How We Use Wikipedia: Studying Readers' Behavior with Navigation Traces

Tiziano Piccardi

Acceptée sur proposition du jury:

Prof Karl Aberer, président du jury
Prof Robert West, directeur de thèse
Prof Tanja Käser, rapporteur
Prof Markus Strohmaier, rapporteur
Prof Ryen W. White, rapporteur

Lausanne, EPFL, 2022

Abstract

In the information age, the Web and the growing global connectivity drastically simplified our access to information. Learning and fact-checking from online resources is nowadays part of our daily routine. Studying the dynamic associated with online content consumption is critical to understanding human behavior and informing future platforms' design.

In this thesis, we provide a comprehensive overview of online knowledge-seeking, a specific instance of information-seeking, by describing the behavioral pattern of Wikipedia readers. Despite the importance and pervasiveness of Wikipedia as one of the largest platforms for open knowledge, surprisingly little is known about how people navigate and interact with its content.

This thesis is organized around two major contributions. We start with a large-scale characterization of the navigation patterns on Wikipedia in English, and then we introduce the tools we developed to conduct our analyses.

In the first part, we shed light on the navigation patterns with three large-scale studies based on passively collected digital traces. Using billions of requests collected in Wikipedia's logs, we measure how readers reach articles, transition between pages, and leave the platform. We provide a complete overview of the readers' behavior by characterizing the frequent navigation dynamics and the level of engagement with different types of external links on the page. Then, given the observed role of Wikipedia as a gateway to the Web, we quantify the hypothetical economic value of the traffic received by external websites.

In the second part, we present the tools that we developed to make our analysis possible and support future work in this field. First, we introduce WikiPDA, a cross-lingual topic modeling method able to generate a shared topics space for all editions of Wikipedia. Then, we present WikiHist.html, an effort to make publicly available the full Wikipedia history in HTML format.

We conclude by discussing the implications of our findings and presenting future research opportunities enabled by our contributions.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Evolution has optimized humans for knowledge-seeking, and humans have in turn optimized the world around them to facilitate access to knowledge. Many of the most consequential evolutionary, cultural, and technological advances in humans have enhanced their ability to find, ingest, process, and transfer knowledge. From the development of language and writing systems to modern telecommunication, humans are constantly pushing the boundaries of knowledge sharing.

In our history, as part of this constant effort of sharing knowledge, encyclopedias have played a crucial role. Since antiquity, humans have developed ways to keep track and share what we know about the world. From the ancient Pliny's *Naturalis Historia* that served as an editorial model to the development of the modern concept of encyclopedia in the France of the enlightenment, this effort served the same ideal. In the 18th century, the philosopher Denis Diderot defined encyclopedias as a way to disseminate knowledge to people that live with us and will come after us, in a virtuous cycle *"so that the work of preceding centuries will not become useless to the centuries to come"* [42].

Fast-forward to the last century, with tremendous technological progress, the way we think about accessing knowledge changed drastically. In 1945, Vannevar Bush [23] sketched his vision of an information management device —the Memex— that would allow users to retrieve information quickly and enhance their memory by interlinking documents following the associations in the human brain. In the last decades, digitalization brought us close to his visionary idea. From the availability of expert-curated encyclopedias on memory support like CD-ROMs such as Microsoft Encarta to the development of online crowdsourced Web encyclopedias like Wikipedia, our access to knowledge became ubiquitous and effortless.

Given the central importance of knowledge seeking to human nature —epitomized by the view of humans as informavores [115]—, understanding how humans seek information and engage with knowledge is of key significance across disciplines, both in the basic and applied

sciences. In the basic sciences, biologists, psychologists, anthropologists, among others, stand to gain fundamental insights into how humans function, whereas in the applied sciences, such insights can enable the design of more effective tools and information environments, such that humans can more readily find relevant knowledge in an ever-surging flood of information.

In recent years, with the increase of computer literacy and access to the Internet worldwide, the Web has become a common destination to find the information we need. Still, despite the extent of this behavior, little is known about humans' strategies when looking for knowledge online. This gap in our understanding naturally begs the question: ***How* do people seek knowledge online?**

To address this fundamental question, Wikipedia plays a crucial role. Besides simplifying and democratizing access to knowledge, Wikipedia represents the ideal candidate to investigate human behavior around knowledge. Online knowledge-seeking is a complex process that involves search engines, browser history, and bookmarks, and it may refer to a variety of different information needs. This thesis focuses on **encyclopedic knowledge-seeking**, which represents an important special case of human knowledge-seeking. Thanks to a rich network of concepts that people can navigate and interact with, researchers can finally collect data that gives us an unprecedented view of human behavior around encyclopedic content. The diverse set of elements that Wikipedia articles contain, such as links, references, and infoboxes, can help researchers unveil all the facets of the readers' information needs.

A crucial aspect that supports our effort in modeling knowledge consumption in the wild is that Wikipedia is accessed daily by millions of people worldwide. These visits leave a digital trace in the usage logs. After the proper anonymization, the traces can be used to obtain a comprehensive behavioral overview. By analyzing these digital fingerprints available in the logs, we can access a level of detail not possible before. For example, in contrast to in-lab studies, using the logs allow us to study how users navigate the content in a realistic setup by accessing a rich set of geographical and temporal properties without altering the readers' experience. In our novel work, we systematically analyze the encyclopedic knowledge-seeking patterns by leveraging large-scale datasets collected from English Wikipedia's logs. We studied digital traces both from the server and the client-side, which offer novel insights on human interactions with knowledge.

Since these logs are passively collected, they are uniquely suited for providing a complete mirror of real-world, self-motivated encyclopedic knowledge seeking. In addition, they offer a way to link requests of the same readers, allowing us to combine multiple pageloads into sessions and study how the navigation evolves within a session. Ultimately, given the large-scale volume of this data, which includes the activities of millions of readers, we can model the user behavior at a population scale and measure subtle behavioral patterns and small-sized effects, which would not be detectable via traditional methods [161]. Chapter 2 put our work and additional advantages of this data in the wider context of previous research. Relying on large-scale passively sensed user traces may have its downsides compared to traditional

methods. We consider and discuss the advances and disadvantages in the light of our findings in Chapter 8.

This thesis aims to advance our comprehension of the dynamics associated with online knowledge consumption with profound implications for Wikipedia and the web experience beyond Wikipedia. In practice, complementary to previous work focused on the motivations to visit Wikipedia [108, 171], this thesis aims to elucidate the mechanism of content consumption. In other words, differently from previous research focused on "why" people use Wikipedia [108, 171], we expand the understanding of knowledge-seeking patterns by focusing on the question of "how".

Describing the mechanics of how we access online content has a significant impact on the work of different players. Researchers interested in modeling online human behavior can benefit greatly from knowing how Web users behave in the wild to develop novel models about information consumption. At the same time, platform designers aware of the common patterns used to navigate content can envision new ways to improve the experience on the platform and implement a workflow more aligned with the readers' information needs. Additionally, the entire Wikipedia platform can benefit from a deeper understanding of the readers' behavior. Wikipedia is a dynamic system where readers, editors, and content are intimately connected in a self-reinforcement loop [52]. Improving the readers' experiences and increasing content consumption leads to more community involvement. The editors of Wikipedia, conscious of the type of content that readers need, can make informed decisions and adapt the priority of the articles and the portion of the pages that need improvements. In turn, better and more comprehensive content cause to increase the consumption thanks to more readers finding what they need on Wikipedia.

