of factors driving users to Wikipedia, such as current events, media coverage of a topic, personal curiosity, work or school assignments, or boredom. Finally, we match survey responses to the respondents' digital traces in Wikipedia's server logs, enabling the discovery of behavioral patterns associated with specific use cases. For instance, we observe long and fast-paced page sequences across topics for users who are bored or exploring randomly, whereas those using Wikipedia for work or school spend more time on individual articles focused on topics such as science. Our findings advance our understanding of reader motivations and behavior on Wikipedia and can have implications for developers aiming to improve Wikipedia's user experience, editors striving to cater to their readers' needs, third-party services (such as search engines) providing access to Wikipedia content, and researchers aiming to build tools such as recommendation engines.

Keywords: Wikipedia; survey; motivation; log analysis

1. INTRODUCTION

Wikipedia is the world's largest encyclopedia and one of the most popular websites, with more than 500 million pageviews per day. It attracts millions of readers from across the globe and serves a broad range of their daily information needs. Despite this, very little is known about the motivations and needs of this diverse user group: why they come to Wikipedia, how they consume the content in the encyclopedia, and how they learn. Without this knowledge, creating more content, products, and services that ensure high levels of user experience remains an open challenge [3, 8, 20, 43].

Background and objectives. A rich body of work has investigated motivations and behavior patterns of users on the Web [11, 21]. Specific attention has been cast on a few major sites including search engines [6, 33, 41], and social networking sites such as Twitter [15, 22] and Facebook [34]. Yet, surprisingly little is known about the motivations, needs, and behaviors of Wikipedia readers, possibly keeping Wikipedia from reaching its full potential.

The vast literature on user behavior in Wikipedia (*cf.* Okoli *et al.* [30] for an overview) has focused on content production. It mainly investigates editors' motivations [1, 29], patterns of editing behavior [16], and the quality of content [18, 39]. Much less is known about content consumption, even though readers make up the majority of Wikipedia users. The limited work on readers has focused on topics such as content preferences [26, 32, 37], search queries leading to Wikipedia [40], or navigation patterns [23, 31, 36, 42]. In contrast, the present work aims at understanding *why we read Wikipedia*.

Materials and methods. We present a robust taxonomy of use cases for reading Wikipedia, constructed through a series of surveys based on techniques from grounded theory [38]. Initially, we administered a survey to elicit free text responses to the question, *Why are you reading this article today?* Based on the responses, we designed a taxonomy covering three major dimensions that can be used to characterize the observed use cases. After validating the robustness of our taxonomy, we study the prevalence of use cases as measured by a large-scale multiple-choice survey on English Wikipedia. To correct for various forms of representation bias in our pool of respondents, we use inverse propensity score weighting adjustment. We then enrich the survey data by linking each survey response to the respondent's behavior traces mined from Wikipedia's webrequest logs. An illustration of how both survey and log data are collected can be found in Fig. 1. Finally, we use the joined survey and log data to identify characteristic behavior patterns for reader groups with specific intentions via subgroup discovery [14, 19].

Contributions and findings. The following are our three main contributions: (i) We present a robust taxonomy for characterizing use cases for reading Wikipedia (Sec. 2), which

[*]Both authors contributed equally to this work.
[†]Robert West is a Wikimedia Foundation Research Fellow.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052716

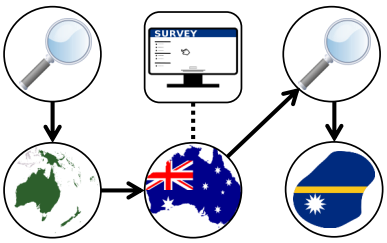

Figure 1: *Example Wikipedia reading session.* The user arrives from a search engine and visits the article about *Oceania;* she then navigates to *Australia,* where she responds to our survey. Afterwards, the reader goes back to the search engine and finally visits the article about *Nauru.* This paper studies survey responses as well as webrequest logs.

captures users' motivations to visit Wikipedia, the depth of information they are seeking, and their familiarity with the topic of interest prior to visiting Wikipedia. (ii) We quantify the prevalence and interactions between users' motivations, information needs, and prior familiarity via a large-scale survey yielding almost 30,000 responses (Sec. 4.1). (iii) We enhance our understanding of the behavioral patterns associated with different use cases by combining survey responses with digital traces recorded in Wikipedia's server logs (Sec. 4.2).

Our analysis lets us conclude that there is a variety of motivations bringing readers to Wikipedia, which can be characterized by distinct behavioral patterns. For example, users visiting Wikipedia out of boredom view multiple, topically different articles in quick succession. While no motivation clearly dominates, it is generally the case that Wikipedia is used for shallow information needs (fact look-up and overview) more often than for in-depth information needs. Also, prior to reading an article, readers are familiar with its topic about as often as not.

The outcomes of this research can help Wikipedia's editor and developer communities, as well as the Wikimedia Foundation, to make more informed decisions about how to create and serve encyclopedic content in the future.

2. TAXONOMY OF WIKIPEDIA READERS

Our research relies on a taxonomy of Wikipedia readers, something that was previously absent from the literature. We designed and analyzed a series of surveys based on techniques from *grounded theory* [38] to build a robust categorization scheme for Wikipedia readers' motivations and needs. In this section, we explain the individual steps taken and the resulting taxonomy.

Building the initial taxonomy. We started with an initial questionnaire *(Survey 1)*, where a randomly selected subgroup of English Wikipedia readers (sampling rate 1:200, desktop and mobile, 4 days, about 5,000 responses) saw a survey widget while browsing Wikipedia articles. If the reader chose to participate, she was taken to an external site (Google Forms) and asked to answer the question *"Why are you reading this article today?"* in free-form text (100-character limit).

To arrive at categories for describing use cases of Wikipedia reading, five researchers performed three rounds of hand-coding on a subset of the 5,000 responses, without discussing any expectations or definitions ahead of time. In the first stage, all researchers worked together on 20 entries to build a common understanding of the types of response. In the second stage, based on the discussions of the first stage, tags were generously assigned by each researcher individually to 100 randomly selected responses, for a total of 500 responses tagged. All 500 tagged responses were reviewed, and four main trends (motivation, information need, context, and source) were identified, alongside tags associated with each response. In the third stage, each researcher was randomly assigned another 100 responses and assessed if they contained information about the four main trends identified in the previous stage and if the trends and tags should be reconsidered. The outcome of these stages revealed the following three broad ways in which users interpreted the question; we use them as orthogonal dimensions to shape our taxonomy:

- *Motivation*: work/school project, personal decision, current event, media, conversation, bored/random, intrinsic learning.
- *Information need*: quick fact look-up, overview, in-depth.
- *Prior knowledge*: familiar, unfamiliar.

