Different Topic, Different Traffic: How Search and Navigation Interplay on Wikipedia

Dimitar Dimitrov[1], Florian Lemmerich[2], Fabian Flöck[3] and Markus Strohmaier[4]

[1] *GESIS – Leibniz Institute for the Social Sciences, dimitar.dimitrov@gesis.org*

[2] *RWTH Aachen University, florian.lemmerich@cssh.rwth-aachen.de*

[3] *GESIS – Leibniz Institute for the Social Sciences, fabian.floeck@gesis.org*

[4] *RWTH Aachen University & GESIS – Leibniz Institute for the Social Sciences, markus.strohmaier@cssh.rwth-aachen.de*

ABSTRACT

As one of the richest sources of encyclopedic information on the Web, Wikipedia offers large-scale article access data that allows us to compare articles with respect to the two main paradigms of information seeking, *i.e.*, search by formulating a query, and navigation by following hyperlinks. Using such data from the English Wikipedia, we study access behavior by employing two main metrics, namely (i) searchshare – the relative amount of views an article received by search –, and (ii) resistance – the ability of an article to relay traffic to other Wikipedia articles – to characterize articles. We demonstrate how articles in distinct topical categories differ substantially in terms of these properties. For example, architecture-related articles are often accessed through search and are simultaneously a "dead end" for traffic, whereas historical articles about military events are mainly navigated. We further link traffic differences to varying network, content, and editing activity features. Lastly, we measure the impact of the article properties by modeling access behavior on articles with a gradient boosting approach and explore explicit importance of individual features. Our results constitute a step towards understanding human information seeking behavior, and may contribute to identify focal points for future improvements of Wikipedia and similar systems.

Keywords: Search Behavior, Navigation Behavior, Log Analysis, Wikipedia

ISSN 2332-4031; DOI 10.34962/jws-71

1 Introduction

Before the age of the World Wide Web (Berners-Lee *et al.*, 2000), information was predominantly consumed in a linear way, *e.g.*, starting at the first page of a book and following the laid out narrative until the end. With the introduction of hypertext (Nelson, 1965) in digital environments, the way people consume information changed dramatically (Leu *et al.*, 2012; Coiro and Dobler, 2007; Leu *et al.*, 2005; Mangen, 2008). While on early websites, users still predominantly visited a main page through a fixed address and were sometimes even bounded by a more directory-like navigation structure, the rise of search engines and tighter interlinking of websites have corroded the linear consumption paradigm even further. Today, users access a single website through a multitude of webpages as entry points and can usually choose from numerous paths through the available linked content at any time. In such a setting, understanding at which (kind of) pages users typically begin and end their journey on a given website, vs. which pages relay traffic internally from and to these points, provides several useful insights. On one hand, it has high practical importance since it provides the first and last contact opportunity; pages could be shaped to leverage their function as an entry point (*e.g.*, by prioritizing improvements of navigational guidance for these pages to retain visitors), or as an exit point (*e.g.*, by surveying visitors for their user experience before leaving, or by providing increased incentives to continue navigation). On the other hand, knowledge about entry, relay and exit points is also closely tied to the relation of the major information seeking strategies, *i.e.*, search and navigation: the first page visited in a session on a website is frequently reached via search engine results, after a query formulation, while navigation has been often used when the exact information need cannot be easily expressed in words (Furnas, 1997; Furnas *et al.*, 1987). Understanding under which circumstances search or navigation dominate the users' information seeking behavior can help in developing an agenda for improving the web content in order to optimize visitor rates and retention.

Scope and research questions. Information consumption on the Web has been of special interest to researchers since the Web's earliest days (Kumar and Tomkins, 2010; Kumar and Tomkins, 2009). While both search (Waller, 2011; McMahon *et al.*, 2017; Spoerri, 2007) and navigation (Dimitrov *et al.*, 2017; Gildersleve and Yasseri, 2018; Lehmann *et al.*, 2014; Lamprecht *et al.*, 2016; Lamprecht *et al.*, 2017) have been investigated thoroughly in related work, they were mostly looked at separately. Consequently, so far little is known about which parts and content types of a specific website (inter)act in which structural roles, begetting different information access patterns.

In this work, we analyze how these patterns manifest on the online encyclopedia Wikipedia. With more than 5 million

articles, Wikipedia is one of the primary information sources for many Web users and through its openly available pageview data provides an essential use case for studying information seeking behavior, as made apparent by numerous studies (Dimitrov *et al.*, 2016; Lamprecht *et al.*, 2017; Paranjape *et al.*, 2016). Yet, there is a lack of understanding how search and navigation as the two major information access forms *in combination* shape the traffic of large-scale hypertext environments, such as the world's largest online encyclopedia. To this end, we are interested in answering the following research questions: (i) How do search and navigation interplay to shape the article traffic on Wikipedia? Given an article, we want to know how its acting as a search entry point is related to (not) relaying navigation traffic into Wikipedia, and vice versa. This also addresses the issue of how search and navigation contribute to the article's popularity. Beyond these characteristics of the system in general, we also examine which specific properties of articles influence their roles in the search-vs-navigation ecosystem. We hence ask: (ii) Which article features (*i.e.*, topic, network, content and edit features) are indicative of specific information access behavior?

Materials, approach and methods. Building our analysis on large-scale, openly available log data for the English edition of Wikipedia, we propose two metrics capturing individual traffic behavior on articles, *i.e.*, (i) searchshare – the amount of views an article received by search –, and (ii) resistance – the ability of an article to channel traffic into and through Wikipedia (*cf.* Section 2).

We use searchshare and resistance to first explore the relation between search and navigation and their effect on the popularity of articles independent of their content (*cf.* Section 3). Depending on these two measures, we assign articles to four groups describing the role they assume for attracting and retaining visitors. Subsequently, we characterize the influence of several article attributes, including the general topical domain, edit activity and content structure on the preferred information access form (*cf.* Section 4). A fine-grained bow tie membership analysis of Wikipedia's traffic is also performed (*cf.* Section 4.5). Finally, we fit a tree-based gradient boosting model to determine the impact of article features on the preferred user access behavior (*cf.* Section 5).

Contributions and findings. Our contributions are the following: (i) Concerning the general (collective) access behavior on Wikipedia, we provide empirical evidence that for the most viewed articles, search dominates navigation regarding the number of articles accessed, and regarding received views. For the tail of the view distributions, navigation appears to become more and more important. (ii) We link article properties, *i.e.*, position in the Wikipedia network, number of article revisions, and topic to preferred access behavior, *i.e.*, search or navigation. Finally, (iii) we quantify the strength of the relationship between article properties and preferred access behavior.

Our analysis suggests that (i) while search and navigation are used to access and explore different articles, both types of information access are crucial for Wikipedia, and (ii) that exit points of navigation sessions are located at the periphery of the link network, whereas entry points are located at the core. (iii) Edit activity is strongly related with the ability of an article to relay traffic, and thus with the preferred access behavior.

Our results may have a variety of applications, *e.g.*, improving and maintaining the visual appearance and hyperlink structure of articles, identifying articles exhibiting changes in access behavior patterns due to vandalism or other online misbehavior. We consider our analysis as an initial step to better understand how search and navigation interplay to shape the user access behavior on platforms like Wikipedia and on websites in general.

Differences to original version. Please note that this paper is an invited journal version of the article "Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia" by Dimitrov *et al.* published at WebSci'18: 10th ACM Conference on Web Science (Dimitrov *et al.*, 2018). This version extends the original publication mainly in three parts: traffic entropy (*cf.* Section 3), bow tie analysis (*cf.* Section 4.5), and model understanding (*cf.* Section 5). The results of the newly added analyses corroborate the results of the original version.

2 Transition Data and Definitions

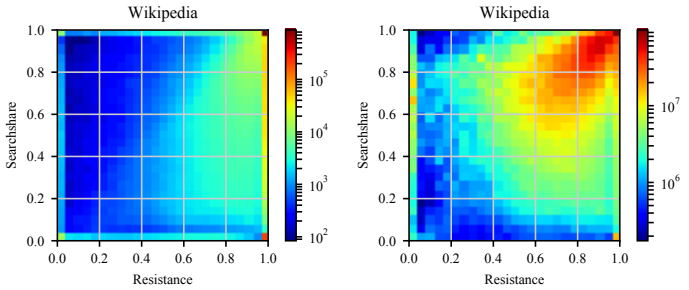
Below, we give an overview of the used dataset capturing the traffic on Wikipedia articles and define *searchshare* and *resistance* as our main metrics for describing the individual article traffic behavior.

2.1 Transition Data

For studying the access behavior on Wikipedia articles of the English language version, we use the clickstream dataset published by the Wikimedia Foundation (Wulczyn and Taraborelli, 2016). The used dataset contains the aggregated transition counts between webpages and Wikipedia articles in form of (*referrer, resource*) pairs extracted from the server logs for August, 2016, and is limited to pairs that occur at least 10 times. The referrer pages are either external (*e.g.*, search engines, social media), internal (other Wikipedia pages), or missing (*e.g.*, if the article is accessed directly using the browser address bar). The navigation targets are purely internal pages.[1] Since we are interested in contrasting Wikipedia article access from search engines and navigation (see also our discussion in Section 7), we focus our analyses only on those articles in the clickstream dataset that have received views through search or internal navigation, setting aside remaining view sources (mostly "no referrer"). Accordingly, we define *total views* of an article as the sum of all page accesses by either search or navigation.

The resulting dataset consists of 2,830,709 articles accessed through search 2,805,238,298 times and 14,405,839 transitions originating from 1,370,456 articles and accounting for 1,251,341,103 views of 2,149,104 target articles. In total, the dataset consists of 3,104,702 articles viewed 4,056,579,401 times, with a ratio of 69% stemming from search and 31% from internal navigation – in line with previous reports on the clickstream data (Dimitrov *et al.*, 2017; Lamprecht *et al.*, 2016).