Furthermore, this thesis touches also upon the broader value of Wikipedia. Wikipedia is an integral part of the Web ecosystem, and this thesis offers the first evaluation of its economic value as a gateway to the broader Web. By estimating the value in monetary terms of the traffic relaid by Wikipedia to external websites, we add a critical piece of evidence to the often-underestimated discussion about the importance of Wikipedia for the Web.

Lastly, the content of this thesis offers additional contributions to Wikipedia research, such as methodological advances beneficial to the effort in modeling Wikipedia. The final part of this thesis describes WikiPDA, a cross-lingual topic model, and WikiHist.html, a dataset with the full revision history in HTML that we developed to support our work and then publicly released to foster further research.

In summary, our work offers a large-scale quantitative overview of how we consume online knowledge relying on passively sensed digital user traces. This work also empowers researchers with tools they can use to understand our habits and envision how to improve our Web experience. Below, concrete scientific contributions are outlined.

1.2 Summary of contributions

This thesis is focused on two major contributions. The first part provides a complete overview of the navigation patterns, the interactions with the references, and external links. We aim to offer a comprehensive view of the natural readers' behavior on Wikipedia by focusing on three stages of their interaction: reaching the content, navigating it, and leaving the platform. We characterize the features associated with users' interest and the volume of traffic incoming and outgoing from Wikipedia—both in terms of page loads and estimated economic value. Second, this thesis introduces two contributions that we developed to support these studies and that we release to the community: a cross-lingual topic model and a large-scale dataset with the entire Wikipedia history in HTML format.

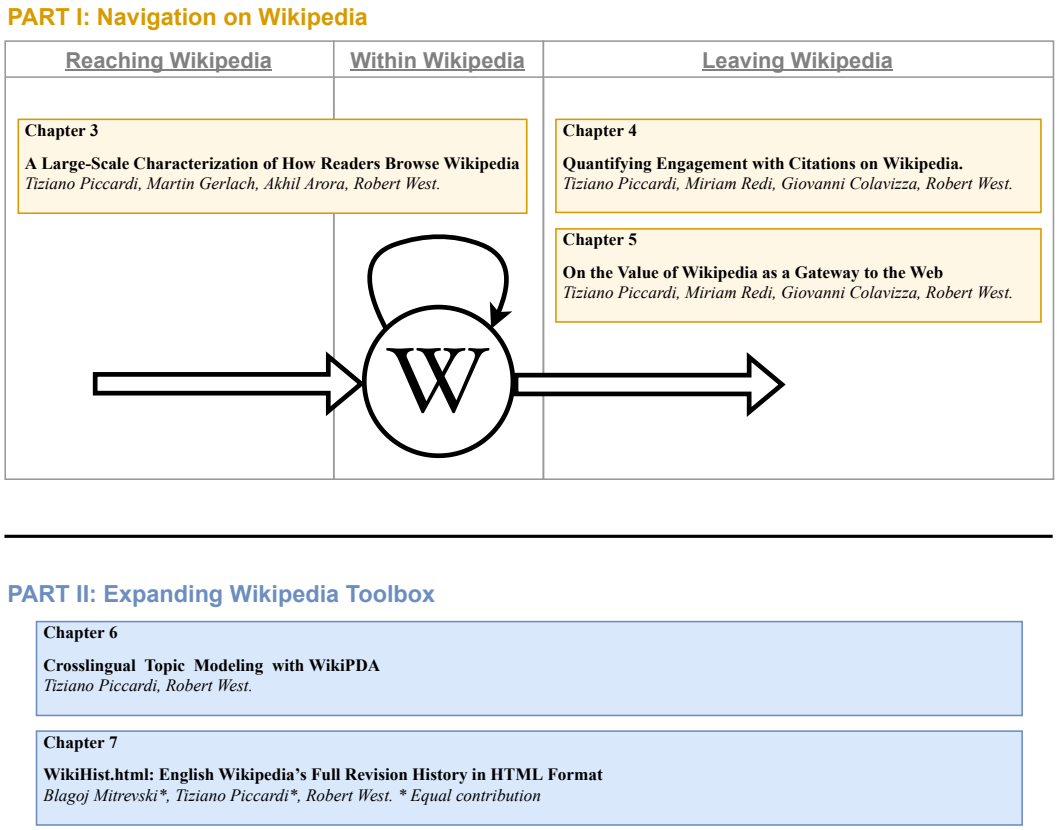


Figure 1.1: Thesis outline

The document structure is as follows. Chapter 2 provides an overview of the related work by putting our contributions in context. Then, Chapter 3 describes a large-scale analysis based on server logs focused on how readers reach Wikipedia and navigate its content. Chapter 4 and 5 describe how readers leave Wikipedia by characterizing how users engage with citations and external links in the articles. Finally, Chapters 6 and 7 introduce WikiPDA and WikiHist.html, two tools that we developed to support our work. Fig. 1.1 summarises the outline of this thesis with the original publication that served as the basis for each chapter.

1.2.1 How Readers Browse Wikipedia (Chapter 3)

Adapted from

A Large-Scale Characterization of How Readers Browse Wikipedia
Tiziano Piccardi, Martin Gerlach, Akhil Arora, Robert West.
arXiv preprint arXiv:2112.11848 (2021)

To understand how people navigate content on Wikipedia when seeking information, we characterize the consumption patterns with a large-scale quantitative study. Using billions of page requests from Wikipedia's server logs, we measure how readers reach articles, move between articles, and combine these patterns into more complex navigation paths. We organize this study in three steps. First, we investigate the individual requests to observe how articles are reached from internal or external resources. Then, we aggregate the requests of the same reader sorted by timestamp and describe the common consumption patterns. Finally, we introduce two approaches to aggregate the navigation sessions as trees using the HTTP referrer, and as sequences using temporal proximity. We characterize the behavior based on both methods and list their advantages and disadvantages.

The key findings can be summarised as follows:

1. *Search engines are the navigation hubs.* The most common way to access Wikipedia is through a search engine. Additionally, we found evidence that they play an important role in the navigation within Wikipedia. Even when the link to the desired page is in the current article, often readers tend to transition to the next content by using external searches.

2. *Revisiting patterns are frequent.* As observed in previous studies on Web consumption, readers tend to access the same content frequently. When a reader loads a new article on Wikipedia, in 11% of the cases, it is the same as the previous one.

3. *Time and device are associated with different behavior.* The broad coverage of Wikipedia allows it to fulfill different information needs. Articles associated with entertainment receive more attention during the evening and night, while articles about STEM are loaded more frequently during the day. At the same time, the volume of traffic from mobile is more than double compared to desktop devices during the evening.

4. *Sessions are short.* Most of the sessions (68–78% depending on the aggregation method) are composed of a single pageload.

5. *Different topics have different navigation patterns.* The topic of the first page is a predictor of the properties of the navigation session. Readers who reach Wikipedia to read an article about entertainment load more pages compared to starting from STEM content.

6. *Navigation has a higher chance to terminate in low-quality articles.* The last pages of internal navigation generated by a sequence of clicks tend to be of lower quality than the average. This finding suggests that poor quality content is associated with the abandonment of the current exploration path.