Assessing the robustness of the taxonomy. We conducted two surveys similar to *Survey 1* on the Spanish and Persian Wikipedias which resulted in similar observations and dimensions as above. Additionally, we assessed the robustness of the above taxonomy in two follow-up surveys. First, we ran a survey identical to Survey 1 to validate our categories on unseen data (*Survey 2;* sampling rate 1:2000, mobile, 3 days, 1,650 responses). No new categories were revealed through hand-coding.

Second, we crafted a multiple-choice version of the free-form surveys (*Survey 3;* sampling rate 1:200, desktop and mobile, 6 days, about 10,500 responses). It comprised three questions with the following answer options in random order (the first two questions also offered "other" as an answer, with the option to enter free-form text):

- *I am reading this article because...*: I have a work or school-related assignment; I need to make a personal decision based on this topic (*e.g.*, buy a book, choose a travel destination); I want to know more about a current event (*e.g.*, a soccer game, a recent earthquake, somebody's death); the topic was referenced in a piece of media (*e.g.*, TV, radio, article, film, book); the topic came up in a conversation; I am bored or randomly exploring Wikipedia for fun; this topic is important to me and I want to learn more about it (*e.g.*, to learn about a culture). Users could select multiple answers for this question.
- *I am reading this article to...*: look up a specific fact or to get a quick answer; get an overview of the topic; get an in-depth understanding of the topic.
- *Prior to visiting this article...*: I was already familiar with the topic; I was not familiar with the topic, and I am learning about it for the first time.

Only 2.3% of respondents used the "other" option, and hand-coding of the corresponding free-form responses did not result in new categories. We thus conclude that our categories are robust and use the resulting classification as our *taxonomy of Wikipedia readers* in the rest of this paper.

Table 1: *Features.* This table describes all features utilized in this work. *Survey features* capture responses to our survey questions; *request features* capture background information about the respondent mined from webrequest logs; *article features* describe the requested Wikipedia article; and *session/activity features* are derived from the entire reading session and beyond-session activity.

	feature	description
survey	motivation	Type of motivation for reading an article, as selected by respondent in survey. As multiple responses were allowed, we work with boolean dummy variables for each motivation.
	information need	Information need for reading an article, as selected by respondent in survey.
	prior knowledge	Prior knowledge about the topic before visiting the article, as selected by respondent in survey.
request	country	Country code of respondent in survey (e.g., USA) derived from the IP address.
	continent	Continent of respondent in survey (e.g., North America) derived from the IP address.
	local time weekday	Local weekday of survey request detected by timezone information (Mon-Sun).
	local time hour	Local hour of survey request detected by timezone information (0-24).
	host	Requested Wikipedia host: "desktop" (en.wikipedia.org), or "mobile web" (en.m.wikipedia.org).
	referer class	Referer class of request (none, internal, external, external search engine, or unknown).
article	article in-degree	The topological in-degree of an article.
	article out-degree	The topological out-degree of an article.
	article pagerank	The unnormalized pagerank of an article; calculated with damping factor of 0.85.
	article text length	The text length of an article as extracted from HTML—number of characters.
	article pageviews	The sum of pageviews for the article in same time period as survey.
	article topics	Probability vector for 20 topics as extracted by LDA from a bag-of-words representation. Topics were manually labeled as follows: (1) transportation & modern military, (2) biology & chemistry, (3) South Asia, Middle East, (4) mathematics, (5) 21st century, (6) TV, movies, & novels, (7) Britain & Commonwealth, (8) East Asia, (9) Spanish (stubs), (10) war, history, (11) geography (unions, trade), (12) literature, art, (13) education, government, law, (14) 20th century, (15) sports, (16) United States, (17) numbers, (18) technology, energy, & power, (19) music, and (20) geographical entities. We use the probabilistic topic distribution as 20 individual features for each article.
	article topic entropy	Measures the topic specificity of an article from LDA probability vector.
session/activity	session length	The number of requests within the session.
	session duration	Total time spent in the session in minutes.
	avg. time difference	Average time difference between subsequent session requests (i.e., dwelling time).
	avg. pagerank difference	Average pagerank difference between subsequent session requests (i.e., stating whether readers move to periphery or core).
	avg. topic distance	Average Manhattan distance between topic distributions for subsequent session requests (i.e., capturing topical changes).
	referer class frequency	For each referer class (see survey features): frequency in session.
	session article frequency	The number of times requested article for survey response occurs within the session.
	session position	Relative position inside a session when answering the survey.
	num. of sessions	The total number of sessions for respondent in survey time period.
	num. of requests	The total number of requests for respondent in survey time period.

3. DATASETS AND PREPROCESSING

Here, we describe utilized datasets and preprocessing.

3.1 Survey

To quantify the prevalence of the driving factors specified by our taxonomy, we ran an additional *large-scale survey* on English Wikipedia consisting of the same three questions on motivation, depth of information need, and prior knowledge as *Survey 3* (Sec. 2). The survey was run at a sampling rate of 1:50 from 2016-03-01 to 2016-03-08 on all requests to English Wikipedia's mobile and desktop sites. It was not shown on non-article pages (discussion pages, search pages, *etc.*), on the main page of Wikipedia, and to browsers with *Do not Track* enabled. Potential survey participants were identified by assigning a token to their browsers and eventually showing a widget with an invitation to participate in the survey. Once shown, the reader could ignore it, dismiss it, or opt in to participate which would take the reader to an external site (Google Forms), where she would see the three questions described in Sec. 2. A unique, anonymous ID was passed to Google Forms for each user, which would later be used to link the survey responses to users' webrequest logs (Sec. 3.2). A privacy and consent statement[1] providing details about the collection, sharing, and usage of the survey data was shown to all users prior to submitting their responses. Overall, our dataset consists of survey answers from 29,372 participants after basic data cleaning such as removing duplicate answers from the same users. Whenever we write "survey" throughout the rest of this paper, we refer to the survey described here.

[1] `https://wikimediafoundation.org/wiki/Survey_Privacy_Statement_for_Schema_Revision_15266417`

3.2 Webrequest logs

Ultimately, we aim to to understand how users' motivation, desired depth of knowledge, and prior knowledge (*i.e.*, their answers to our survey) manifest themselves in their reading behavior. The data collected through the survey alone, however, does not provide any information on the respondent's behavior beyond the single pageview upon which the survey was presented.