[1] Leaving a Wikipedia page is treated as the end of the visit in the logs, whether by clicking on an external link or closing the page.



(a) Number of articles (b) Total views

Figure 1: Articles and article views by access behavior. For a given searchshare (y-axis) and resistance (x-axis), the figure shows (a) the number of articles and (b) the sum of their views in each heatmap square bin. Warm colors denote high values, using a logarithmic scale. We observe that search dominates navigation in terms of number of accessed articles (note the single top data bin in (a)) and that a substantial amount of articles exhibits high resistance values. When focusing on views, we see a more spread-out pattern, evidencing that a relatively small amount of articles attracts a substantial amount of search views and channels them onward to other articles (upper left side of (b)), corresponding to the *search-relay* group (*cf.* Table 1).

2.2 Definitions

To achieve a fundamental understanding of the parts that search and navigation each play for the distribution of views in Wikipedia, we take a look at the functional roles articles can assume for the overall traffic flow in respect to their *searchshare* and *resistance*.

Searchshare. A high *searchshare* value indicates that search is the predominant paradigm of accessing an article, and thus that the article acts as an *entry point* for a site visit. In contrast, articles with a low value receive most of their views from users visiting them by means of navigation. The *searchshare* metric is defined as

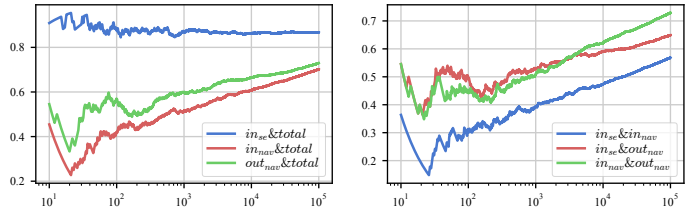
$$searchshare(a) = \frac{in_{se}(a)}{in_{se}(a) + in_{nav}(a)} \quad (1)$$

where $in_{se}(a)$ is the number of pageviews an article a received directly from search engine referrers, and $in_{nav}(a)$ is the number of views from navigation as recorded in the Wikipedia clickstream.

Resistance. A low *resistance* value signals that an article forwards most of its received traffic to other articles within Wikipedia, hence does not block the flow of incoming traffic onward. A high value in turn indicates that an article acts as an *exit point*. Thus, it rarely relays users to other Wikipedia articles. These articles are traffic sinks in the Wikipedia information network. We define the resistance metric as

$$resistance(a) = 1 - \frac{out_{nav}(a)}{in_{se}(a) + in_{nav}(a)} \quad (2)$$

where $out_{nav}(a)$ is the number of pageviews that had article a as a referrer. Additionally, we restrict the values to be in the interval [0,1]. This is necessary since a small number of articles generates more out-going traffic than they receive pageviews, *e.g.*, due to a user opening several links in a new tab each.



(a) Search and navigation vs. total (b) Search vs. navigation

Figure 2: Ranking overlap. Four rankings are shown, according to the total number of pageviews (*total*), the number of pageviews coming from search (in_{se}) as well as in- (in_{nav}) and out-navigation (out_{nav}). The y-axis indicates overlaps between pairs of rankings, considering the top-k articles of each ranking as marked on the x-axis (log-scaled, top articles on the left). As a result, the overall ranking of total pageviews shows a very high overlap with the incoming search ranking. The top pages by search and navigation differ substantially. Notably, being a distribution point of traffic (high out_{nav}, *cf.* (b)) is correlated most to receiving search, but only for top out_{nav} articles, with lower ranks being supplied with traffic predominantly through in_{nav}.

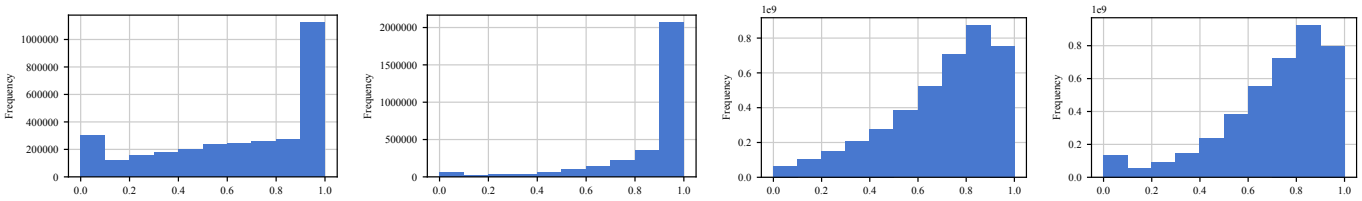
3 General Access Behavior

In this section, we investigate how exogenous and endogenous traffic contribute to article popularity on Wikipedia, and we study the distribution of traffic features. We provide a first overview of the general access behavior on Wikipedia regarding search and navigation, aided by a division of articles into four groups with respect to searchshare and resistance; in Section 4, we will subsequently take a deeper look at dissimilarities between different types of articles.

Search and navigation in relation to total views. As can be expected from related research on Wikipedia and similar online platforms, the distribution of pageviews over articles is long-tailed with a heavy skew towards the head (80% views generated by the top-visited 5.2% of all articles). To better investigate the relationship between search and (incoming and outgoing) navigation on the articles popularity, we calculate the cumulative overlap (intersection) of the descendingly ranked articles at each rank k, divided by k; this is an adaptation of the Rank Biased Overlap[2] measure.

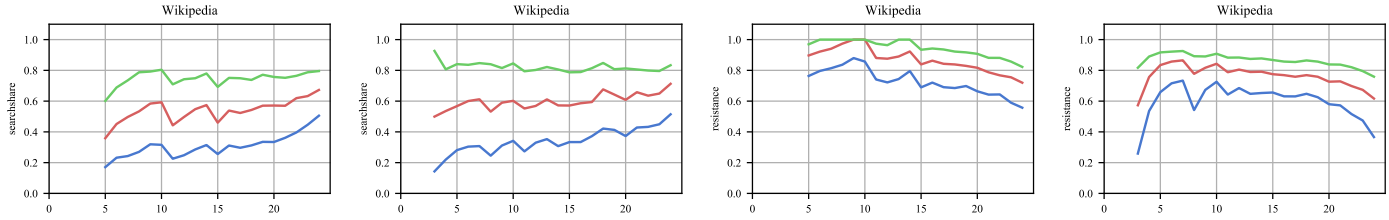
Figure 2(a) shows that the top k articles ordered by search traffic (top-k-search) are highly overlapping with the top articles by total views (top-k-total) at any k, underscoring the general

[2]Rank Biased Overlap (Webber *et al.*, 2010) is a common metric for similarity between rankings using cumulative set overlap in cases where the two lists do not necessarily share the same elements (as is the case here). Top-weighting as can be specified for RBO is neither suited nor necessary for the distinction of different k that we aim for here.



(a) Searchshare (b) Resistance (c) Searchshare-weighted (d) Resistance-weighted

Figure 3: Traffic feature distributions. Figures (a) and (b) show an unweighted histogram of searchshare and resistance, while (c) and (d) respectively weight articles by their pageview counts. Most articles have a very high value for searchshare and resistance. However, extreme values close to 1.0 in (a), (b) stem mostly from rarely visited pages.



(a) in-entropy vs. searchshare (b) out-entropy vs. searchshare (c) in-entropy vs. resistance (d) out-entropy vs. resistance

Figure 4: Traffic entropy. The figure shows the first (blue), second (red), and third (green) quartile of searchshare as function of articles' in- (a) and out-entropy (b), and resistance as function of articles' in- (c) and out-entropy (d). Entropy values are divided into 25 bins. For articles with high in- and out-entropy, we also observe high searchshare. The higher the in-entropy the lower the resistance. The resistance is the highest for an average out-entropy while it drops for extreme out-entropy values.

importance of search as a driver of incoming views. In-navigation, in contrast, is not a deciding factor to belong to the top most visited pages, but sees an extreme increase in the influence on overall views for articles up to top-k-total around 8000, at which point the increase continues, but levels off. Apparently, while search is the overall main driver for traffic, in-navigation rapidly becomes a more central source of traffic beyond the extremely popular articles. Turning to navigation passed on *from* articles to other articles, we can glean from Figure 2(a) that (i) while the very top of viewed articles contribute little in relation to their accumulated views to the internal traffic flow of Wikipedia (low overlap for out_{nav}&$total$), we (ii) see a rapid and constant drop in the amount of traffic "dying" at a given page with increasing top-k-total.

Further, while it is not surprising that the outgoing traffic accumulates generally in line with the overall received views, up until around top-k-total 1,500,000 it is generated at a rate *surpassing* the relative increase of total views, with the highest ranks of top-k-total contributing comparably little to it, just as to in-navigation. These observations are in line with Figure 2(b), where we see that a higher rank in receiving navigation - rather than from search - is more strongly correlated with distributing views to other articles for the largest portion of pages, after top-k-total 3000; up until that point, the largest share of channeled traffic stems from search views. As bottom line, we see a pattern that points to a small number of pages at the extreme top of the pageview counts that are mostly searched, but *in relation to their popularity* rather isolated in terms of navigation; with in- and out-navigation similarly gaining notably in correlation with overall views for lower top-k-total ranks.

Traffic feature distributions. Figure 3 depicts the system-wide distribution of searchshare and resistance. Pages are generally much more searched than navigated to (searchshare median = 0.74, mean = 0.66) as seen in Figure 3(a). It is also apparent from Figure 3(b) that most articles do not tend to forward much of their received traffic internally, with the median for resistance for all articles lying at 1.0 and the mean at 0.88. This general tendency prevails when these scores are weighted by their received views (Figures 3(c) and 3(d)), but a notably less skewed distribution emerges, implying that – even when accounting for regression-to-the-mean effects – a majority of views is acquired via search and that a majority of views hits rather high-resistance targets.