1.2.2 How Readers Engage with Citations on Wikipedia (Chapter 4)

Adapted from

Quantifying Engagement with Citations on Wikipedia.
Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, Robert West.
Proc. of The World Wide Web Conference (WWW) 2020

As an encyclopedia, Wikipedia is not a source of original information but was conceived as a gateway to secondary sources. According to Wikipedia's guidelines, facts must be backed up by reliable sources that reflect the full spectrum of views on the topic. Although citations lie at the very heart of Wikipedia, little is known about how users interact with them.

To close this gap, we built client-side instrumentation for logging all interactions with links leading from English Wikipedia articles to cited references during one month and conducted the first analysis of readers' interaction with citations on Wikipedia. We use matched observational studies of the factors associated with reference clicking to identify the causal relation between features of the article and level of engagement.

The key findings can be summarised as follows:

1. *Engagement on citation is low.* About one in 340 page views results in a reference click (0.29% overall; 0.56% on desktop; 0.13% on mobile).
2. *Readers engage more with references in low-quality articles.* Clicks occur more frequently on shorter pages and pages of lower quality, suggesting that references are consulted more commonly when Wikipedia itself does not contain the information sought by the user.
3. *Page popularity is associated with less engagement.* Overall, the popularity of an article is associated with less engagement, suggesting that different types of readership for popular content.
4. *Different topics have different levels of engagement.* Referenced content about recent events, people's lives, and offering open access receive more clicks. Readers show more interest in links containing a recent date in the description or containing words such as "married", "wife", "dies", "pdf", and "free".

1.2.3 On the Value as a Gateway to the Web (Chapter 5)

Adapted from

On the Value of Wikipedia as a Gateway to the Web
Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, Robert West.
Proc. of The World Wide Web Conference (WWW) 2021

By linking to external websites, Wikipedia can act as a gateway to the broader Web. To understand its role in the larger picture of Web navigation, we perform a detailed analysis of usage logs gathered from Wikipedia users' client devices.

First, we characterize the general engagement with all external links on the page in terms of clicks volume and speed — defined as the time gap between the page load and the first click. We organize the external links according to their location in 3 groups: infobox, article's body, and references. Then, we focus on the official links available in the infoboxes. Since they represent the clear intention to know more about the entity described in the article, we create a classifier to identify them and measure their level of engagement. Finally, we conclude by quantifying the traffic that Wikipedia relays to these official websites in economic terms.

The key findings can be summarised as follows:

1. *Infobox links have higher and faster engagement.* Infoboxes typically contain a summary of key facts about the entity described in the article, and they always appear at the top of the page. Readers engage more this the links available in this area (30 times more than with references) and faster, with a median time of around 18 seconds (compare to 51 seconds of the references).

2. *Official links are a special case.* Official links listed in infoboxes have by far the highest click-through rate, with a click-through rate of 2.47%, compared to 0.03% of the references. In particular, official links associated with articles about businesses, educational institutions, and websites have the highest CTR. In contrast, official links associated with articles about geographical content, television, and music have the lowest CTR.

3. *Wikipedia acts as a stepping stone for Web navigation.* We investigate patterns of engagement with external links, finding that Wikipedia frequently serves as a stepping stone between search engines and third-party websites, effectively fulfilling information needs that search engines do not meet.

4. *The traffic relaid by Wikipedia would be worth several million.* Use used Google Ads to quantify the hypothetical economic value of the clicks received by external websites from English Wikipedia. The website owners would need to pay between $7 and 13 million per month to obtain the same volume of traffic via sponsored search.

1.2.4 Crosslingual Topic Modeling with WikiPDA (Chapter 6)

Adapted from

Crosslingual Topic Modeling with WikiPDA
Tiziano Piccardi, Robert West.
Proc. of The World Wide Web Conference (WWW) 2021

Extracting a list of topics from a set of documents is a common task that enables researchers to answer a broad set of questions. In the case of Wikipedia articles, examples span from measuring the semantic distance between documents, investigating how Wikipedia covers different subjects, and monitoring evolving trends. A common approach is to rely on Latent Dirichlet allocation (LDA) that generates a probability distribution over a set of topics learned by the textual corpus. A substantial limitation of using this approach off-the-shelf on Wikipedia is its inability to learn shared topics for all the 300 languages available. LDA relies on a bag-of-words model, and the basic assumption is that the training corpus is in the same language.

To bridge this gap, we present Wikipedia-based Polyglot Dirichlet Allocation (WikiPDA), a cross-lingual topic model that learns to represent Wikipedia articles written in any language as distributions over a common set of language-independent topics. It leverages the fact that Wikipedia articles link to each other and are mapped to concepts in the Wikidata knowledge base, such that, when represented as bags of links, articles are inherently language-independent. WikiPDA works in two steps, by first densifying bags of links using matrix completion and then training a standard monolingual topic model. A human evaluation shows that WikiPDA produces more coherent topics than monolingual text-based LDA, thus offering cross-linguality at no cost. We demonstrate WikiPDA's usefulness in two applications: a study of topic biases in 28 Wikipedia editions and cross-lingual supervised classification. Finally, we highlight WikiPDA's capacity for zero-shot language transfer, where a model is reused for new languages without any fine-tuning.

The key findings and our contributions can be summarised as follows:

1. *Each Wikipedia language edition has measurable topic biases.* We measure the distance of Wikipedia editions, finding that the topic similarity can be attributed both to geographical and cultural proximity (e.g., Spain-Portugal) and to technological choices such using the same bots (e.g., Lsjbot).

2. *WikiPDA can be used for supervised training and zero-shot language transfer.* The resulting topics vectors can be used to train supervised models. This property allows extending monolingual models like ORES [67], the official Wikimedia topics' classifier, to all languages. Additionally, WikiPDA can do zero-shot language transfer. The representation based on bag-of-link allows applying the same model to new languages not used in the training set.

3. *WikiPDA is open-source.* We release the full method description with the pre-trained models and a library that can generate topics distribution and prediction of ORES labels.

1.2.5 English Wikipedia's Full Revision History in HTML Format (Chapter 7)

Adapted from

WikiHist.html: English Wikipedia's Full Revision History in HTML Format
Blagoj Mitrevski, Tiziano Piccardi*, Robert West. * Equal contribution*
Proc. of Conference on Web and Social Media (ICWSM) 2020

Wikipedia is written in the wikitext markup language. When serving content, the MediaWiki software that powers Wikipedia converts wikitext to HTML, thereby inserting additional content by expanding macros (templates and modules). Hence, researchers who intend to analyze Wikipedia as seen by its readers should work with HTML rather than wikitext. Since Wikipedia's revision history is publicly available exclusively in wikitext format, researchers have had to produce HTML themselves, typically by using Wikipedia's REST API for ad-hoc wikitext-to-HTML parsing. This approach, however, (1) does not scale to very large amounts of data and (2) does not correctly expand macros in historical article revisions. We solve these problems by developing a parallelized architecture for parsing massive amounts of wikitext using local instances of MediaWiki, enhanced with the capacity of correct historical macro expansion.

The key contributions can be summarised as follows:

1. *WikiHist.html is a public dataset.* By deploying our system, we produce and release WikiHist.html, English Wikipedia's full revision history in HTML format. We publish the full dataset and the code to reproduce our results.
2. *WikiHist.html allows having a full picture of the evolution of Wikipedia.* We highlight the advantages of WikiHist.html over raw wikitext in an empirical analysis of Wikipedia's hyperlinks, showing that over half of the wiki links present in HTML are missing from raw wikitext and that the missing links are important for user navigation.