In order to be able to analyze respondents' reading behavior in context, we connect survey responses to the webrequest logs maintained by Wikipedia's web servers, where every access to any Wikipedia page is stored as a record that contains, among others, the requested URL, referer URL, timestamp, client IP address, browser version, and city-level geolocation inferred from the IP address. Since the logs do not contain unique user IDs, we construct approximate user IDs by concatenating the client IP address and browser version; *cf.* discussion in Sec. 5.2.

As the information needs and reading behavior of the same user may change over time, we operate at an intermediate temporal granularity by decomposing a user's entire browsing history into *sessions,* where we define a session as a contiguous sequence of pageviews with no break longer than one hour [12]. To reconstruct the session in which a user took our survey (*cf.* Fig. 1), we retrieved from the webrequest logs all records with the user's (approximate) ID, ordered them by time, chunked them into sessions according to the aforementioned one-hour rule, and returned the session that contains the record with the specific URL and timestamp of the survey response.

3.3 Wikipedia article data

Different articles are consumed in different ways. Hence, the properties of articles viewed by survey respondents play an important role in our analysis. To extract these properties, we utilized the public dump of English Wikipedia released on 2016-03-05,[2] such that article revisions closely match those seen by survey participants.

The dump contains wiki markup, whereas browsers receive HTML code generated from this markup. Since the markup may contain templates that are expanded only upon conversion to HTML, some page content is not immediately available in markup form. In order to obtain a more complete representation, we retrieved the full HTML of the article contents using Wikimedia's public API. In addition to the textual article content, we extracted the network of articles (5.1M) connected by hyperlinks (370M).

3.4 Features

Throughout this work, we study features extracted directly from the survey responses (Sec. 3.1), from underlying webrequest logs of article requests and extracted sessions (Sec. 3.2), and background Wikipedia article data associated with requested articles (Sec. 3.3). We list and describe all features utilized in this work in Table 1.

For topic detection, we fit a Latent Dirichlet Allocation (LDA) [4] model on bag-of-words vectors representing articles' textual content (with stopwords removed) using online variational Bayes. To find a balance between complexity and interpretability, we decided to work with 20 topics. We assigned labels to topics by manually inspecting the topics' word distributions and their top Wikipedia articles.

3.5 Correcting survey bias via webrequest logs

The goal of this work is to study the motivations and behaviors representative of Wikipedia's entire reader population. However, deducing properties of a general population from surveying a limited subpopulation is subject to different kinds of biases and confounders, including *coverage bias* (inability to reach certain subpopulations), *sampling bias* (distortions due to sampling procedure), and *non-response bias* (diverse likelihood of survey participation after being sampled as a participant).

Consequently, an important step in our analysis is to account for potential biases in survey responses. Finding suitable adjustments has been a decade-long research effort in the survey methodology community [5]. Since strata methods such as poststratification [9] are less well suited to control for a large number of features, we opt for *inverse propensity score weighting* [2] as an alternative. This technique assigns control weights to each survey response, thus correcting bias with respect to a control group (Wikipedia population). The rationale behind this procedure is that answers of users less likely to participate in the survey should receive higher weights, as they represent a larger part of the overall population with similar features. For determining participation probabilities *(propensity scores),* we use gradient-boosted regression trees on individual samples to predict if they belong to the survey *vs.* the control group, using all features of Sec. 3.4. (We provide additional methodological details in the appendix.) By using background features (*e.g.*, country, time) plus digital traces (*e.g.*, sessions), and by building a representative control group, we have an advantage over traditional survey design, which is often limited to few response features such as gender and age, as well as to small control groups.

When discussing results in the next section, we shall see (Fig. 2) that our weight adjustment changes the relative shares of survey responses only slightly, with general trends staying intact. Hence, we shall utilize only weighted survey responses for inference on statistical properties from this point on. Additionally, we use the so-called *effective sample size* (*cf.* appendix) when calculating standard errors, confidence intervals, and statistical tests, in order to account for differing standard errors of weighted estimators.

4. RESULTS: WHY WE READ WIKIPEDIA

This section discusses results on why users read Wikipedia.

4.1 Survey results

We start with a discussion of the responses to our survey.

Survey responses. First, we examine the percentages of survey respondents with specific motivations, information needs, and prior knowledge. We visualize the results in Fig. 2, focusing on the green (right) bars representing weighted survey responses (sorted by popularity).

With respect to *motivation,* we find that Wikipedia is consulted in a large spectrum of use cases and that no clearly dominant motivation can be identified. Prominently, extrinsic situations trigger readers to visit Wikipedia to look up a topic that was referenced in the media (30%), came up in a conversation (22%), is work or school-related (16%), or corresponds to a current event (13%). At the same time, readers have intrinsic motivations, such as wanting to learn something (25%), being bored (20%), or facing a personal decision (10%). We also find that the "other" option was only rarely selected, further confirming the robustness of the taxonomy of readers introduced in Sec. 2.

The results also show that Wikipedia is visited to satisfy different kinds of *information needs.* Interestingly, shallow information needs (overview [39%] and quick fact-checking [38%]) appear to be more common than deep information needs (21%). As for *prior knowledge,* we observe nearly identical shares of readers being familiar (50%) *vs.* unfamiliar (47%) with the topic of interest.

Survey response correlations. Next, in Table 2, we study whether certain combinations of motivations, information needs, and prior knowledge occur more frequently than expected, quantified by the *lift, i.e.*, the ratio between observed and expected frequencies.

Table 2a suggests that different *motivations* are coupled with different *information depths.* Specifically, in-depth in-