Relation between searchshare and resistance. We observe a light positive correlation (pearson = 0.26, spearman = 0.33) indicating that the more likely an article is used to start a session, the more likely it is also to be the last article accessed in a session. Figure 1 depicts this association for all articles in our dataset.

To explore this relation further, we assign each article to one of four groups, determined by the *mean* of both searchshare and

resistance as the thresholds.[3] We label each group according to its traffic behavior, *i.e.*, (i) *search-relay* articles that are often searched while simultaneously contributing to further navigation (above-mean searchshare, below-mean resistance); (ii) *search-exit* articles with above-average searchshare that are often accessed from search but do not lead to users navigating further (above-mean searchshare, above-mean resistance); (iii) *navigation-exit* articles that receive their traffic mostly from navigation but cannot channel traffic to other pages (below-mean searchshare, above-mean resistance); (iv) *navigation-relay* articles that are mainly accessed from within Wikipedia and able to pass traffic on internally (below-mean searchshare, below-mean resistance). Table 1 reports the share of articles and views pertaining to each group. We observe that a small group of highly visited articles is able to inject considerable amounts of traffic (search-relay) into Wikipedia while about a fifth of the articles' role is mainly to channel traffic internally (nav.-relay). On the other hand, exit points receive less views while covering a much bigger portion of Wikipedia articles. Overall, these observations are in line with Figure 3.

Traffic entropy. In order to assess how concentrated the incoming and outgoing traffic of an article is over all links leading to and from it, respectively, we employ entropy as a measure of (in)equality. High entropy reflects a traffic pattern for which all links are almost equally likely to deliver or transmit traffic to and from an article, whereas low entropy values indicate that almost all of the traffic is flowing in and out over a small number of links. Figure 4 shows how searchshare and resistance are related to in- and out-entropy; entropy is calculated by ignoring all article links that are not transitioned at all and is discretized into 25 bins. For each bin, the 1st, 2nd and 3th quartile for searchshare and resistance are depicted. For articles with high in- and out-entropy, we observe increased searchshare (*cf.* Figure 4(a) and (b)). This is an indication that articles that attract external search traffic are also capable to attract traffic from various internal navigation sources. Importantly, they disperse traffic over their outgoing links into the network. We also observe that with increasing in-entropy resistance decreases (*cf.* Figure 4(c)). This suggest that articles that receive traffic from different sources are also good relay points. Exit points are articles with rather average out-entropy, whereas extreme out-entropy values – both high and low – are associated with low resistance. Thus, there are two types of relay articles: articles that spread the traffic through many links and articles that channel the traffic through very few links (*cf.* Figure 4(d)).

Summary. Our analysis shows that search dominates navigation with respect to the number of articles accessed and visit frequency. However, the less viewed an article is, the more significant navigation becomes as an information access form. Further, only popular articles are able to relay traffic while the majority of the articles acts as exit points for user search and navigation sessions. Low resistance articles (relay articles) can do both spread the traffic through many links and channel it through only a few links. Search-relay articles are able to attract also traffic from internal navigation. Moreover, they are spreading the traffic in many directions.

Table 1: Article group sizes and views. For each group, the table shows the percentage of articles and their received views. The majority of the articles are less visited and act as exit points of user session, whereas only popular articles are able to further relay traffic.

	search-exit	search-relay	nav.-relay	nav.-exit	total
articles	43%	9%	21%	27%	100%
views	17%	37%	39%	7%	100%

[3] A delimitation by median yields groups with the sole resistance value 1.0 and was therefore not used. Cut-offs at 0.5 would have created extremely unbalanced groups.

4 Characterizing Access Behavior

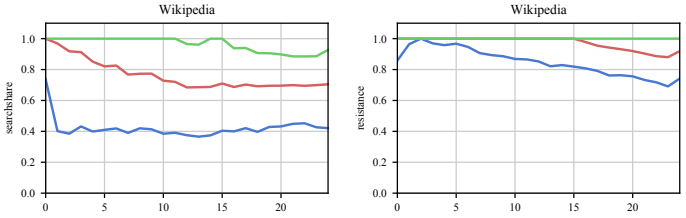
In the previous section, we analyzed the general Wikipedia information access behavior, setting aside individual page attributes. However, Wikipedia articles have different properties that may influence the way they are retrieved (*cf.* Section 4.1). To this end, we analyze the general Wikipedia access behavior dependent on the article network (*cf.* Section 4.2), and content and edit properties (*cf.* Section 4.3). Subsequently, we highlight differences between general access behavior on Wikipedia and on Wikipedia topics dominated by search and navigation, respectively (*cf.* Section 4.4). The provided characterization of user access behavior is complemented by a bow tie analysis of Wikipedia's traffic from search and navigation (*cf.* Section 4.5).

4.1 Wikipedia Article Data and Features

To study the influence of the content on the preferred access behavior, we focus on a snapshot of all Wikipedia articles contained in the main namespace of the English language version from August, 2016[4]. We obtained the articles using the Wikipedia API[5]. The collected article data represent the HTML version of each article on which the transitions data used to study the Wikipedia traffic has been generated (*cf.* Section 2.1). By parsing and rendering the HTML version of the articles, we are able to extract article features capturing aspects related to the content of the articles. The dataset contains roughly 5 million articles connected by 391 million links.

For these Wikipedia articles, we determine a wide variety of features describing their characteristics. We categorize these features into three different groups, *i.e.*, (i) network properties, (ii) content and edit properties and (iii) article topics. The network features consist of *in-, out-* and *total degree* of the article in the Wikipedia link network as well as the k-core value for this network as a typical centrality measure. Regarding the content and edit properties, we calculated for each article the *number of sections, tables, figures and lists* contained in the article. These features capture visual appearance of the article, whereas the *number of revisions and editors* represent the content production process. We also consider the article *age* measured in years to account for differences between mature and young articles. To account for

[4] https://archive.org/details/enwiki-20160801
[5] https://www.mediawiki.org/wiki/API:Main_page



(a) k-core vs. searchshare (b) k-core vs. resistance

Figure 5: Network position. The figure shows the first (blue), second (red), and third (green) quartile of searchshare (a) and resistance (b) as function of the article position in the network indicated by its k-core. K-core values are divided into 25 bins. The access behavior on articles is influenced by their position in the network. The more central an article, the lower its searchshare and resistance – i.e., the more traffic it relays through the network.

the amount of information provided in an article, we calculate its *size in kilobytes*. The features capturing the content production process are extracted from the TokTrack dataset (Flöck *et al.*, 2017) and consider the period between article creation and the end of August 2016. As the Wikipedia article categories are often too specific[6], we fit a Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) model on article texts using Gensim (Řehůřek and Sojka, 2010) bag of words article vectors with removed stop words. To allow for manual interpretation of the topics, we fit a model for 20 topics. Subsequently, we asked five independent researchers to provide topic labels based on the top words and Wikipedia articles for each topic and summarized their labels. Section 4.4 describes the extracted topics. The following analyses are based on a random sample of 50000 articles.

Table 2: Network features. For each network feature, the table shows the *median* feature values of the articles in the respective group. The article network properties influence the preferred access behavior. Nav.-relay articles act as intersections for the traffic as they occupy central network positions and provide lots of in- and outgoing links. Search-relay articles are similarly well-connected, which is important for injecting traffic into Wikipedia. Exit points (search-exit and nav.-exit articles) lack connectivity and are unable to channel external and internal traffic, respectively.

M	search-exit	search-relay	nav.-relay	nav.-exit	overall
in-deg.	14	38	54	18	22
out-deg.	33	56	71	35	41
degree	51	105	131	57	69
k-core	44	76	95	49	57

[6]I.e., very specific categories of articles are not linked to the relevant super-category; in other cases, two conflicting categories are linked or fitting categories are missing completely.

4.2 Network Features

To understand the role of the network features, we compute the median of the features for each of the four article groups *search-exit*, *search-relay*, *nav.-relay*, and *nav.-exit* (*cf.* Section 3). The results are shown in Table 2. We can observe that articles with below-average searchshare and resistance (article group *nav.-relay*) have higher median values across all network features, *i.e.*, they are located more in the center of the network and consistently have more incoming and outgoing links. Although search-relay articles are not as well connected as nav.-relay, their relatively central position in the network and high number of outgoing connections is important in order to inject traffic into Wikipedia. By contrast, articles that are often used as exit points (*search-exit* and *nav.-exit*) are located more in the periphery of the network (low k-core value), are less often linked to, and contain less outlinks themselves, which eventually results in higher resistance values, signifying the termination of user sessions.

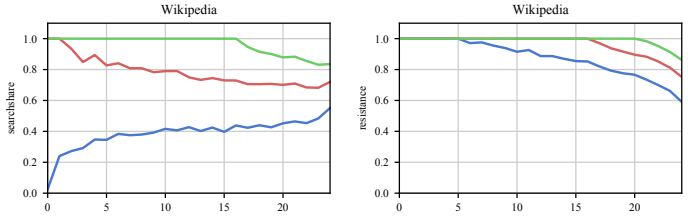
For further analysis, we sort the articles according to their k-core value and discretize them into 25 equally-sized bins. For each bin, we compute the quartiles for searchshare and resistance, as seen in Figure 5. Looking at the median (center red line), we find that for articles with increasing k-core values the searchshare indeed decreases (*cf.* Figure 5(a)). However, this effect stops at around 50% of the dataset, *i.e.*, for half of the articles, which are located in high k-core network layers, the searchshare is mostly independent from the exact centrality. Regarding the resistance, there exists a substantial amount of nodes with a resistance of 1.0 for all k-core values, *cf.* the green line indicating the upper quartile. However, for the more central nodes, an increased number of pages have a significantly lower resistance (*cf.* Figure 5(b)).