2 Background and related work

General information seeking and human navigation on the Web have been investigated largely, producing a rich body of literature. Our work is strongly related to previous work aiming to understand how people interact with online content. Specifically, our work falls in the context of modeling information-seeking behavior and navigation patterns on the Web (Sec. 2.1) with particular attention to Wikipedia (Sec. 2.2). This section is organised around three navigation stages: how readers reach Wikipedia (Sec. 2.2.1), navigate the content (Sec. 2.2.2), and leave the platform (Sec. 2.2.3).

Additionally, to conduct our studies, we developed tools and datasets that we released to contribute to the current landscape of resources for researches (Sec. 2.3). This chapter reviews the previous work and the connection with our contributions.

2.1 Information seeking and Web content

Given the Internet diffusion, a significant portion of the information we consume comes from the Web. This section provides an overview of the models developed to describe our general information-seeking behavior and how we interact with content on the Web.

2.1.1 Information-seeking behaviors

In the past, information-seeking behavior received attention from sociologists and cognitive psychologists. In the 80s, Wilson [208] popularised the concept of *information needs* and defined a model to describe our behavior when we look for information. He realized that information needs are challenging to observe and their definition unclear and hard to formalize, but at the same time, the strategies we use to find a piece of information are observable and easier to model. The model proposed is designed for offline information seeking, but it incorporates generic roles such as *information systems* and *information resources* that can be applied to the online world. Wilson kept evolving the model with additional revisions [206, 207] to conciliate it with the progress in information science.

A different but complementary approach coming from cognitive psychology argues that despite being in the digital age with tremendous progress in multiple aspects of our daily lives, many human behaviors are rooted in the needs of our animal ancestors. Machlup [115] described humans as *informavores*, comparing our need to find and consume information in a similar way as we need food. This idea inspired Pirolli *et al.* to develop the information foraging theory [145], that applying the behavioral ecology model of optimal foraging to information, describes humans behaving as predators in the information space.

Animals, in their search for food, tend to prefer strategies that maximize the food intake with the lower investment of time and energy. Similarly, when seeking information, we rely on our intuition —or "built-in" foraging strategies— to pick the best path. This idea was formalized by Chi *et al.* [28] with the concept of information scent that explains how we find this information. Like predators following scents to find the food they need, we look for cues to find the paths that maximize the chances of leading us to the desired piece of information. When our foraging mechanism picks up the information scent, we follow it, whereas when it loses its strengths and our expectation of success decreases, we lose interest in that path and look for a different source.

More recently, researchers moved their attention to the Web and the application of information foraging theory to information networks and the social Web. They investigated how this concept can be used to increase community engagement on Twitter [178] and Facebook [159], and guide the Wikipedia editors community [117].

Complementary to this line of research, Kitajima *et al.* [91] propose a cognitive model to represent the information-seeking behavior in the Web. They propose a theoretical framework focused on the comprehension of the text and images as the driver of the navigation. They define the scent followed by the users as the relatedness of the link or image with the desired goal. They describe that the chances the user will follow a path are associated with the similarity —in semantic space— of the link with the destination, the user's familiarity with the path, and the literal matching with the text.

2.1.2 Navigation patterns on Web

We spend a large fraction of our digital life browsing the World Wide Web. Since its mass adoption in the mid-90s, researchers have invested time and resources to model its topology and how we browse it. Understanding these properties has deep implications for the development and evolution of the Web, the design of search engines, and the economy driven by online businesses.

Structure of the Web. In early works, Kumar *et al.* [98, 99] formally defined the concept of Web Graph as a directed graph composed by hyperlinks and systematically analyzed the properties of the early network. They observed how the in- and out-degree of Web pages follow low power distributions, and that the Web has a topology that resembles a bow tie. Many documents

belong to a giant strongly connected component, and two sets of comparable size point either toward the center of the graph or its periphery. Recent work [54] shows that this model is still relevant today with the addition that the large bow-tie structure can be decomposed into local mini bow-ties topologies. In complementary works, Albert *et al.* [4] focused on measuring the diameter of the Web, discovering that two random documents could be reached on average by clicking 19 links, and Huberman *et al.* [81] described a universal power-law distribution that model the number of pages per website.

Modeling users' navigation. Characterizing the user navigation on this massive graph is a challenging task because of the limited availability of data, the technical difficulties, and the important privacy implications. Previous work focused on modeling navigation patterns based on passively collected server logs of large websites or by using modified browser versions, including extensions like toolbars.

An investigation from the early days of the Web shows that users navigate a small area within the visited websites with frequent backtracking patterns using the browser's back button [25]. A recurrent finding is that people tend to revisit the same content multiple times [6, 181] making recency is a strong predictor of the next pages visited. This behavior has been observed in many online activities, such as search engine logs [183, 190], browsing websites [1, 180], and consumption of multimedia content [18]. Studies show that up to 40% of the searches received by Yahoo are repeated queries and that navigation histories contain between 50% and 80% of pages visited multiple times [31, 75, 180]. It has been observed that the rate of revisiting web pages is associated with the frequency their content is updated [2]. People tend to abandon this repeat consumption behavior when the time gap between the revisit events increases, interpretable as a loss of interest of the user that is getting bored of the content [18]. Additionally, researchers show that visual properties impact the navigation of the users on the Web. The position of a link can impact the level of engagement on search engines results [35], and e-commercial content [91].

Fu *et al.* [53] used information foraging theory to build a cognitive model of the users navigating on a website. They implemented a program that accurately simulated the behavior of a human on two sample websites. The model decides to click a link by estimating its relevance for the content navigation. They validated their findings with 74 real navigation traces, observing a high level of accuracy.

In recent work, Crichton *et al.* [36] described the navigation traces of more than 250 Web users, observing that the navigation patterns evolve over time and that it is highly centralized with 50% of the Web consumption happening on 1% of the websites.

Previous work also investigated user navigation in the context of search engine usage. In a longitudinal log-based investigation of the navigation following a web search, White *et al.* [205] found that people's behavior shows a high level of variability. Using navigation traces of thousands of users for five months, they found that navigation behavior can be grouped

into two classes of extreme users, namely navigators and explorers. Regardless of what they searched, navigators tend to revisit the same domains and progress sequentially in a predictive way. On the other hand, explorers tend to submit multiple queries during navigation and jump between multiple domains.

Other approaches to model the behavior of users on the Web come from applications of graph theory, and they focus on the mechanics of how the information is accessed. A prevalent example is the random surfing model that describes how users move across the information space of the Web, not following only the available links but by frequently jumping from location to location by entering the URL of the destination directly. The model assumes that the user tends to stay in the neighborhood of the page where they land on the websites and quickly lose interest and leave. Geigl *et al.* [56] found that with the rise of search engines that send the user directly to the desired content, navigation is often limited to the landing page and can be approximated with a random surfer model.

Overall, these findings show that human mobility on the Web has predictable patterns [97], and many models have been developed to predict the user intention or next actions. Although researchers found that Web users are not strictly Markovian (the page visited next does not depend exclusively on the current state) [29], many prediction models approximate the navigation of users on a network with Markov chains [39, 114, 146] and hybrid models [13, 83, 90, 131].

2.1.3 Content engagement

Expanding our knowledge on the patterns associated with online content consumption also has important implications for understanding and modeling engagement. Being able to quantify user engagement is crucial for websites, especially for those with an advertising-based business model [10]. Researchers from various fields have investigated ways to define what engagement means in the online world [136] and to measure users' attention, interest, and interaction with websites [10].