[2] `https://archive.org/details/enwiki-20160305`

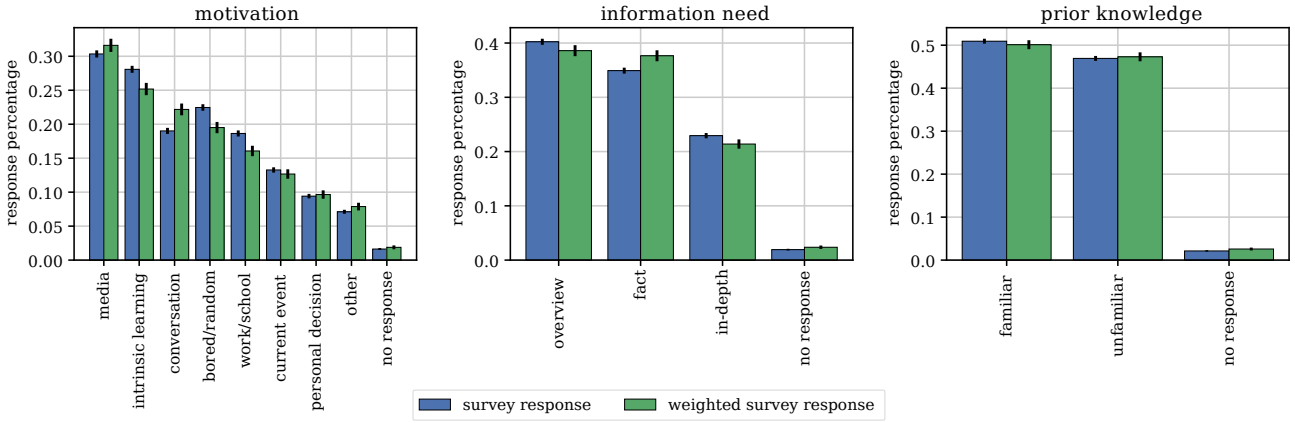

Figure 2: ***Survey responses.*** **This figure visualizes the share of answers for the three parts of the user survey: motivation, information need, and prior knowledge. The blue bars (left) reflect the raw unweighted responses, and the green bars (right) depict the bias-corrected weighted responses (propensity score weight adjustment). Error bars visualize the 95% confidence intervals using effective sample size for the weighted responses. In general, results suggest popularity of both extrinsic and intrinsic motivation, as well as high, but balanced, relevance for certain information needs and prior knowledge. The results based on bias-correction weighting reflect minor changes in responses without drastically changing the general direction.**

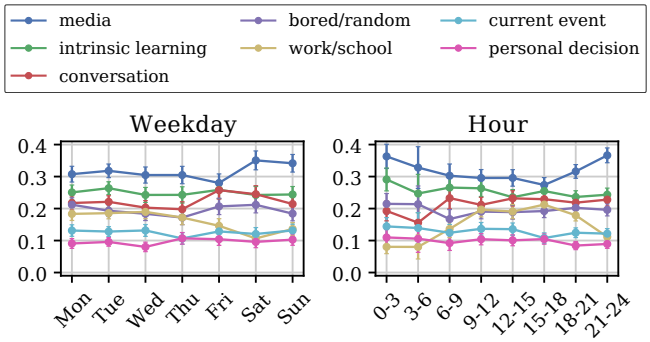

Figure 3: ***Motivation day and time.*** **This figure visualizes how the relative share of motivation (y-axes) changes over the course of a week and the course of a day (x-axes). Error bars depict the 95% confidence interval with effective sample size.**

formation needs prevail when readers are driven by intrinsic learning (lift 1.62); quick fact look-ups are associated more strongly with conversation and work/school motivations than one would expect *a priori;* and gaining an overview of a topic appears to be especially important for readers motivated by media coverage and for the bored.

In Table 2b, we find weaker correlations between *motivations* and levels of *prior knowledge,* apparent from lifts closer to 1 and a lack of significance. However, certain trends still emerge; *e.g.*, when readers research a topic from the media, they are more likely to be unfamiliar with the topic (lift 1.22). In contrast, readers whose goal is learning are more likely to be familiar with the topic (lift 1.14).

As a corollary of the above correlations, we also observe patterns when contrasting *prior knowledge* with *information need* (Table 2c). We find that familiar readers are more likely to look up quick facts (lift 1.13) and aim at getting in-depth knowledge about a topic (lift 1.15) than one would expect. Contrarily, unfamiliar readers are more likely to first aim at getting an overview of the topic (lift 1.22) instead of directly going into depth (lift 0.87).

Survey responses over time. Next, we study how the prevalence of motivations, information needs, and prior knowledge changes over time. For *motivations,* shown in Fig. 3, we find relatively stable trends over the course of a week or day. Three notable exceptions, however, emerge. First, on weekends (Saturday, Sunday) and at night, there is a higher share of readers who are led to Wikipedia by media coverage; this is potentially due to a higher likelihood of being exposed to media during these time periods. Similarly, conversations play a more important role on Fridays and Saturdays, possibly since people go out, meet with friends, and are involved in conversations that lead to consulting Wikipedia. By contrast, reading an article for work or school reasons has a relatively lower share towards the weekend, but peaks at daytime hours, probably because people work and go to school on working days and during daytime hours.

Additionally, results on *information need* show overall quite stable trends over a week and over a day without clear outliers, also due to larger confidence intervals (results not visualized). For *prior knowledge,* we identify small upward trends on weekends and evening hours for already being familiar with the topic, compared to being unfamiliar. However, error bars are again too large to justify stronger claims.

4.2 Webrequest-log results

Our previous results suggest that Wikipedia is visited for a variety of use cases that differ not only in their motivation triggers, but also in the depth of information needs, and readers' prior familiarity with the topic. In this section, we investigate correlations of survey responses with behavioral patterns based on request, article, and session features (Sec. 3.4). In doing so, we reveal characteristic differences and develop stereotypes for motivational groups.

Methodology. Due to our large set of features at interest (Sec. 3.4), we investigate behavioral reader patterns based on rule mining techniques, specifically *subgroup discovery* [14, 19]. The general goal of subgroup discovery is to find descriptions of subsets of the data that show an interesting (*i.e.*, significantly different) distribution with respect to a

predefined target concept from a large set of candidates. In our scenario, we perform a series of subgroup searches, each using one survey answer option as the target. To create the search space of candidate subgroup descriptions, we use all features described in Sec. 3.4. For the topic features, we consider a topic as present in an article viewed by a user if our topic model provided a probability for this topic above 20%. Other numeric features are binarized in five intervals using equal-frequency discretization. Due to missing values and multiple occurrences of values, bin sizes can significantly deviate from 20% of the dataset for some features. To select the most interesting subgroups, we use the *lift* as a quality function [10]. This measure is computed as the ratio between the likelihood of a survey answer in the subgroup and the respective likelihood in the overall dataset. As an example, a lift of 1.3 means that the respective survey answer is 30% more likely to occur in the subgroup than in the overall data. Additionally, we apply a filter to remove all subgroups that could not be shown to be significant by a χ^2 test with a Bonferroni-corrected threshold of $\alpha = 0.05$.