4.3 Content and Edit Features

Next, we characterize the article groups in terms of the article content and edit history which account for the content presentation and content production process. Table 3 reports the median values of these features in the four article groups. We can observe that the content features (number of tables, number of sections, size of the article) are modestly increased for relay articles, *i.e.*, articles that contain more content tend to be less often exit points

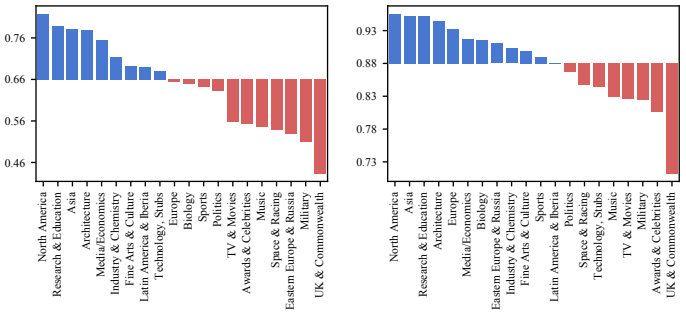
Table 3: Content and edit features. For selected content and edit features, the table shows the median feature values of the articles in the respective group. The content production process influences the access behavior as search- and nav.-exit points have low edit activity, and offer less content. On the other hand, relay articles are more frequently edited, and congruently, are generally more extensive.

M	search-exit	search-relay	nav.-relay	nav.-exit	overall
editors	21	52	46	21	25
rev.	38	97	86	37	46
sections	6	7	7	4	6
tables	3	3	4	3	4
age	9	11	10	8	9
size	41	50	54	41	44



(a) revisions vs. searchshare (b) revisions vs. resistance

Figure 6: Edit activity. The figure shows the first (blue), second (red), and third (green) quartile of searchshare (a) and resistance (b) as function of the article editors' activity indicated by the number of revisions. Revisions values are divided into 25 bins. Except for the most edited articles, high edit activity has a negative effect on the resistance, which on the other hand has a positive effect on navigation indicated by the lower searchshare.



(a) **searchshare** (b) **resistance**

Figure 7: Access behavior for all topics. Topics are ordered from highest (left) to lowest (right) for searchshare (a) and resistance (b). Values over (blue) and below (red) the respective mean value are colored respectively. There are pronounced differences in the dominant access behavior on different Wikipedia topics.

of navigation sessions. By contrast, the revision history plays a more important role: we can see that articles in the *search-relay* group have (as a median) more than twice the number of editors and revisions compared to exit articles, and tend also to be somewhat older. Articles in the group *nav.-relay* show similar, but slightly lower values with the same tendency. The median feature values for both "exit" article groups are very similar and show slightly lower editor and revision numbers. Overall, content and edit features provide strong indicators for articles relaying traffic (as opposed to being exit points), but only weak indicators for being accessed by search or by navigation.

We will have a more detailed look at an exemplary edit feature, *i.e.*, the number of revisions. Analogously to above (network features), we assign the articles to one of 25 bins according to their revision count, compute the distribution of searchshare and resistance for each bin, and plot the quartiles. The results are shown in Figure 6. We can see that the median searchshare continuously decreases with increasing number of revisions. The effect is in particular significant for very low number of revisions (*cf.* Figure 6(a)). Additionally, the spread of the distribution – measured by the interquartile range (IQR) – also substantially decreases the more revisions an article has. This can likely be explained by *regression to the mean* since articles with less revisions receive overall less views, making more extreme searchshare values more likely. With regard to the resistance, we can observe that specifically high number of revisions correlate with a lower resistance scores (*cf.* Figure 6(b)). The number of editors, and the age of an article is highly correlated with the number of revisions and reveal a very similar behavior with respect to searchshare and resistance.

4.4 Topic Features

Search-related popularity, navigability as well as other characteristics related to traffic might be highly dependent on the topical domain of an article. We hence investigate the access behavior across Wikipedia's numerous article themes, represented by the 20 topics we have extracted. Table 4 provides descriptive statistics of these topics. With 32% "TV and Movies" is the topic with the most views while consisting of a mere 7.5% of all articles on Wikipedia. "Technology, Stubs" and "Architecture" show an opposing dynamic, providing a large amount of articles, but relatively few views.[7] Overall, the amounts of articles and view counts are not

Table 4: Topic statistics. The table shows the percentage of articles and views for each topic. Additionally, it reports the median age in years, number of editors and revisions, and the size of the articles in kB. Not surprisingly, popular articles are generally longer in terms of text, edited more and by higher number of editors, and relatively old.

topic	% articles	% views	*M* age	*M* editors	*M* rev.	*M* size
Technology, Stubs	19.3	7	7	20	36	38
Architecture	12.4	5	8	31	61	56
Sports	12.0	8	7	37	86	68
Politics	8.1	8	8	47	103	60
TV&Movies	7.5	32	8	96	197	55
Fine Arts&Culture	7.2	6	8	49	100	49
Biology	7.0	6	7	29	57	46
Music	6.9	9	8	64	136	53
Research&Education	4.8	2	7	40	87	47
Media/Economics	3.3	4	8	54	115	49
Military	3.1	4	8	47	105	65
Industry&Chemistry	3.0	6	9	66	126	55
North America	1.2	0	9	23	39	52
Space&Racing	1.2	2	8	52	114	73
Europe	0.9	0.0	7	19	36	52
Asia	0.7	0.0	5	8	14	60
Lat. America&Iberia	0.5	0.0	7	20	33	58
UK&Comm.	0.5	0.0	7	19	40	43
East. Europe&Russia	0.4	0.0	7	15	24	49
Awards&Celebrities	0.0	0.0	6	22	42	41

[7] "Technology, Stubs" is a compound of general Wikipedia:Stub articles and often short technology articles that were not sufficiently distinguishable by LDA. We exclude it from discussion here due to its ambivalent nature.

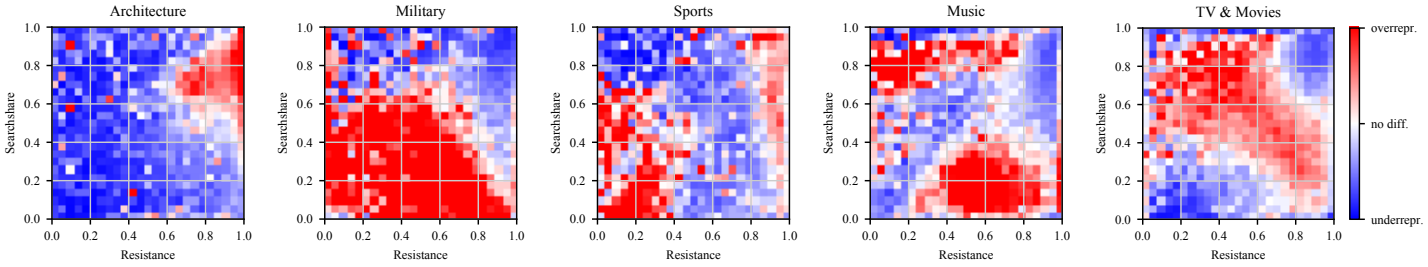
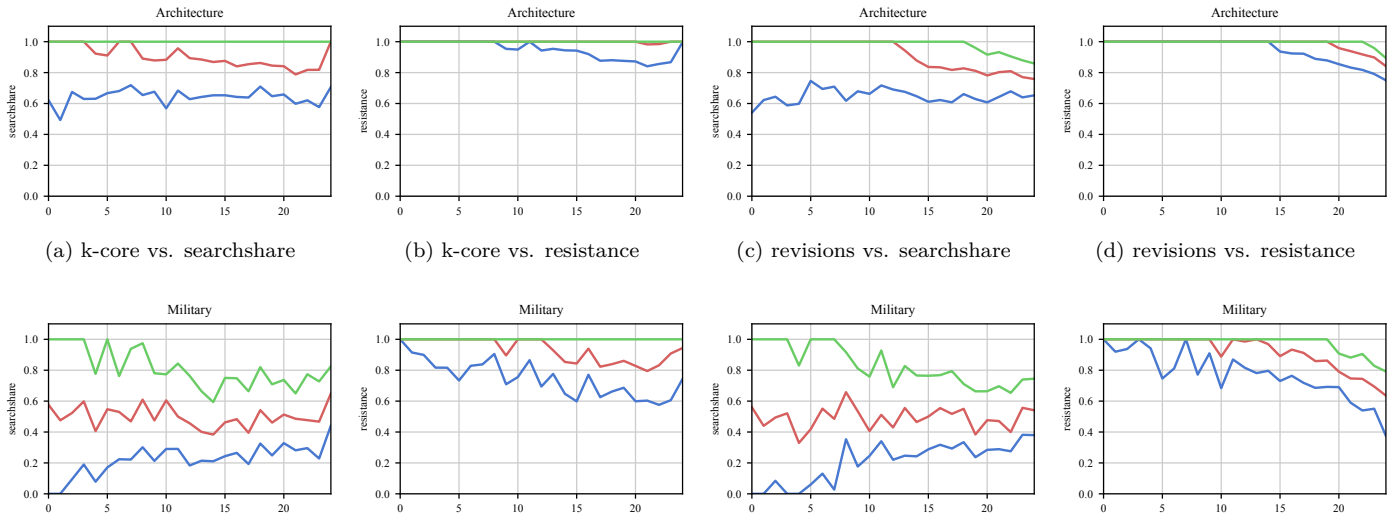


Figure 8: Relative difference of individual topics to the overall view distribution of searchshare vs. resistance (*cf.* Figure 1(b)). White denotes no relative difference, blue denotes underrepresentation (down to 0), while red denotes overrepresentation (max. over all topics at 2). The figure highlights the differences between search-heavy and navigation-heavy topics compared to the all-articles baseline. "Architecture", exhibiting above-mean searchshare and resistance (*cf.* Figure 7) stands representative for six similarly distributed topics and mainly attracts search hits that it cannot pass on. "Military" shows an almost inverted pattern, mostly receiving as well as producing internal navigation. The bi-focal distribution of "Sports" can be found in "Politics" and "Fine Arts & Culture" as well, while patterns for "Music" and "TV & Movies" are more unique.