The concept of engagement is application-specific and customized for the intended purpose of the platform. Common metrics used to assess the level of interaction or interest for online content include click-through rate (CTR), representing the ratio of clicks to impressions of a link, and dwelling time [111], commonly defined as the time spent consuming the content before user actions. Other works have tried to predict engagement with content in social media based on social interest metrics, such as the number of post comments or likes [14, 30, 80]. Researchers in information retrieval have also investigated methods to estimate users' satisfaction and engagement with textual and visual Web search engines [84, 174, 217]. In computational advertising, existing works have tried to improve ad serving based on target engagement metrics [16, 212], or to predict ad click-through rates directly [110].

2.2 Navigation on Wikipedia

Wikipedia exists in more than 300 languages, but most investigations are based on the English edition of the platform. With its knowledge base of more than 6M articles, the English version is the most visited language edition with the broadest coverage. In January 2019, Wikipedia in English had around 171M links added manually by the editors in the wikitext, raising to 475M [125] when considering all the links readers could use to navigate using the HTML representation. A giant component dominates the global network topology, i.e., the largest strongly connected component includes more than 55% of the nodes. In line with the finding of "recursive" bow-tie structure [54], the topology of the Wikipedia links network resembles itself the organization of the Web [73, 101]. Wikipedia is semantically rich in content with a dense network of links, making it the ideal candidate to study how we navigate a knowledge space and interact with information.

2.2.1 Getting into Wikipedia

Wikipedia's traffic is influenced by its connections to the larger Web ecosystem and its interdependence with external platforms. Like every web resource, besides typing the URL explicitly, people can reach its articles by various origins, such as clicking links available on external websites, social media, or retrieved through a search engine.

In particular, search engines relay most of the incoming traffic received by Wikipedia, representing the preferred way to access its content. The relationship between Wikipedia and search engines was investigated by McMahon *et al.* [120]. They focused on Google and described the mutual dependency existing between the two platforms: Wikipedia contributes to Google's success by answering a large portion of the queries posed by its users, and Wikipedia depends largely on Google for its readership. At the same time, they found a critical tradeoff in this relationship. On the one hand, Wikipedia's content improves Google search results, for example, via content snippets like knowledge panels; on the other hand, this might keep users that already satisfied their information need from visiting Wikipedia itself. In follow-up work, Vincent *et al.* [191] investigated further this connection observing that Wikipedia articles are very frequent on the first page of Google results, accounting for 67%-84% of the queries; depending on the type of searched content. These findings are consistent with a similar observation described in work conducted more than ten years earlier by Laurent *et al.* [104] on Google, Yahoo, and MSN that showed that Wikipedia appeared on the first page of results for 71%-85% of medical queries. This interdependency between search engines and Wikipedia also results in a high correlation between the volume of searches for one term and pageloads of the relative article. Yoshida *et al.* described this phenomenon, observing a correlation of 0.72% between the data obtained through Google Trend and the Wikipedia pageloads .

Besides search engines, links to Wikipedia are frequently posted on social media and Q&A websites. A study by Gómez *et al.* [61] found that Wikipedia is the second most common domain in the links posted on StackOverflow, the popular Q&A website for programmers. This relation

with StackOverflow was further explored by Vincent *et al.* [192] to estimate the value received by the 2 counterparts. They found evidence that the StackOverflow community benefits from the open knowledge available on Wikipedia by observing that the posts containing links to one of its articles have around double the level of engagement. Additionally, they found a similar dynamic for the links posted on Reddit. The posts containing a link to Wikipedia on Reddit have up to 5 times more upvotes and generate twice as much discussion compared to the website average. The benefit offered in the other direction, from these platforms to Wikipedia, is less obvious. The presence of links in external platforms has an influence [129] on the attention received by Wikipedia in terms of raw pageloads. However, this interest does not last, and it is not translated into actions since Wikipedia does not record an increase of edits [192] for the linked articles. This aspect was also investigated by Morgan *et al.* [128] in a study aiming to quantify if articles that experience a spike of attention caused by social media are at high risk of vandalization. They investigated the impact of the traffic received from Facebook, Twitter, Reddit, and YouTube, finding that the views received are not converted into edits that would justify the need for more patrollers.

2.2.2 Within Wikipedia

Motivations and content popularity. Given the central role of Wikipedia for our access to knowledge, in recent years, researchers investigated the motivations and patterns associated with the consumption of its content. In a foundational work, Singer *et al.* [171] investigated why people read Wikipedia. Basing their analysis on data collected through surveys, they found that readers are motivated by a variety of different factors such as current events, media coverage of a topic, personal curiosity, work or school assignments, or boredom. By examining the participants' activities in the server logs, they observed different behavioral patterns based on their motivation. People exploring Wikipedia out of boredom tend to have long sessions with fast transitions between articles, whereas readers interested in learning a subject spend more time on a few relevant pages. In follow-up work, Lemmerich *et al.* [108] extended this analysis across 14 languages finding that Wikipedia has an important role as a source for knowledge for countries with low Human Development Index, where the readers exhibit an in-depth reading behavior. This finding was confirmed by TeBlunthuis *et al.* [182] that using client-side instrumentation investigated the time readers spend reading an article. They found that readers from the Global South spend more time reading the content of the articles than users from other geographical locations.

Similar work in modeling how people behave when reading Wikipedia conducted by Lehmann *et al.* [107] concluded that Wikipedia users have reading patterns grouped into four categories: exploration, focus, trending, and passing. Additionally, they found a misalignment of attention between readers and editors, observing that the most popular articles are not always the most edited. They discovered that articles related to entertainment are the most popular content on Wikipedia, confirming a previous work conducted by Spoerri [175] more than ten years earlier that showed that at least half of the most visited articles are about entertainment and

sexuality. In a similar work focused on the Australian population, Waller [197] reached similar conclusions measuring the highest popularity for content associated with popular culture such as music and TV shows. Additionally, they described that different population segments have diverse information needs, and the ratio of the topics read varies based on their lifestyle.

Wikipedia readers also exhibit preferences in the type of images on the page. Rama *et al.* [151] observed that one in 29 page-loads results in a click on at least one image, and readers click more on images associated with articles about visual arts, transports, and biographies of less well-known people. Images play an important role also in content navigation: links with an image in the preview tooltip have a lower click-through rate, suggesting that the readers satisfy their information need from the image without loading the entire article.

Other works on modeling the content popularity observed that articles experience sequences of bursts of attentions caused by external factors [152] such as an Academy Award nomination and that number of page views is sensitive to internal design changes like the introduction of the preview feature [27]. Similarly, during the COVID-19 pandemic, exogenous factors like introducing mobility restrictions impacted the type of content people sought on Wikipedia. Ribeiro *et al.* [157] observed an increase of topics associated with entertainment. However, not all interventions impact content popularity. The awareness campaign to promote Wikipedia in Hindi [27] showed, for example, no significant change in the traffic recorded.