As a result, we obtain a list with the top k interesting subgroups for each survey answer T. For each subgroup S, we can compute various statistics: the (relative) *size* $P(S)$ of the subgroup, *i.e.*, the share of users that are covered by the subgroup description, the share $P(S|T)$ of subgroup users among those who answered with T in the survey, the *target share* $P(T|S)$ in the subgroup, *i.e.*, the share of users within the subgroup that reported the respective answer, and the *lift,* which is defined as $P(T|S)/P(T) = P(S|T)/P(S)$. Note that the absence of a feature in the discussion does not mean that it was not considered, but that it is not among the most significant subgroups.

Table 2: *Survey response correlations.* Each cell depicts the row-normalized share of responses that have also selected a given column as answer (without "other" and non-responses). The bottom rows highlight the overall share of responses for a given column as expectation. Values in brackets reflect the lift ratio of observed *vs.* expected frequency. The last column indicates significance (* < 0.001, ** < 0.01, * < 0.05) for the hypothesis test of independence of observed frequencies (contingency table with row frequencies and complement of all other rows) and expected frequencies (as in the last table row) using a χ^2 test using the effective sample size.**

(a) Motivation vs. information need

information need motivation	fact	in-depth	overview	sig.
media	0.38 (1.00)	0.19 (0.87)	0.43 (1.12)	***
intrinsic learning	0.29 (0.76)	0.35 (1.62)	0.35 (0.92)	***
conversation	0.43 (1.13)	0.20 (0.93)	0.36 (0.94)	***
bored/random	0.31 (0.83)	0.23 (1.05)	0.45 (1.17)	***
work/school	0.39 (1.04)	0.23 (1.09)	0.36 (0.93)	
current event	0.36 (0.95)	0.28 (1.30)	0.35 (0.92)	***
personal decision	0.32 (0.85)	0.29 (1.35)	0.38 (0.97)	***
response perc.	0.38	0.21	0.39	

(b) Motivation vs. prior knowledge

prior knowledge motivation	familiar	unfamiliar	sig.
media	0.42 (0.83)	0.58 (1.22)	***
intrinsic learning	0.57 (1.14)	0.41 (0.87)	***
conversation	0.49 (0.98)	0.49 (1.04)	***
bored/random	0.53 (1.07)	0.45 (0.95)	
work/school	0.52 (1.04)	0.46 (0.97)	
current event	0.52 (1.03)	0.46 (0.98)	
personal decision	0.50 (0.99)	0.48 (1.02)	
response perc.	0.50	0.47	

(c) Prior knowledge vs. information need

information need prior knowledge	fact	in-depth	overview	sig.
familiar	0.43 (1.13)	0.25 (1.15)	0.32 (0.83)	***
unfamiliar	0.34 (0.90)	0.19 (0.87)	0.47 (1.22)	***
response perc.	0.38	0.21	0.39	

Motivation. We start with characterizing groups with specific *motivations* as reported in the survey. In particular, we provide detailed results for two exemplary motivational groups (work/school and bored/random; Table 3) and only shortly summarize results for other motivations.

Users who intend to use Wikipedia for work or school are more frequently observed for specific topics of articles, namely war & history, mathematics, technology, biology & chemistry, and literature & arts. For the first two of these topics, users are more than twice as often motivated by work or school tasks as on average. While these topics cover a wide range of different areas, all of them are more related to academic or professional activities than for leisure. Additionally, this type of motivation is more often reported by users accessing Wikipedia's desktop version. This could be expected since many work/school activities are performed in office settings. Furthermore, we can see that this motivation occurs more often for users who are referred by external search engines multiple times in a session, and by users who stay longer on an individual page, which can be seen as a potential indicator for intensive studying.

By contrast, users who describe their motivation as bored/random, are more likely to use internal navigation within Wikipedia and to spend only little time on the individual articles. Also, they tend to switch topics between the individual articles more often (as indicated by the subgroup with a high average topic distance). These are telltales for less focused browsing behavior. Bored users also view more articles on Wikipedia both within the survey session and overall during the respective week. Finally, this motivation can also be observed more frequently for articles that cover specific topics, such as sports, 21st century, and TV, movies, & novels. Clearly, these topics are more leisure-oriented and are in stark contrast to the previously discussed topics favored by users who use Wikipedia for work or school.

Due to limited space, we only outline findings for other motivations: For example, motivation via media is significantly more often observed for the topics TV, movies, & novels (lift 1.37) and 21st century (lift 1.26), for popular articles, *i.e.*, articles with a high number of pageviews (lift 1.17), and for articles in the periphery of the Wikipedia link network according to pagerank (lift 1.14). The motivation of looking up something that came up in a conversation is more frequently reported for users with a single Wikipedia article request within a session (lift 1.08) and for users of the mobile version of Wikipedia (lift 1.08). The current-event motivation is more likely for articles about sports (lift 1.97),

Table 3: ***Top subgroups for the motivations "work/school" and "bored/random".*** **Each table shows the top subgroups with significantly different shares of users with a certain motivation T. For each subgroup S, we display the relative size $P(S)$ of the subgroup (*i.e.*, the share of users covered by the subgroup description), the share $P(S|T)$ of the subgroup among those with motivation T, the target share $P(T|S)$ in the subgroup, and the *lift* measure, defined as $P(T|S)/P(T) = P(S|T)/P(S)$. Rows are ranked by lift. The last column indicates significance (*** < 0.001, ** < 0.01, * < 0.05) for the hypothesis test of independence between subgroup and target motivation using a χ^2 test with effective sample size and Bonferroni correction.**

(a) T: "motivation = work/school"; $P(T) = 19.5\%$

Subgroup S	$P(S)$	$P(S\|T)$	$P(T\|S)$	lift	sig.
topic (mathematics)	7.9%	17.1%	34.8%	2.17	***
topic (war, history)	4.4%	9.6%	34.7%	2.16	***
topic (technology)	13.2%	23.7%	28.8%	1.79	***
topic (biology, chemistry)	8.6%	14.0%	26.2%	1.63	***
host = desktop	35.5%	57.8%	26.1%	1.63	***
article pagerank $\geq$ 9.98	20.0%	32.4%	26.1%	1.62	***
avg. time difference $\geq$ 9.40	7.7%	11.5%	24.0%	1.50	***
avg. pagerank difference < -4.35	7.6%	11.2%	23.6%	1.47	***
topic (literature, art)	10.1%	14.7%	23.5%	1.46	***
avg. time difference: [3.60:9.40[	7.7%	11.0%	23.1%	1.44	***
num. (referer=search) $\geq$ 2	20.5%	28.5%	22.4%	1.39	***
session duration $\geq$ 6.60	18.0%	24.2%	21.6%	1.34	***