(a) k-core vs. searchshare (b) k-core vs. resistance (c) revisions vs. searchshare (d) revisions vs. resistance

(e) k-core vs. searchshare (f) k-core vs. resistance (g) revisions vs. searchshare (h) revisions vs. resistance

Figure 9: Relation of traffic features with network and content features. For a topic dominated by search ("Architecture") and one dominated by navigation ("Military"), the figure shows the first (blue), second (red), and third (green) quartile of the article searchshare and resistance as function of its position in the network indicated by its k-core and editors' activity indicated by the number of revisions. Articles are divided into 25 bins by k-core and revision values. Apart from base-level differences of searchshare and resistance, the topics exhibit comparable trends, with the exception of searchshare not being influenced as much by network position or edit activity features for "Military" articles.

strongly correlated. Consistent with previous research, we also observe that the popular articles are in general longer, relatively old, and revised more often by more editors (Spoerri, 2007).

A look at the distribution of searchshare and resistance in the overview provided by Figure 7 reveals the different access behaviors for Wikipedia topics. To examine these pronounced differences further, we set out to highlight the dissimilarities of the overall searchshare vs. resistance distribution for total views – as shown in Figure 1(b) – with the same distribution for the individual topics. To do so, we create heatmaps pinpointing the *relative differences* of each topic to the baseline of the overall distribution. This is achieved by performing a bin-wise division of a topic's normalized view count for a given searchshare-resistance bin with the respective normalized bin for the general Wikipedia traffic behavior. The resulting heatmaps are shown in Figure 8 for selected topics. They draw a clear picture of the over- and under-representation of certain article types (in terms of views) in each topic against the whole-system baseline. "Architecture"

in Figure 8 stands as one representative for a group of topics ("Biology", "Industry & Chemistry", Research & Education, "Space & Racing") that all exhibit a very similar distribution with their article views occurring at high searchshare and high resistance, *i.e.*, these topics are mostly searched and not used for further navigation. In stark contrast, views for "Military" topics occur to the largest part in comparably low-resistance articles, that are mostly navigated to (views for "UK & Commonwealth" are distributed almost analogously). "Sports" reveals a similar inclination for *nav-relay* types of articles attracting views, yet sports articles also frequently get accessed by search and abandoned immediately (closely related patterns: "Fine Arts & Culture" and "Politics"). Lastly, "Music" and "TV & Movies" exhibit remarkably idiosyncratic distribution patterns, not mirrored by another topic. "Music" attracts many views in a *search-relay* fashion, but on the other hand also explicitly acts as a "dead end" for internal navigation.

As "maximally different" topics in respect to these traffic patterns and with overall high view counts, we select "Architecture" for search-heavy topics, and "Military" for navigation-heavy topics to conduct a deeper analysis regarding article network, content and edit properties. While "Architecture" includes articles covering popular buildings, landmarks and municipalities, "Military" consists of articles covering significant historic events often associated with violence such as wars and notable battles, along with many articles dedicated to military units, personnel and equipment (*cf.* (Samoilenko *et al.*, 2017)). For the general access behavior concerning the network, content and edit features, we again assign the articles to one of 25 bins according to their k-core and revision counts, compute the distribution of searchshare and resistance for each bin, and plot the quartiles (*cf.* Figure 9). For "Architecture", searchshare (a) initially decreases for increasing k-core but sees an uptick for very central nodes, and a very similar behavior can be observed for resistance (b). "Military" is characterized by generally lower levels of both metrics, yet shares the trend of decreasing resistance with increasing k-core (e), meaning that for both topics, the more central articles in the network are able to channel visitors into Wikipedia, with the topmost central nodes excluded from this trend. Being edited more implies decreasing resistance for both topics ((d), (h)), although this trend reveals itself only for much higher revision counts for "Architecture", most likely to its generally higher resistance. Edit counts have no clearly distinguishable influence on "Military" articles' searchshare, for "Architecture" it, however, implies lower searchshare.

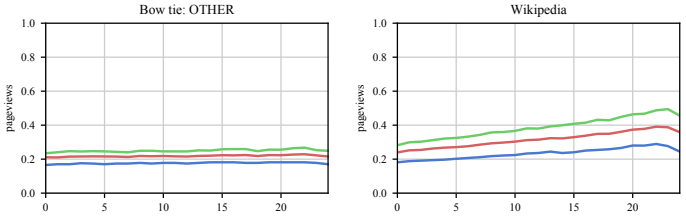
4.5 Bow tie analysis

To gain a more comprehensive view on Wikipedia's traffic from a network theoretic perspective, we perform a bow tie membership and pageviews analysis of the transitions network D_{trans} (edges are navigational transitions connecting Wikipedia articles) and of the network D_{total} (extends D_{trans} to contain also searched articles). The bow tie model has been introduced by Broder *et al.* to describe the graph structure of the Web (Broder *et al.*, 2000) and has already been applied to assess the navigability of Wikipedia's network (Lamprecht *et al.*, 2016). According to this model, a directed graph can be decomposed into seven different components: SCC—the set of mutually reachable nodes, IN—the set of nodes having paths to SCC but not part of it, OUT—the set of nodes not in SCC but reachable from it, TUBES—the set of nodes that are on a path from a node in IN to a node in OUT, (IN and OUT)TENDRILS—leading away from IN and towards OUT, and OTHER—accommodates all disconnected components in the graph. The results of the bow tie membership and views analysis are presented in Table 5. D_{trans} has a rather small number of articles in the IN component (3.7%) that generates only 0.6% of the outgoing views and a big OUT component (42.2%) that receives only 6.2% of the views. Moreover, we see that SCC accounts for 51.8% of all nodes, and for 93.8% of the incoming and 99.2% of the outgoing views from navigation. This suggests that Wikipedia's traffic forms a network with about half of all accessed articles being members of SCC. These articles act as navigation-relay points as they attract and generate most of the views from navigation. On the other hand, transitions leading outside the SCC to a large number of navigation-exit articles in OUT account only for a small portion of all views. As extending D_{trans} to D_{total} adds only nodes and no edges to D_{trans}, it is interesting to observe how the node membership and views percentages in each component changes. Network-theoretically, articles in IN can act as relay points and contribute to increasing the number of views of the articles in SCC by converting search to navigation traffic. However, as for D_{trans}, this component remains rather small and receives few views. While the proportions of SCC and OUT do not change notably with respect to views, these components change in terms of their sizes to reflect the emergence of a big OTHER component (28.2%) accounting for 1.7% of all views. The nodes in OTHER cannot act as navigation-relay articles, as they are disconnected from the internal navigational traffic (no incoming transitions). Moreover, they are unable to convert search traffic to internal navigation (searchshare=1.0, resistance=1.0).

A possible explanation of this pattern might be a look-up user behavior indicating a search-exit functional role for those articles. Nevertheless, this traffic access pattern makes these articles especially interesting with respect to their Wikipedia

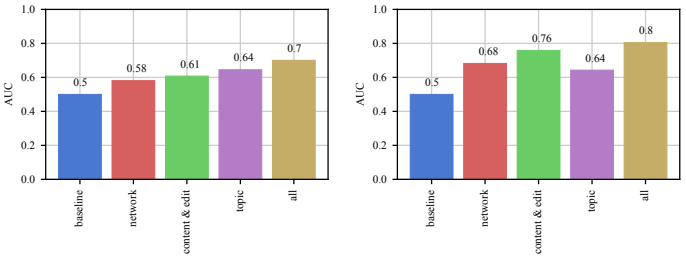
Table 5: Bow tie analysis. SCC and OUT are the biggest components for both networks. Articles in SCC act as relay points as they attract both internal navigation and external search traffic. Only a small portion of search and navigation traffic is distributed among a large number of exit point articles in OUT. Although articles in IN can act as search-relay points and ingest external search traffic into SCC, they are too few and not attractive enough to users. The TUBE component is not shown since it is empty.

component	D_{trans} % articles	D_{trans} % views in_{nav}	D_{trans} % views out_{nav}	D_{total} % articles	D_{total} % views
IN	3.7	0.0	0.6	2.7	0.8
SCC	51.8	93.8	99.2	37.7	92.6
OUT	42.2	6.2	0.2	30.7	4.8
TL_IN	0.4	0.0	0.0	0.3	0.0
TL_OUT	0.6	0.0	0.0	0.4	0.1
OTHER	1.3	0.0	0.0	28.2	1.7



(a) k-core vs. pageviews (b) k-core vs. pageviews

Figure 10: Bow tie: OTHER. The figure shows the first (blue), second (red), and third (green) quartile of pageviews as function of the article's k-core. Pageviews are logarithmically transformed and then normalized. Network properties values are divided into 25 bins. Independent of their k-core (a), articles in OTHER receive a stable amount of external traffic through search while the total traffic (from search and navigation) increases the more central an article is in the network (b). Similar results are obtained for in- and out-degree.



(a) searchshare (b) resistance

Figure 11: Results. The figure shows the model performance (ROC AUC) for (a) searchshare and (b) resistance. Predicting searchshare is more challenging than predicting resistance. The article topic determines the preferred access behavior (search or navigation). However, position in the network, content maturity and presentation of the article are indicative of resistance, and thus if an article will be an entry-exit point or a relay point for the traffic.

network properties that might reveal other reasons leading to a breakdown of navigation. Interestingly, independent of their positions and connectivity in the Wikipedia network, the articles from the OTHER component receive the same amount of views (*cf.* Figure 10 (a)). This is exceptional, as in the normal case the total views (views from search and from navigation) increase the more central and connected an article is (*cf.* Figure 10 (b)). After manual examination, we found that often articles in OTHER are: (i) presenting the content in a way not necessarily compliant with the Wikipedia guidelines, (ii) in general small pages, *i.e.*, stubs, or (iii) disambiguation pages. For example, 1st, 2nd, and 3rd Massachusetts Regiment are articles from OTHER for which navigation breaks down. As these articles belong to the topic "Military", their expected access patterns should be similar to articles accessed by navigation. With about 40 in-coming, 90 out-going links and a k-core about 100, these articles are also well connected. However, these articles are not interconnected via links in the running text but only via links in their navigational boxes at the bottom of each page which users are less likely to be found and clicked by users (Dimitrov *et al.*, 2015; Dimitrov *et al.*, 2016).