Natural navigation. The analysis, modeling, and prediction of human navigation *inside* Wikipedia has been considered in previous studies [43, 60, 74, 102, 170, 189]. Multiple approaches have been used to study human navigation on the platform. In early works Reinoso *et al.* [155] use server logs to characterize the traffic recorded over six months. They described the daily and weekly patterns and illustrated how this data could be exploited to provide valuable insights to understand Wikipedia readers' behavior and design an efficient and scalable infrastructure [156]. One of the downsides of working with raw server logs is the need for privileged access to sensitive information such as IPs and geo-locations that, for privacy reasons, should be granted with care. To promote research in this direction and overcome this limitation, the Wikimedia Foundation releases with monthly frequency the public clickstream [210] for multiple languages editions. This dataset contains transition counts for pairs of articles, giving researchers valuable insights into how readers move from article to article. The clickstream is an aggregated and filtered version of the server logs to preserve the readers' privacy, but in our recent work not included in this dissertation, we proved that it approximates the real navigation with a good level of accuracy [8]. We showed on common tasks such as link prediction and topic similarity that although the differences are measurable and statistically significant, the conclusions obtained using the clickstream are the same as for the private data, and the differences for the metrics used are within 10%.

Researchers used the public clickstream to study how different topics relay more traffic than others [43]. Dimitrov *et al.* [44] found that most pages attract traffic from external sources and not from internal navigation. This finding suggests that a common reader's behavior consists

of an individual lookup of articles and that engaging in long navigation is not common. When the navigation goes beyond the first article, the topic plays a role, and articles about historical military events relay more traffic than content related to architecture. Similarly, Gildersleve *et al.* [59] found that different types of articles such as lists and disambiguation pages relay traffic in different ways by acting as distributors of traffic. Lamprecht *et al.* [102] investigate the impact of the article layout on the navigation, observing that readers tend to click more links located at the top of the page. This positional bias in the user preferences is also described by Dimitrov *et al.* [46]. Additionally, they found that readers prefer links that lead to the periphery of the network and about semantically similar content.

The clickstream was also used to generate synthetic data through biased random walks. Rodi *et al.* [158] generated synthetic navigation sequences by selecting the next article according to the clickstream probability and simulating an organic interruption of the navigation based on the traffic relayed by the article. They described how readers' navigation paths tend to start general and become incrementally more semantically focused at every step.

Other approaches to understanding readers' navigation rely on data shared by volunteers recruited for the study. Lydon-Staley *et al.* [112] used a philosophical taxonomy created to model curiosity to identify different types of behavior on Wikipedia. By asking participants to navigate Wikipedia for 15 minutes every day, they found evidence that the readers' behavior can be classified into two categories historically called hunters and busybodies based on the properties of the network explored.

Targeted navigation. A different approach to characterize human navigation relies on digital traces obtained via *games with a purpose* [193] (GWAP). These games are a popular human-computation technique to collect human-generated data in a gamified environment. They offer a convenient way to collect data providing users entertainment. GWAP have a large set of use cases that go from crowd-sourced protein folding [34] to the investigation on search queries formulation [3]. In the context of Wikipedia navigation, two successful games are Wikispeedia [202] and TheWikiGame [185]. In these games, the players start from a random article and are tasked to reach a target page in as few clicks as possible by following internal links only. The trajectories are then collected as sequences and reveal how people move across Wikipedia content with the advance to overcome the limitation of the public clickstream that models only one step of the readers' navigation.

In contrast to natural navigation, these trajectories, denoted as *targeted navigation* posits an unambiguous definition of success (i.e., reaching the target article). They let researchers determine if the navigation is terminated and study the strategies used by the player to traverse the information network. West *et al.* [200] and Helic [73] found that participants tend to make progress toward the destination in the first part of the exploration by jumping toward high degree nodes. These articles act as hubs of the network and maximize the probability of finding a page closer to the target. Once a hub is reached, people advance to the destination using content features and traverse the semantic space with smaller step sizes. These features can

predict the destination of the search, with important implications for the design of tools that can assist people in reaching the desired content. These navigation strategies make humans very efficient in finding the shortest paths between two concepts on a knowledge network. Interestingly, this high performance does not necessarily require background knowledge on the topic. It has been observed [199] that simple automatic agents relying on basic features of the articles have performance comparable to humans.

Another advantage of a clear termination state is that it enables researchers to model how people drift away from the best path and understand when users will abandon the exploration. In follow-up work, Scaria *et al.* [162] found that in both successful and unsuccessful paths, humans tend to move to high degree nodes. A progressive increase of the semantic distance from the target indicates that the user lost the right track, and out of frustration, the navigation will be interrupted soon after. This finding was further investigated by Koopmann *et al.* [93] that proved that using features from the articles and the underline hyperlinks graph to train an RNN, the success of a navigation game is is predictable from its early trajectory.

The paths obtained through targeted navigation gave researchers valuable insights into how humans navigate information networks, but it does not necessarily represent how readers navigate Wikipedia in a natural setup. As we will see in Chapter 3, natural navigation defined as sequences of internal clicks tends to be short, with the majority composed by a single pageload. Especially in the light of behavioral models like random surfers and the effectiveness of search engines, a frequent behavior consists in leaving Wikipedia and entering again through another external search.

Applications of navigation traces. In terms of applications, navigation traces have proven useful as a tool to improve the website navigability by identifying missing links [101, 138, 201] and other usability issues that normally require the work of domain experts [58]. Similarly, navigation logs can be used to compute semantic relatedness of pages by studying what content is typically accessed together [38, 172].

2.2.3 Leaving Wikipedia

Besides an extended network of internal links, Wikipedia contains many links to external resources. External links enrich articles with additional content that should not or cannot be included in Wikipedia itself. There are various reasons to add external links, with linked content ranging from official websites to news articles used as references and copyrighted material. When readers decide to continue their navigation to external websites by following links available on Wikipedia, they can pick them mainly from three different page areas: infoboxes, the articles' body, and references.

The patterns associated with how readers use these links have not been widely investigated because obtaining this data is challenging. Intercepting clicks to external content requires access to the server logs of websites frequently linked in the articles, like in the case of DOI

URLs [116], deploying client-side instrumentation on Wikipedia or access browsers history logs. This dissertation fills this gap with a comprehensive overview of the dynamics of leaving the platform through external links. Our findings are described in Chapters 4 and 5.

2.3 Tools and datasets for Wikipedia-related research

Beyond its role as a free source of knowledge for millions of readers, Wikipedia is a crucial dataset for scientific development. Wikipedia and all the associated Wikimedia projects are at the core of the research in many different disciplines for the modeling of human behavior and the development of language models, knowledge graphs, and AI models [11, 32, 40, 55, 63, 139, 163, 166]. This crucial role for the research community is confirmed by the large volume of papers published in recent years. As of December 2021, a search on Google Scholar for the keyword *wikipedia* returns more than 2M entries, with 21K of them having the name explicit in the title[I].

To facilitate working with the data from Wikipedia, the Wikimedia Foundation and the close research community regularly release datasets and tools.

Content of the articles. Besides the public dataset describing behavioral patterns of the readers, such as pageviews count and the clickstream introduced in the previous section, the Wikimedia Foundation releases regularly updated datasets with the content of all Wikipedia articles.

Contributors write the articles in a markup language called Wikitext, that the PHP engine of MediaWiki —the software behind Wikipedia— converts for the browsers into HTML. To simplify the analysis of the documents, the analytics team releases a downloadable snapshot in XML format every month containing the Wikitext of the articles in all languages[II]. The archive includes the last revisions of each article at the moment of the data release, which is typically the beginning of each month.

For more resource-intensive longitudinal studies, the Wikimedia Analytics team also releases the dataset with all the revisions since the launch of Wikipedia in 2001. This historical data allows researchers to study how the content on Wikipedia evolved in time. Consonni *et al.* [33] used this data to generate WikiLinkGraphs, a dataset containing the evolution of internal-links network in 9 languages from 2001 to 2018, with a monthly granularity.