(b) T: "motivation = bored/random"; $P(T) = 16.1\%$

Subgroup S	$P(S)$	$P(S\|T)$	$P(T\|S)$	lift	sig.
referer class: internal	9.4%	14.0%	29.0%	1.49	***
num. of requests $\geq$ 8	11.8%	16.6%	27.5%	1.41	***
topic (sports)	5.9%	8.0%	26.1%	1.34	**
num. (referer=internal $\geq$ 1	17.1%	22.7%	25.9%	1.33	***
session position: [0.33:0.75[	7.5%	9.8%	25.6%	1.31	**
avg. topic distance $\geq$ 1.08	7.5%	9.8%	25.2%	1.29	*
topic (21st century)	25.1%	32.1%	25.0%	1.28	***
session length $\geq$ 3	22.2%	28.3%	24.8%	1.27	***
avg. time difference: [0.68:1.56[	7.7%	9.7%	24.7%	1.27	*
num. (referer=none) $\geq$ 2	9.7%	12.2%	24.5%	1.26	*
topic (tv, movies, novels)	34.1%	41.4%	23.7%	1.21	***
# article pageviews $\geq$ 63606	19.8%	23.5%	23.1%	1.19	**

21st century (lift 1.49), and education, government, & law (lift 1.49). It is also more common for articles with many page views (lift 1.68), possibly because articles on current events are trending. Users who aim at intrinsic learning show a topic preference for more scholarly topics such as literature & art (lift 1.30), mathematics (lift 1.24), and technology (lift 1.21). Finally, the geographical origin of a user also has an effect: the motivations personal decision, current event, and intrinsic learning are reported significantly more often for users from Asia (mostly India; lifts 1.46, 1.44, and 1.20).

Information need. Overall, the investigated subgroups are more homogeneous with respect to the reported information need. We can, however, find some notable (anecdotal) exceptions: Users from Asia describe their *information needs* significantly more often as acquiring in-depth information (lift 1.51). For users who want to obtain an overview of a topic, using the desktop version of Wikipedia is more common than for the average user (lift 1.13) Also, topics play a certain role: fact look-ups, for example, are more often observed for the sports topic (lift 1.08). Session features that describe user behavior across multiple page visits do not lead to any significant differences in information need.

Prior knowledge. Regarding readers' *prior knowledge,* we can observe that users feel familiar with topics that are more spare-time oriented, such as sports (lift 1.21), 21st century (lift 1.08), and TV, movies, & novels (lift 1.07). They also feel more familiar about articles that are popular, *i.e.*, have many pageviews (lift 1.11), are longer (lift 1.10), and are more central in the link network (out-degree, in-degree, or pagerank; lifts 1.11, 1.09, and 1.08). Naturally, the answer "unfamiliar" is more often reported for the exact opposite of these subgroups. Features that describe a user behavior over multiple article views do not lead to significant deviations.

4.3 Summary of results

Prevalence of use cases. We have shown that Wikipedia is read in a wide variety of use cases that differ in their motivation triggers, the depth of information needs, and readers' prior familiarity with the topic. There are no clearly dominating use cases, and readers are familiar with the topic they are interacting with as often as they are not. Wikipedia is used for shallow information needs (fact look-up and overview) more often than for deep information needs. While deep information needs prevail foremost when the reader is driven by intrinsic learning, and fact look-ups are triggered by conversations, we saw that overviews are triggered by bored/random exploration, media coverage, or the need for making a personal decision.

Use cases over time. Motivations appear to be mostly stable over time (days of the week and hours of the day), with a few exceptions: motivations triggered by the media are increased over the weekends and at nights, conversation triggers are increased over the weekends, and work/school triggers are increased on week days and during the day.

Behavioral patterns. By connecting survey responses with webrequest logs, we identified certain behavioral patterns:

- When Wikipedia is used for work or school assignments, users tend to use a desktop computer to engage in long pageviews and sessions; sessions tend to be topically coherent and predominantly involve central, "serious" articles, rather than entertainment-related ones; search engine usage is increased; and sessions tend to traverse from the core to the periphery of the article network.
- Media-driven usage is directed toward popular, entertainment-related articles that are frequently less well embedded into the article network.
- Intrinsic learning tends to involve arts and science articles with no significant navigational features; conversations bring infrequent users to Wikipedia, who engage in short interactions with the site, frequently on mobile devices.
- People who use Wikipedia out of boredom or in order to explore randomly tend to be power users; they navigate Wikipedia on long, fast-paced, topically diverse link chains; and they often visit popular articles on entertainment-related topics, less so on science-related topics.
- Current events tend to drive traffic to long sports and politics-related articles; the articles tend to be popular, likely because the triggering event is trending.
- When Wikipedia is consulted to make a personal decision, the articles are often geography and technology-related, possibly due to travel or product purchase decisions.

5. DISCUSSION

Every day, Wikipedia articles are viewed more than 500 million times, but so far, very little has been known about the motivations and behaviors of the people behind these pageviews. The present study is the first comprehensive attempt to help us understand this group of users by combining a survey with a log-based analysis.

The work most closely related to ours is by Lehmann *et al.* [26], who extracted Wikipedia navigation traces from Yahoo! toolbar logs (which may be considered a biased sample of the complete logs we have access to) with the goal of discovering a set of usage patterns according to which articles are consumed. Using clustering techniques, they concluded that there are four types of articles: trending articles, articles read in a focused manner, articles read by exploring users, and articles users just quickly pass through. Lehmann *et al.*'s work is entirely "unsupervised", in the sense that they have no ground truth of the actual underlying user motivations and needs.

We, on the contrary, have elicited the ground truth through our survey and can thus arrive at stronger and more actionable conclusions, which we discuss next. We do so by first highlighting implications and directions for future work (Sec. 5.1), and then reflecting on our methodology and pointing out its limitations (Sec. 5.2).

5.1 Implications and future directions

This research has already had considerable impact within the Wikimedia Foundation, where it has informed several items on the product development agenda, and we hope that it will further inspire Wikimedia developers, academic researchers, and volunteers to build tools for improving the user experience on Wikipedia.

Predicting motivation and desired depth of knowledge. A tool immediately suggested by our results could involve statistical models for real-time inference of user session motivations from behavioral traces as captured in the webrequest logs. Such models could be trained in a supervised fashion with features of Sec. 3.2 as input, and survey responses as output, and could form the basis for products and services for supporting the needs of Wikipedia readers more proactively. For instance, if an editor working on an article could be shown an estimate of the distribution of the motivations and desired depths of knowledge on behalf of the readers of the article, she can take this information into account to tailor the content to the needs of the audience or attempt to change the distribution of the audience's motivation by creating specific types of content in the article. Such a tool could have large impact, considering that, currently, editors contribute to content on Wikipedia without much knowledge of the users who will eventually read it.