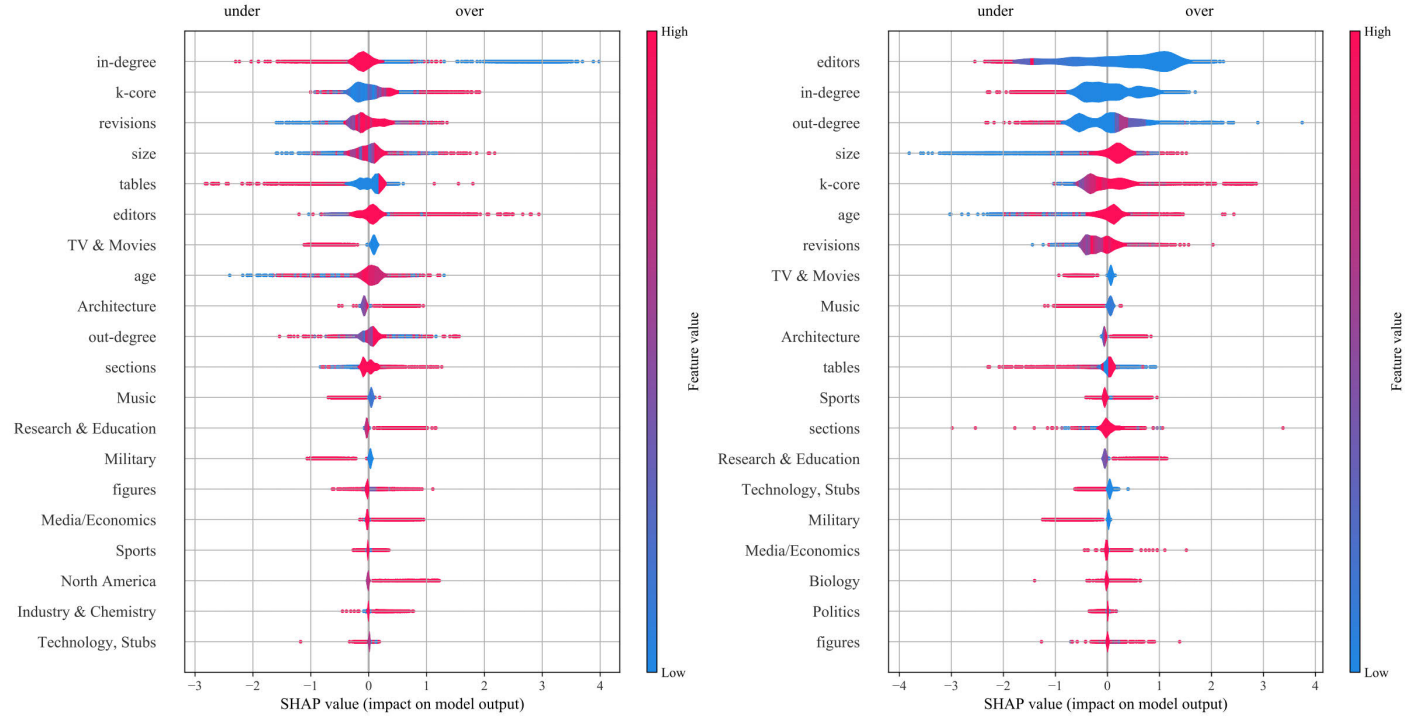
Summary. The results presented in this section suggest that the content heavily influences the access behavior on Wikipedia. Particularly, topical domains are accessed differently, *i.e.*, users prefer to access articles about architecture and landmarks mainly through search, whereas more historical articles about military actions are navigated. Moreover, mature articles with high revision numbers and articles located in the core of the network are more likely to channel traffic through Wikipedia, whereas articles located at the network periphery act as exit points. A bow tie analysis suggests that transitions form a network with a SCC attracting almost all of users' attention.

5 Modeling Access Behavior

Our previous analysis characterized the user access behavior on Wikipedia articles with respect to their traffic from search and navigation dependent on the article features. However, this analysis does not reveal the impact of the feature groups on the access behavior. To this end, we set out to model the access behavior on articles in order to measure the relative advantage of each feature group with respect to the models' predictive performance. The higher the predictive performance of a feature group, the higher the relative advantage of the group is on the role articles play with respect to the traffic (entry-exit and relay articles), and thus on the preferred information access form (search and navigation). This approach of estimating the influence of a feature on the predictive performance is widely used and adopted also in many advance methods for measuring feature importance (Fisher *et al.*, 2018). We tried out different machine learning approaches and report the results for the best performing model approach *i.e.* tree-based gradient boosting; to fit the model we resort to the scikit-learn implementation[8]. As a final step using SHAP values, we inspect the importance of individual features on the model output, *i.e.*, predicted class.

Modeling searchshare. We ask, given a Wikipedia article, if it is possible to classify it as dominated by search, *i.e.*, $searchshare > 0.66$ or dominated by navigation, *i.e.*, $searchshare \leq 0.66$. The threshold used for the separation is the searchshare mean (*cf.* Section 3). In our experiments, we consider three different sets of article features: (i) network features, *i.e.*, in-, out-degree, k-core, (ii) content & edit features, *i.e.*, article size and age, number of revisions and editors, number of sections, tables, figures and lists, and (iii) topic. For predicting the preferred access form, we fit a separate tree-based model using gradient boosting for each feature group and evaluate the models' performance with ROC AUC. The models are trained using 10-fold cross validation on a balanced dataset. On this dataset random guessing results in 50% accuracy, which is used as a baseline. Figure 11(a) shows the individual performance for each feature group, as well as the performance for the combination of all features. We observe that

[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html



(a) searchshare (b) resistance

Figure 12: SHAP feature values. The figure shows the SHAP values for the twenty most important features for (a) searchshare and (b) resistance. Features are shown in decreasing importance order. Each dot shows the SHAP value of a training sample. Dots stack up to visualize density of samples. Dots are colored according to their feature values. SHAP values patterns suggest that both models make plausible predictions corroborated by previous results (*cf.* Section 4).

modeling searchshare is difficult even with all features (AUC = 0.70). The network features are the least indicative (AUC = 0.58). The topic feature predicts searchshare best (AUC = 0.64), which suggests strong user preferences for specific information access forms, *i.e.*, search or navigation for different topics.

Modeling resistance. To model the resistance of Wikipedia articles, *i.e.*, their ability to relay traffic, we treat an article as a relay point if $resistance \leq 0.88$ and an exit point if $resistance > 0.88$. Again, the separation of the articles is based on the resistance mean (*cf.* Section 3). We consider the same feature groups as for modeling searchshare and use random guessing as our baseline. For classifying the articles, we again utilize 10-fold cross validation to train separate tree-based gradient boosting models for each feature group on a balanced dataset. The performance is measured in terms of ROC AUC. Figure 11(b) shows the individual classification performance for that task for each feature group. The content and edit features are the most important (AUC = 0.76). This makes the case for an influence of the way content is presented to the user on lowering or increasing the resistance of a page. Unlike for searchshare, the network features are indicative of the resistance of an article (AUC = 0.68). This suggests that the network position of an article influences the extent to which it channels traffic. The topic plays only a small role, which again highlights the importance of the quality of the content presentation and production process.

Model understanding. To aid a more fine-grained understanding of how the individual features contribute to the overall model performance and to predicting one class over the other, we provide SHAP summary plots. SHAP summary plots are based on SHAP (SHapley Additive exPlanation) values, which combine local explanations (Ribeiro *et al.*, 2016) and game theory (Štrumbelj and Kononenko, 2014) to ensure consistent feature importance attribution. Compared to traditional feature importance attribution methods such as gain, split count and permutation, SHAP values are consistent with model changes. This means that if a model changes to rely more on a given feature, the SHAP value of a feature also increases (Lundberg *et al.*, 2018). To compute SHAP values, we utilize SHAP Tree,[9] an efficient algorithm for tree-based models such as Gradient Boosting and Random Forest.

Figure 12 shows SHAP summary plots for (a) searchshare and (b) resistance for the models using all features. The features are sorted in decreasing order of their global model importance. Each dot shows the SHAP value of a training sample (Wikipedia article) for a given feature. Dots stack up to visualize density of samples with a given SHAP value. Each dot is colored to indicate the feature's value. High SHAP values of a feature indicate predicting over the mean searchshare and resistance, while low SHAP values indicate predicting under the mean searchshare and resistance. The in-degree of an article has the biggest impact on

[9] https://github.com/slundberg/shap

the searchshare model's output (*cf.* Figure 12(a)), although we saw that network features as a group do not aid in increasing predictive performance necessarily. While high in-degree is indicative of predicting under-average searchshare, the trail of low in-degree articles with high SHAP values suggests that extreme low in-degree values are responsible for predicting over the average searchshare, which is congruent with an intuitive explanation of fewer opportunities for in-navigation. Regarding the number of tables, we see that samples stack around zero SHAP value for high and low feature values. However, there is a small number of articles with a high number of tables that exhibit extremely negative SHAP values. This indicates that a high number of tables in an article can also lower the predicted searchshare value, whereas low table numbers cannot increase it. With respect to article topics, we observe that articles of the topic "TV & Movies" tend to bias the model towards predicting lower than average searchshare. The pattern is very similar to the SHAP values pattern for "Military". These observations support our previous findings indicating that articles from these topics are rather relay points as they are able to forward external traffic to other Wikipedia articles and also attract internal traffic. Almost the opposite pattern can be observed for the topics "Architecture", "Research & Education" and "Media/Economics" as articles from these topics are rather predicted to have over the average searchshare. Regarding resistance, the most important feature is the number of editors followed by the article's in- and out-degree (*cf.* Figure 12(b)). This suggests that the output class of the model is influenced by the content and edit behavior features followed by the way the article is embedded in Wikipedia's information network. More specifically, a high number of editors and many incoming and outgoing navigation possibilities lower an article's predicted resistance. The influence of the number of tables is even more evident for resistance than for searchshare, *i.e.*, extreme values of this feature can not only increase but also lower the odds for predicting under or over average resistance, respectively. Articles belonging to the topics "TV & Movies" "Military" and "Music" impact the model to predict under the average resistance and thus are indicative of relay points. On the other hand, the articles from topics "Architecture" and "Research & Education" as for searchshare exhibit the opposite SHAP value patterns and impact the model to predict over the average resistance.