Other datasets focus on the citations available on the page. Singh *et al.* [173] released a complete dataset of all the scientific references, whereas Zagovora *et al.* [215] focused on their historical evolution and their contribution to Altmetric score[III].

[I] https://scholar.google.com/scholar?q=intitle:wikipedia
[II] https://dumps.wikimedia.org/
[III] https://www.altmetric.com/

One of the limitation of these datasets is that they are sourced from the original wikitext of the articles, which may be misaligned with the HTML version received by browsers. The engine of MediaWiki that converts the wikitext to HTML allows editors to create modular snippets of code in the form of templates. The code of these templates can be written in wikitext and encoded in a special Wikipedia namespace or obtained by running external modules written in LUA. This dynamic mechanism for page generation makes the editors' work reusable, allowing them to add complex content in the HMTL without writing it explicitly in the wikitext. Examples of content in the articles generated by templates are infoboxes, references, and the links in the navigation boxes. The problem with this approach is that it is impossible to extract the exact content received by the browser —that the readers see— without expanding them or executing the scripts. As described in Chapter 7, we bridged this gap by releasing a dataset containing the full history of Wikipedia in HTML.

Topics and semantic of Wikipedia articles. Documents on Wikipedia cover a large variety of topics. Having a classification taxonomy over the articles allows many tasks, such as improving the platform's organization, understanding what encyclopedic content is available, and helping readers and editors find content. In the past, researchers approached this project in many different ways.

A significant body of work focused on generating taxonomies automatically from the category network of Wikipedia. These categories, curated and assigned to articles by editors, are organized in a hierarchical structure that represents a form of semantic specialization. In an early attempt to use the Wikipedia categories to support the creation of ontologies, Ponzetto *et al.* [148, 149] developed a method to extract a large scale taxonomy by exploiting the implicit *is-a* relation of the categories hierarchy. In a similar spirit, Gupta *et al.* [65] improved the taxonomy generation process by exploiting a set of heuristics on syntactic features of the categories names. Thanks to follow-up improvements[50, 64], these methods can work with more than one language edition. Wikipedia categories have also been used as one of the source datasets to create knowledge bases such as YAGO [176] and DBPedia [11]. By mixing automatic inference and human labeling, these databases describe complex relations between concepts ranging from *subclass-of* to *born-in-year*. A major difficulty of relying on the category network is that it is user-contributed, and it needs careful cleaning before being used [140]. The network constantly evolves, and it has logical issues such as cycles, relations that can not be interpreted as *is-a*, and lacking shared norms in assigning the articles.

Vrandečić *et al.* [194] developed a different approach to represent Wikipedia as a structured knowledge base called Wikidata that involves the community in a collaborative effort. Wikidata is today the knowledge backbone of Wikipedia, and it summarises, in a machine-readable form, the concepts available on Wikipedia. Concepts are described in a language-independent format, and contributors can add properties to enrich the database. Wikidata is described in more detail in Chapter 6.

An alternative approach comes from ORES (Objective Revision Evaluation Service) [67], the official Wikimedia scoring platform. Originally developed to monitor vandalism and the contributions' quality, it evolved into a complete prediction model to infer the topics of any article's revision. The model is trained to return the probability of an article belonging to the 64 classes organized in taxonomy with 4 top categories: Culture, Geography, History and Society, and STEM. Halfaker *et al.*, manually curated the classes by clustering and mapping the WikiProjects to a set of cohesive topics. WikiProjects are portals organized by editors to group articles that belong to the same topic. They come at different levels of granularity, from generic topics like *WikiProject Science* to very specific like *WikiProject Poker*. By mapping the articles assigned to these projects to the relative 64 classes, they trained a model that returns the distribution of probabilities that the article belongs to each of them. Currently, ORES supports only articles from the English edition of Wikipedia, but recent work from Johnson *et al.* [85] aims to expand the model to all the available languages.

Since ORES is based on supervised training, it can accurately assign labels to the input article, but it cannot discover new topics beyond the 64 classes. This topic discovery can be achieved with unsupervised methods such as LDA (Latent Dirichlet allocation) [21] that represents each document as a mixture of topics learned from the corpus. LDA is a powerful generative model, but it has a major shortcoming: its basic implementation does not support topics discovery in a multi-language corpus. Regardless of the semantic of the content, documents in different languages would be considered of different topics. Chapter 6 introduces WikiPDA that overcomes this limitation and allows to obtain a topics' distribution for Wikipedia articles in every language.

Navigation on Wikipedia Part I

3 How Readers Browse Wikipedia

3.1 Introduction

Wikipedia is a unique platform to understand the dynamics of knowledge-seeking online. Given the time we spend online looking for information, comprehending the consumption mechanisms is increasingly critical to learning more about our information needs and designing a better Web experience. In this chapter, we focus on the mechanics of "how" readers consume content on the platform by characterizing how people reach Wikipedia and how they move across its content.

Previous work dedicated to shedding light on human knowledge-seeking behavior has faced important limitations: surveys [214], and thinking-out-loud studies [130] are prone to cognitive biases, like humans generally perform poorly at introspection [134]. Lab-based experiments [112] typically involve small samples consisting of biased populations (e.g., university students) and are thus frequently not representative and might lack statistical power. Studies based on surrogate tasks (e.g., navigation games [202]), although measuring navigation-related skills, do not capture real-world, self-motivated knowledge seeking and may thus lack external validity [137]. Finally, studies based on aggregated and filtered public versions of real-world knowledge-seeking traces (as page-to-page transition counts instead of full traces [46, 59]), although capturing local, page-level choices accurately, may lack relevant reader-specific preferences (e.g., a full navigation trace).

Our study relies on passively collected digital traces, and with billions of monthly views, the representativeness of the human activities recorded in the logs far surpasses any lab-based studies.

In contrast to prior work, which has leveraged Wikipedia's server logs to shed light on specific aspects of reader behavior (including reasons for visiting Wikipedia [108, 171], studying variation in dwell time [182], and measuring geo-localized collective behavior [187]), the present work is the first to employ the logs in a principled, broad analysis to systematically elucidate the nature and structure of encyclopedic knowledge-seeking pathways.

By analyzing billions of navigation traces extracted from the logs (Sec. 3.2), we span three levels of aggregation in our research questions:

RQ1 Unigram level: How do readers reach Wikipedia articles? (Sec. 3.3)

RQ2 Bigram-trigram level: How do readers transition from one article to the next? (Sec. 3.4)

RQ3 Session level: What is the structure of entire reading sessions? (Sec. 3.5)

We find that Wikipedia navigation traces expose a wide variety of structures; although shallow sessions consisting of single pageloads dominate, we observe a long tail of long, complex traces, whose depth and shape vary systematically with topic, device type, and time of day. We highlight that Wikipedia navigation does not happen in isolation, but is embedded in sessions where users transition fluidly to and from the external Web. This aspect, as well as other differences that emerge, distinguishes real-world, in-the-wild Wikipedia usage from the targeted navigation behavior captured by lab-based studies. Finally, we find strong evidence that users stop navigating when reaching low-quality articles.

These results have important implications for Wikipedia and beyond. Understanding how readers explore content on Wikipedia is critical for framing its role in fulfilling information needs and for making design decisions regarding its structure, format, accessibility, and supportive tools such as recommender systems. Going beyond Wikipedia, these findings deepen our understanding of how humans navigate information when seeking knowledge.