Similarly, predicting the distribution over depths of knowledge sought by the readers of an article could offer opportunities for creating different versions the article, *e.g.*, for those who are interested in quick look-ups *vs.* in-depth readers. This could enhance the usability of Wikipedia articles particularly on mobile devices with smaller screens and low-bandwidth connections.

The above task of using digital traces to predict survey responses has been called *amplified asking,* and it is known to be difficult [35]. This has been confirmed by our preliminary attempts, where we have achieved accuracies only slightly better than simple baselines. This may be partly explained by the fact that user motivations may change during a session, and while the survey captures the motivations at the article level accurately, it fails to capture possible transitions between motivations during a session. For instance, a session might start with a school or work project in mind, but the user might then transition to procrastinating by exploring Wikipedia randomly, which would not be captured in our current setting. Also, prediction is complicated by the fact that, even for a fixed article, user motivations might vary widely. For instance, of the 222 users taking the survey upon reading the article about Donald Trump, 38% read the article out of boredom, 32% in response to media coverage, 24% because of a conversation, 23% due to current events, 17% because the topic was personally important to them, *etc.*

Despite these difficulties, future work should investigate the problem of predicting user intentions in more depth.

5.2 Methodological limitations

We discuss certain limitations of present research next.

Survey selection bias. A general caveat with surveys is that one typically cannot guarantee that whether a subject participates or not is a fully random choice. Certain covariates may be associated with both participation rates and responses given, leading to biased conclusions. We made a best effort to correct for this bias by adjusting survey responses based on a random sample of all Wikipedia pageviews drawn from Wikipedia's webrequest logs (Sec. 3.5). However, if the bias-inducing covariates are hidden, one cannot fully correct for the bias. For instance, young users might be both more prone to use Wikipedia for work or school and to participate in our survey; this would over-represent the work/school motivation in our raw survey results, and since we have no information about users' age, we could not correct for this bias. Apart from that, survey answers might be biased by social desirability [7]; *e.g.*, even in an anonymous survey, users might be reluctant to admit they are visiting Wikipedia out of boredom.

Unique visitors and level of analysis. Wikipedia does not require users to log in, nor does it use cookies in webrequest logs to maintain a notion of unique clients. Hence, we need to rely on an approximate notion of user IDs based on IP addresses and browser versions (Sec. 3.2), which makes the attribution of pageviews to users and the construction of sessions imperfect. In particular, we might not recognize that two pageviews are by the same user if they use several devices or if their IP address changes for other reasons; and we might conflate several users if they share the same device or IP address (*e.g.*, via a proxy). Currently, we limit the impact of such errors by analyzing the data on a session level and operating at relatively short time scales (an inactivity of more than one hour ends the session being studied). If the attribution of pageviews to unique users becomes more precise in the future, we could study user behavior at longer time scales, which would, *e.g.*, allow us to understand and support long-term learning needs. Also, our current method aims at giving each user session equal weight. An alternative approach would be to analyze the data on a request level, which would put more emphasis on the motivations and needs of power users.

Cultural issues. The results discussed here pertain to the English edition of Wikipedia. Even within this limited scope, our behavioral analysis hints at subtle cultural and geographical differences; *e.g.*, the bored/random motivation

is particularly frequent in the U.S., whereas current events are a stronger motivator in India. Survey answers might also be influenced by different notions and associations of the survey phrasing across cultures [13]. Since Wikipedia strives to reach beyond cultural and linguistic boundaries, it is important to further investigate these cultural issues. As part of this effort, we are planning to repeat our study in additional language versions of Wikipedia to elicit cultural differences on a larger scale.

6. CONCLUSIONS

In this work, we study why users read Wikipedia. We use survey data to develop a taxonomy of Wikipedia usage along three dimensions: *motivation*, *information need*, and *prior knowledge*. In a large-scale survey with almost 30,000 participants, we quantify the share of readership for these driving factors. The bias-corrected survey results reveal a broad range of usage scenarios, interdependencies between survey answers, and temporal trends. Combining the survey responses with webrequest logs allows us to characterize motivational groups with behavioral patterns. The outcomes of this study are currently being discussed in the Wikimedia Foundation as a stimulus for developing specialized tools for readers and editors.

APPENDIX: SURVEY BIAS CORRECTION

This appendix covers the details of the survey bias correction.

Propensity score weight adjustment. We use *inverse propensity score weighting* to adjust for potential biases in survey response data with respect to control data [2, 27]. Specifically, we want to infer unbiased estimates of survey answers for the whole Wikipedia readership. Thus, we randomly sampled a large set of Wikipedia readers (25 times the number of survey responses) from the webrequests logs in the survey period. Then, we proceeded to sample one request for each selected user and marked it as an imaginary request reflecting a potential survey response; we also deduced the same set of features as for our survey (except responses). We only sampled requests that are desktop or mobile pageviews in English Wikipedia's main namespace and applied bot-filtering in order to match the original survey.

The propensity score of a single instance then reflects the probability that an instance with these control features (Sec. 3.4) participated in the survey. We approximate it using our control group. For that, a post-stratification approach [27] is infeasible due to the large number of control features we consider. Instead, we model the group membership (survey participant or control group) using gradient boosted regression trees showing promising results in the past in comparison to traditional approaches like logistic regression [24]. Given the features of an instance x, the model predicts a probability $p(x)$ that x belongs to the survey group. We then set the weight w for instance to $1/p(x)$. The rationale behind this procedure is that answers of users that are overall less likely to participate in the survey receive higher weights since they represent a larger part of the entire population with similar features.

Evaluating weights. To evaluate if applied weighting schemes have the intended correcting effect of making the user survey data more representative for the overall Wikipedia, we resort to two scenarios.

First, we check that the resulting weights do not contain drastic outliers dominating subsequent results, which would warrant so-called *trimming* [25]. In that regard, we observe that weights are sufficiently homogeneous distributed with a minimum of 1, a maximum of 190, a mean of 17.6, and a standard deviation of 26.9.

Additionally, we evaluate how well we can recover the mean value of features in the overall population from observed survey response features and our weighting scheme. For that purpose, we compute weighted and unweighted averages of the observed values for the survey users and compare them with the mean of a different random sample as a ground truth. As a result, the average of relative errors is reduced by 86%, from 0.556 in the unweighted case to 0.079 in the weighted case. The reduction is strongly significant ($p \leq 0.001$ according to a Wilcoxon signed rank test). If weighting is applied, then the mean recovered from the weighted survey is never more than 0.2 standard deviations off compared to the actual feature mean in the sample.