Summary. In general, modeling article resistance is easier than modeling searchshare as suggested by the higher ROC AUC values. Modeling searchshare is challenging due to the influence of external events (*e.g.*, the transition data exhibits high view numbers on articles about the Summer Olympics 2016), and the content diversity.

Regarding searchshare, the topic of an article is the most clear-cut explanatory factor of why it is accessed more or less through external search, reflected in the predictive ability of that feature group as well as in the clear distinctions between over and under mean prediction for high and low values of the different topics in the SHAP analysis (with high in-degree adding to more incoming navigation traffic and effectively lowering searchshare). For an article's ability to relay traffic the picture is quite different, with content presentation and community engagement as well as the article's position in the network all being responsible for changes in resistance. Notably, articles that are interacted with by many editors appear to be more likely to be of low resistance, and being more central in the network actually increases the resistance change somewhat, as does higher age and larger size. Low out-degree, on the other hand, is not necessarily associated with a higher probability of being classified as above mean resistance, which might be a first intuitive assumption.

6 Related Work

Since the inception of the Web, researchers have been studying the user content consumption behavior. Initially, content has been accessed by traversing hyperlinks on the Web (Kumar and Tomkins, 2010). This navigational user behavior on the Web and on Wikipedia is often modeled using well-established methods such as Markov chains (Chierichetti *et al.*, 2012; Singer *et al.*, 2015; Singer *et al.*, 2014; Page *et al.*, 1999; Pirolli and Pitkow, 1999) and decentralized search models (Dimitrov *et al.*, 2015; Helic *et al.*, 2013). Numerous navigational hypotheses on Wikipedia have also been presented based on, *e.g.*, click traces stemming from navigational games and on click data from server logs. For example, West and Leskovec observed a trade-off between similarity and popularity to the target article in the user sessions of the Wikispeedia game (West and Leskovec, 2012). Lamprecht *et al.* studied the general navigability of several Wikipedia language editions and showed how the Wikipedia article structure influences the user click behavior (Lamprecht *et al.*, 2016; Lamprecht *et al.*, 2017). Dimitrov *et al.* conducted a large-scale study on the navigational behavior on Wikipedia. They found that users tend to select links located in the beginning of Wikipedia articles and links leading to articles located in the network periphery (Dimitrov *et al.*, 2017; Dimitrov *et al.*, 2016). By constructing a navigational phase space from transition data, Gilderslave and Yasseri studied internal navigation on Wikipedia and identified articles with extreme, atypical, and mimetic behavior (Gildersleve and Yasseri, 2018). The navigational user behavior is also influenced by the structure of the underlying network. In literature there are two models characterizing the topology of a network which are of special interest with respect to navigation. "Small world" networks are networks in which any given node pair is connected via a short chain of nodes (Watts and Strogatz, 1998). This property makes such networks highly navigable and researchers have found that it is the result of a fine balance between the degree and clustering coefficient distribution of the network (Boguna *et al.*, 2009; Kleinberg, 2000). In this paper, we study the traffic flow on Wikipedia by looking at its network not from the "small world" network perspective but from a core-periphery perspective while accounting for search traffic. The second model is the bow tie model that is trying to characterize the structure of the Web and how this structure shapes the traffic flow. In contrast to previous works our study applied the bow tie model not to the underlying network, *i.e.*, Wikipedia link network but to network of transitions between Wikipedia articles. Web content can be also discovered by formulating and executing a search query. Kumar and Tomkins performed an initial characterization of the user search behavior (Kumar and Tomkins, 2009), while Weber and Jaimes studied the search engine usage with respect to the users demographics, topics, and session length (Weber and Jaimes, 2011). Earlier Wikipedia reading behavior studies focused on explaining bursts,

dynamics of topic popularity and search query analysis to Wikipedia (Thij *et al.*, 2012; Ratkiewicz *et al.*, 2010; Spoerri, 2007; Waller, 2011; Lehmann *et al.*, 2014). Singer *et al.* investigated the English Wikipedia readers motivations (Singer *et al.*, 2017). By complementing a reader survey with server log data, they discovered specific behavior patterns for different motivations, *i.e.*, bored readers tend to produce long article sequences spanning different topics. A more recent study by Lemmerich *et al.* extend this line of research by looking into 13 more different Wikipedia language editions (Lemmerich *et al.*, 2019). Their results showed similarities but also substantial differences in the access patterns across different language editions, *i.e.*, readers from countries with a low Human Development Index (HDI) are prevalent to a more in-depth reading of Wikipedia articles compared to readers with a high HDI. McMahon *et al.* focused on the interdependence between search engines, *i.e.*, Google and Wikipedia (McMahon *et al.*, 2017). They showed that Google is responsible for generating high traffic to Wikipedia articles, although, in some cases traffic is reduced due to the direct inclusion of Wikipedia content in search results. Compared to our work, McMahon *et al.* concentrate on the peer production site and not on the content consumption. While there is a long line of research with respect to search and – more so – navigation, they have rarely been studied together which is the focus of this work.

7 Discussion

As a general observation, our results shed light on the different roles of articles with respect to traffic entering and leaving Wikipedia. On one hand, an overwhelming amount of pages attracts mostly direct search traffic and only little internal navigation, thanks to Wikipedia's strong symbiotic relationship with web search engines. Yet, notably, most of that traffic goes to articles that act mainly as exit points, *i.e.*, users do not continue visiting Wikipedia directly afterwards. This is congruent with, but not necessarily because of, a pure "look up" nature of search. Only a very small share of searched articles is responsible for relaying disproportionally large amounts of traffic into the rest of Wikipedia. We see that these articles are well-connected, more edited and more extensive than their exit counterparts, although we cannot yet conclude whether this is because of a "worn path" paradigm, wherein links and content are built because of the natural thematic positioning and suitability of an article to act as an entry point *and* as a bridge to more content, or because the a-priori structure of these article facilitates the observed navigational patterns. A longitudinal study, which we plan for future work, could obtain more detailed insights on this co-evolution of structural features and navigation. Furthermore, our data shows that articles which are able to forward traffic sit mostly at the very (k-)core of the link network. This is however not necessarily the case for being a receiver of navigation traffic, with searchshare values stabilizing already at lower k-cores – and with in-links not being more highly correlated with k-core than out-links. This hints to the fact that – to some extent – users enter Wikipedia by search on more central articles, and then navigate outwards from more central to less central nodes. This is consistent with previous findings studying navigation on Wikipedia (Dimitrov *et al.*, 2017).

Regarding articles with different topical alignments, we see certain evidence that the thematic domain of a user's information pursuit seems connected with the "mode" of how this information is attained. While the highly aggregate data used in this work does not allow for direct inferences as to the type of information retrieval in the continuum between a targeted and well-defined lookup versus a completely serendipitous discovery process, we can nonetheless discern distinct patterns between article topics. Although "Architecture" articles are not more devoid of in- or out-navigation opportunities than "Military" pages, they show far higher amounts of search views and exit points, while the latter are navigated at a constantly high level, regardless of their connectedness. A possible explanation of the navigation-heavy behavior on "Military" articles is that people like to follow paths through events in order to understand historical developments.

Search and navigation are the two dominant information access forms not only on Wikipedia but also on the Web. The findings of this study, however, are yet to be validated for other Wikipedia language editions, where different access behaviors have already been observed (*cf.* Lemmerich *et al.* (Lemmerich *et al.*, 2019)). Secondly, their generalization to the Web requires more in-depth research, not least because webpages might assume very different roles as a result of technical aspects. For example, on Wikipedia, the availability of interlinked articles is ensured by high-quality standards and established guidelines for the editors, whereas on the Web switching from navigation to search might be triggered by encountering a broken link. Still, our results can inform future research designs by providing investigative starting points, *e.g.*, similar traffic pattern differences in topically grouped collections of websites or similar core-to-periphery traffic flow on other platforms.

For our analysis, we utilize publicly available clickstream data about Wikipedia. However, due to privacy restrictions, the data contains only (referrer, resource) pairs that occurred at least ten times during the data collection period. This could lead to a skewed view on the access behavior when contrasting search and navigation. For example, if an article is navigated in total more than ten times over different links, but each individual link is transitioned less than ten times, all of these transitions will not be included in the data. In this case, the searchshare for this article might get substantially overestimated. The data restrictions could also affect the results of the bow tie analysis in making the OTHER component of D_{total} appear to be much bigger than it actually is due to the removal of edges infrequently transitioned over. The traffic entropy analysis might also be affected which can lead to different access patterns for articles. For example, articles that are spreading small portions of traffic over many edges can suddenly appear to channel traffic over just a few edges.

The study conducted in this paper presents fundamental research on the interplay of search and navigation on the English version of Wikipedia. Our findings can hopefully help to pave the way for improvements in Wikipedia and allow for data-driven decision making for further maintenance, resource allocation, and development of intelligent user interfaces. Specifically, we showed that different topics exhibit different traffic patterns. This finding can help developers to better understand the implications of

newly introduced platform features such as the link preview[10]. Its core idea (showing a short abstract of an article) likely leads to decreased navigation to the target articles of previewed links, *e.g.*, because the summary is deemed sufficient to prevent a link-following by the reader. This, in turn, might lead to interrupted navigation chains beyond that specific link in navigation-heavy topical domains such as "Military", effectively reducing traffic numbers for these article types.[11] Moreover, our findings can help Wikimedia to better allocate resources and improve editor guidelines. For example, the visibility of less viewed and often short articles depends to a large extent on traffic from navigation. Furthermore, they tend to act as a "dead end" for the traffic, *i.e.*, navigation-exit points. Editors can be encouraged by Wikimedia to pay special attention to the visual appearance and interlinking of such articles in order to both attract and forward more traffic on Wikipedia. A possible application of the models we presented is the identification of exit point articles even before they go online. The methodology we used to identify topics with search-heavy and navigation-heavy traffic behavior can also be applied to groups of articles or a single article over time to track changes in their traffic patterns and evaluate if changes to the article's content, visual appearance or interlinking provide the intended results or not.

8 Conclusion and Future Work

In this work, we studied the prevalence of user access preferences across articles on Wikipedia. For that purpose, we introduced *searchshare* and *resistance* as two key features to characterize article traffic. While we can identify search as the more dominant access paradigm compared to navigation on Wikipedia overall, we observe heterogeneous behavior at different types of articles. That is, depending on the article topic and other article properties, the share of navigation and search strongly varies, as well as the amount of traffic an article relays to other Wikipedia pages. For example, articles on topics such as "Military" exhibit above average access by navigation, while topics such as "Architecture" show a strong prevalence of search. Furthermore, edit activity on an article and its position in the network is strongly correlated with its ability to relay traffic on Wikipedia. Thus, we find overall that both, search and navigation play a crucial role for information seeking on Wikipedia.

In the future, we plan to extend our studies over time intervals and to other language editions in order to further explore cultural differences in the identified access patterns.

References

Berners-Lee, T., M. Fischetti, and M. L. Foreword By-Dertouzos (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. HarperInformation.

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

Boguna, M., D. Krioukov, and K. C. Claffy (2009). "Navigability of complex networks". *Nature Physics*. 5(1): 74–80.

Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener (2000). "Graph structure in the web". *Computer networks*. 33(1): 309–320.

Chierichetti, F., R. Kumar, P. Raghavan, and T. Sarlos (2012). "Are web users really markovian?" In: *Proceedings of the 21st International Conference on World Wide Web*. ACM. 609–618.

Coiro, J. and E. Dobler (2007). "Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the Internet". *Reading research quarterly*. 42(2): 214–257.

Dimitrov, D., F. Lemmerich, F. Flöck, and M. Strohmaier (2018). "Query for Architecture, Click Through Military: Comparing the Roles of Search and Navigation on Wikipedia". In: *Proceedings of the 10th ACM Conference on Web Science*. Amsterdam, Netherlands: ACM. 371–380.

Dimitrov, D., P. Singer, D. Helic, and M. Strohmaier (2015). "The Role of Structural Information for Designing Navigational User Interfaces". In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM. 59–68.

Dimitrov, D., P. Singer, F. Lemmerich, and M. Strohmaier (2016). "Visual Positions of Links and Clicks on Wikipedia". In: *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee. 27–28.

Dimitrov, D., P. Singer, F. Lemmerich, and M. Strohmaier (2017). "What Makes a Link Successful on Wikipedia?" In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 917–926.

Fisher, A., C. Rudin, and F. Dominici (2018). "All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance". *arXiv preprint arXiv:1801.01489*.

Flöck, F., K. Erdogan, and M. Acosta (2017). "TokTrack: A Complete Token Provenance and Change Tracking Dataset for the English Wikipedia". In:

Furnas, G. W. (1997). "Effective view navigation". In: *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. ACM. 367–374.

Furnas, G. W., T. K. Landauer, L. M. Gomez, and S. T. Dumais (1987). "The vocabulary problem in human-system communication". *Communications of the ACM*. 30(11): 964–971.

Gildersleve, P. and T. Yasseri (2018). "Inspiration, Captivation, and Misdirection: Emergent Properties in Networks of Online Navigation". In: *Complex Networks IX*. Ed. by S. Cornelius, K. Coronges, B. Gonçalves, R. Sinatra, and A. Vespignani. Springer International Publishing. 271–282.

Helic, D., M. Strohmaier, M. Granitzer, and R. Scherer (2013). "Models of Human Navigation in Information Networks Based on Decentralized Search". In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM. 89–98.

Kleinberg, J. M. (2000). "Navigation in a small world". *Nature*. 406(6798): 845–845.

[10] https://blog.wikimedia.org/2018/04/20/why-it-took-a-long-time-to-build-that-tiny-link-preview-on-wikipedia/

[11] Since visiting the full version of the article would have exposed the reader to more content able to stir additional interest in further navigational exploration.

Kumar, R. and A. Tomkins (2009). "A Characterization of Online Search Behaviour". *Data Engineering Bullettin*. 32(2): 2009.

Kumar, R. and A. Tomkins (2010). "A Characterization of Online Browsing Behavior". In: *Proceedings of the 19th International Conference on World Wide Web*. ACM. 561–570.

Lamprecht, D., D. Dimitrov, D. Helic, and M. Strohmaier (2016). "Evaluating and improving navigability of Wikipedia: A comparative study of eight language editions". In: *Proceedings of the 12th International Symposium on Open Collaboration*. ACM. 17:1–17:10.

Lamprecht, D., K. Lerman, D. Helic, and M. Strohmaier (2017). "How the structure of wikipedia articles influences user navigation". *New Review of Hypermedia and Multimedia*. 23(1): 29–50.

Lehmann, J., C. Müller-Birn, D. Laniado, M. Lalmas, and A. Kaltenbrunner (2014). "Reader Preferences and Behavior on Wikipedia". In: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. ACM. 88–97.

Lemmerich, F., D. Sáez-Trumper, R. West, and L. Zia (2019). "Why the World Reads Wikipedia: Beyond English Speakers". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM. 618–626.

Leu, D. J., J. Castek, D. Hartman, J. Coiro, L. Henry, J. Kulikowich, and S. Lyver (2005). "Evaluating the development of scientific knowledge and new forms of reading comprehension during online learning". *Final report presented to the North Central Regional Educational Laboratory/Learning Point Associates*. Accessed: May 2018.

Leu, D. J., H. Everett-Cacopardo, L. Zawilinski, G. McVerry, and W. I. O'Byrne (2012). "New Literacies of online reading comprehension". *The Encyclopedia of Applied Linguistics*.

Lundberg, S. M., G. G. Erion, and S.-I. Lee (2018). "Consistent Individualized Feature Attribution for Tree Ensembles". *arXiv preprint arXiv:1802.03888*.

Mangen, A. (2008). "Hypertext fiction reading: haptics and immersion". *Journal of research in reading*. 31(4): 404–419.

McMahon, C., I. Johnson, and B. Hecht (2017). "The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies". In:

Nelson, T. H. (1965). "Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate". In: *Proceedings of the 1965 20th National Conference*. ACM. 84–100.

Page, L., S. Brin, R. Motwani, and T. Winograd (1999). "The PageRank Citation Ranking: Bringing Order to the Web." *Technical Report* No. 1999-66.

Paranjape, A., R. West, L. Zia, and J. Leskovec (2016). "Improving Website Hyperlink Structure Using Server Logs". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM. 615–624.

Pirolli, P. L. and J. E. Pitkow (1999). "Distributions of Surfers' Paths through the World Wide Web: Empirical Characterizations". *World Wide Web*. 2(1-2): 29–45.

Ratkiewicz, J., S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani (2010). "Characterizing and modeling the dynamics of online popularity". *Physical review letters*. 105(15): 158701.

Řehůřek, R. and P. Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA. 45–50.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 1135–1144.

Samoilenko, A., F. Lemmerich, K. Weller, M. Zens, and M. Strohmaier (2017). "Analysing Timelines of National Histories across Wikipedia Editions: A Comparative Computational Approach". In: *Proceedings of the Eleventh International AAAI Conference on Web an Social Media (ICWSM 2017)*. Montreal, Canada. 210–219.

Singer, P., D. Helic, A. Hotho, and M. Strohmaier (2015). "HypTrails: A Bayesian Approach for Comparing Hypotheses About Human Trails on the Web". In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 1003–1013.

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

Singer, P., F. Lemmerich, R. West, L. Zia, E. Wulczyn, M. Strohmaier, and J. Leskovec (2017). "Why We Read Wikipedia". In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 1591–1600.

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

Štrumbelj, E. and I. Kononenko (2014). "Explaining prediction models and individual predictions with feature contributions". *Knowledge and information systems*. 41(3): 647–665.

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

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

Watts, D. J. and S. H. Strogatz (1998). "Collective dynamics of small-world networks". *Nature*. 393(6684): 440–442.

Webber, W., A. Moffat, and J. Zobel (2010). "A similarity measure for indefinite rankings". *ACM Transactions on Information Systems (TOIS)*. 28(4): 20.

Weber, I. and A. Jaimes (2011). "Who Uses Web Search for What: And How". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. ACM. 15–24.

West, R. and J. Leskovec (2012). "Human Wayfinding in Information Networks". In: *Proceedings of the 21st International Conference on World Wide Web*. ACM. 619–628.

Wulczyn, E. and D. Taraborelli (2016). "Wikipedia Clickstream. figshare." doi:10.6084/m9.figshare.1305770. Accessed: 2017-5-3.