Effective sample size. Int this work, we employ a variety of statistical techniques on the survey data. Yet, the introduction of sample weights for correcting bias in survey responses leads to violations of IID assumptions [28]. Thus, standard errors of estimators are estimated as too small, which in turn leads to confidence intervals being too narrow and statistical tests asserting significance too often if standard procedures are applied. The extent to which the sampling error in the survey for some parameter θ deviates from the expected error from an IID sample due to survey design and correction, is known as the *design effect* (deff) [17]. If the design effect deviates from 1—as it is the case in our survey—then our understanding of sample size for calculating standard errors becomes incorrect. To that end, we consider the *effective sample size* estimating the required sample size of a random sampling survey for achieving the same error as the weighted sample—it is defined as $n_{\text{eff}} = n/\text{deff}$. As we cannot directly calculate deff without knowing the expected sampling error, we use Kish's approximation formula with weights w_i [17]:

$$n_{\text{eff}} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2}$$

For our complete survey data, $n_{\text{eff}} = 8839$. We use this effective sample size throughout this article for calculating standard errors, confidence intervals, and statistical tests. Note that this makes reported confidence interval and statistical hypothesis tests overly careful. For further details, please refer to [28].

Acknowledgements. We thank Dario Taraborelli from Wikimedia Foundation who was indispensable to the early phases of the project. We also thank Jonathan Morgan for helping us with the hand-coding; Jon Katz and Toby Negrin for helping us shape the direction of the research and supporting us throughout; Anne Gomez, Jeff Hobson, Bahodir Mansurov, Jon Robson, and Sam Smith for running the surveys on Wikipedia; and Aeryn Palmer for creating the privacy statements for this research. This research has been supported in part by NSF IIS-1149837, ARO MURI, DARPA NGS2, and Stanford Data Science Initiative.

References

[1] O. Arazy, H. Lifshitz-Assaf, O. Nov, J. Daxenberger, M. Balestra, and C. Cheshire. On the "how" and "why" of emergent role behaviors in Wikipedia. In *Conference on Computer-Supported Cooperative Work and Social Computing*, 2017.

[2] P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

[3] A. Basu. Context-driven assessment of commercial web sites. In *International Conference On System Sciences*, 2003.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[5] J. M. Brick. Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3):329–353, 2013.

[6] A. Broder. A taxonomy of web search. In *ACM SIGIR Forum*, 2002.

[7] T. J. DeMaio. Social desirability and survey. *Surveying Subjective Phenomena*, 2:257, 1984.

[8] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *International Conference on Research and Development in Information Retrieval*, 2010.

[9] A. Gelman and J. B. Carlin. Poststratification and weighting adjustments. In *CiteSeerX*, 2000.

[10] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3):9, 2006.

[11] S. Goel, J. M. Hofman, and M. I. Sirer. Who does what on the Web: A large-scale study of browsing behavior. In *International Conference on Web and Social Media*, 2012.

[12] A. Halfaker, O. Keyes, D. Kluver, J. Thebault-Spieker, T. Nguyen, K. Shores, A. Uduwage, and M. Warncke-Wang. User session identification based on strong regularities in inter-activity time. In *International Conference on World Wide Web*, 2015.

[13] J. A. Harkness, F. J. Van de Vijver, P. P. Mohler, et al. *Cross-cultural survey methods*. Wiley-Interscience Hoboken, 2003.

[14] F. Herrera, C. J. Carmona, P. González, and M. J. Del Jesus. An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems*, 29(3):495–525, 2010.

[15] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Workshop on Web Mining and Social Network Analysis*, 2007.

[16] D. Jurgens and T.-C. Lu. Temporal motifs reveal the dynamics of editor interactions in Wikipedia. In *International Conference on Web and Social Media*, 2012.

[17] L. Kish. *Survey sampling*. John Wiley and Sons, 1965.

[18] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *Conference on Computer Supported Cooperative Work*, 2008.

[19] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. American Association for Artificial Intelligence, 1996.

[20] S. Krug. *Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability*. New Riders, 2014.

[21] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *International Conference on World Wide Web*, 2010.

[22] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *International Conference on World Wide Web*, 2010.

[23] D. Lamprecht, D. Dimitrov, D. Helic, and M. Strohmaier. Evaluating and improving navigability of Wikipedia: A comparative study of eight language editions. In *International Symposium on Open Collaboration*, 2016.

[24] B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, 2010.

[25] B. K. Lee, J. Lessler, and E. A. Stuart. Weight trimming and propensity score weighting. *PloS One*, 6(3):e18174, 2011.

[26] J. Lehmann, C. Müller-Birn, D. Laniado, M. Lalmas, and A. Kaltenbrunner. Reader preferences and behavior on Wikipedia. In *Conference on Hypertext and Social Media*, 2014.

[27] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

[28] P. Mukhopadhyay. *Complex Surveys: Analysis of Categorical Data*. Springer, 2016.

[29] O. Nov. What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64, 2007.

[30] C. Okoli, M. Mehdi, M. Mesgari, F. Å. Nielsen, and A. Lanamäki. The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. *SSRN 2021326*, 2012.

[31] A. Paranjape, R. West, L. Zia, and J. Leskovec. Improving website hyperlink structure using server logs. In *International Conference on Web Search and Data Mining*, 2016.

[32] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Characterizing and modeling the dynamics of online popularity. *Physical Review Letters*, 105(15):158701, 2010.

[33] D. E. Rose and D. Levinson. Understanding user goals in web search. In *International Conference on World Wide Web*, 2004.

[34] T. Ryan and S. Xenos. Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*, 27(5):1658–1664, 2011.

[35] M. J. Salganik. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, 2017.

[36] P. Singer, D. Helic, B. Taraghi, and M. Strohmaier. Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *PloS One*, 9(7):e102070, 2014.

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

[38] A. Strauss and J. Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, 1998.

[39] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, 2008.

[40] V. Waller. The search queries that took Australian Internet users to Wikipedia. *Information Research*, 16(2), 2011.

[41] I. Weber and A. Jaimes. Who uses web search for what: and how. In *International Conference on Web Search and Data Mining*, 2011.

[42] R. West and J. Leskovec. Human wayfinding in information networks. In *International Conference on World Wide Web*, 2012.

[43] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *Conference on Information and Knowledge Management*, 2009.