3.2 Data

The data sources exploited in this study include user traces mined from Wikipedia's server logs and features extracted from articles.

Pageloads. To study how readers navigate Wikipedia, we analyze the server logs of the English edition collected for four weeks between 1 and 28 March 2021. This data contains an entry for each time a Wikipedia page is loaded. It is automatically collected for analytic purposes on Wikimedia's infrastructure and deleted after 90 days.

We limit our analysis to the pageload requests for articles (MediaWiki namespace 0), filtering out requests from bots. To protect readers' privacy, we remove sensitive information in several steps: discarding pageloads from readers who edited or were logged in during the time of data collection; discarding all requests from countries with at least one day with fewer than 300 pageloads; generating (pseudo) user identifiers by hashing IPs and user-agent strings, as done in previous work [138]; and dropping IP, user-agent, and fine-grained geoinformation. In total, these anonymization steps lead to the removal of around 3% of the data. In addition, we perform the following filtering steps. First, we drop pageloads of the *Main_Page* article, as it does not represent any specific entity. These requests may come from users who set Wikipedia

Origin	Desktop	Mobile	Total
Search engines	45.97%	48.77%	47.71%
Wikipedia			
Articles	35.64%	35.75%	35.72%
Main page	1.65%	0.70%	1.06%
Lang. switching	1.62%	0.50%	0.92%
Categories	0.59%	0.25%	0.39%
Search page	0.38%	0.22%	0.29%
Special pages	0.07%	0.01%	0.03%
Portals	0.03%	0.01%	0.02%
Others	0.01%	0.07%	0.03%
Unspecified origin	12.64%	13.03%	12.88%
External websites	1.36%	0.70%	0.95%

Table 3.1: Statistics of referrers of single pageloads.

as the browser's default page. Second, we remove traffic from massively shared IPs, which would make it hard to study individual activities, by dropping all user identifiers with more than 2,800 pageloads, or on average 100 per day, removing 28k (0.0019%) user identifiers. The final dataset contains 6.52B pageloads associated with 1.47B user identifiers.

Article features. To characterize the content viewed by the readers, we collect a set of article features. To ensure alignment between the server logs and the articles' content, we compute the features for the article revisions of the public snapshot released at the end of March 2021.

We obtain article features such as the number of outgoing links, the PageRank, article We obtain article features such as the number of outgoing links, the PageRank, article quality score, and topic. We assign the quality of the articles using the *articlequality* model of ORES[I] [67], the official Wikipedia scoring platform. This model offers a way to obtain a score [66] that summarises the structural properties of the article, such as the number of sections, references, and the presence of infoboxes. Similarly, for the topic, we use two approaches: (1) ORES [67] *articletopic* model's probabilities for 64 manually curated topics, used for assigning topical labels to articles; (2) WikiPDA [144] topic vectors, used for placing articles in a 300-dimensional topic space.

3.3 Unigram level

We use the term "n-gram" to designate a sequence of n subsequent pageloads from the same user. We start our analysis with unigrams (n = 1) and enumerate how readers can reach Wikipedia articles. We classify Web traffic according to HTTP referrers and quantify the frequency of each type (Table 3.1). In total, 4B (61.5%) pageloads have external or empty referrers and are thus entry points to Wikipedia.

[I] https://www.mediawiki.org/wiki/ORES

Search engines. The most common way to reach the content of Wikipedia is through external search engines, at 3.1B pageloads (47.7% of all recorded traffic, or 77.4% of external traffic). This volume reflects the significant value offered by Wikipedia in fulfilling the information needs of search engine users [7, 191].

Wikipedia. Clicks from other articles account for 35.7% of all traffic. Interestingly, as observed in previous work [125], 6.6% of these pageloads happen through links that do not exist in the link network itself, but likely through other interactions such as Wikipedia's search drop-down menu. Content can also be reached from other pages on the Wikipedia platform: (1) the main page, (2) category pages, (3) Wikipedia's internal search, (4) portals, (5) other Wikipedia pages, including talk pages or pages in other languages (language switching).

Unspecified origin. In 12.9% of all traffic, we observe an empty referrer field. Multiple reasons can produce a request without an explicit origin, including direct access via the browser history, redirects from apps, bookmarks, search toolbars, or when the link source has explicitly turned on the *noreferrer* property.

External websites. In total, 0.95% of the requests originated from external websites that are not search engines nor Wikipedia domains (1.55% of the external traffic). Among those, the most common sources are Facebook (15.6%), Reddit (9.6%), YouTube (8.0%), and Twitter (4.3%).

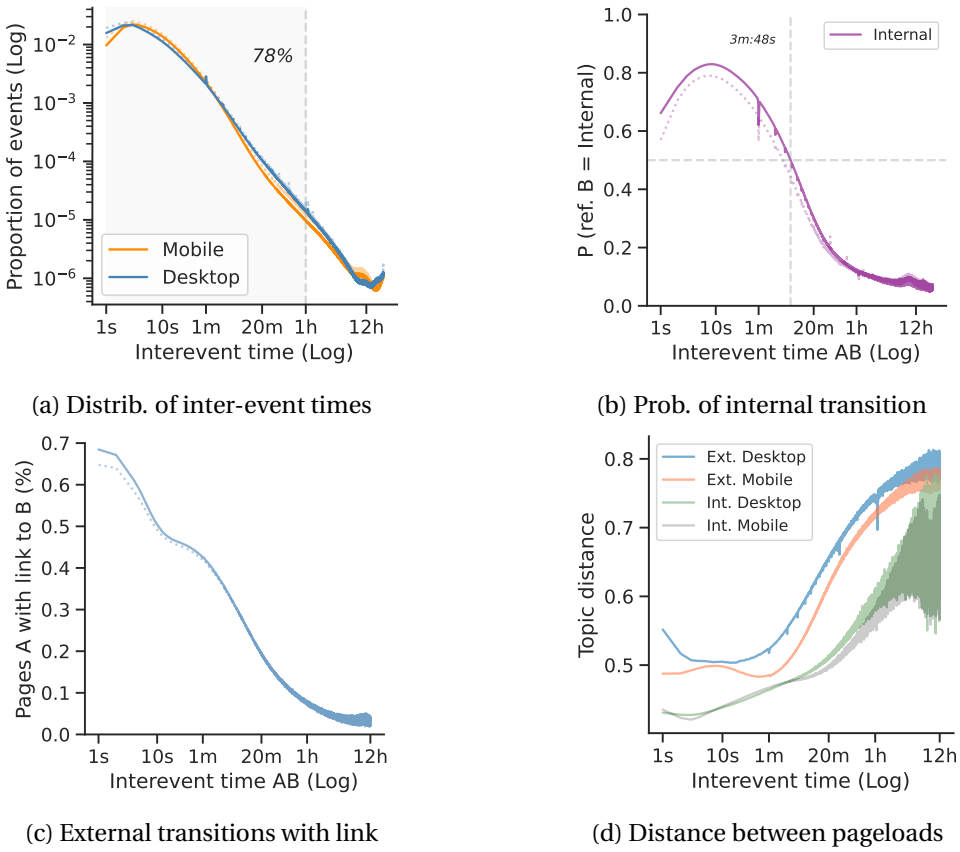
Others. Other external visits (0.015% of the external traffic) come from Android Web views and custom embedded visualizations, with the most common being the Telegram and Reddit sync apps, and Facebook on Android devices.

3.4 Bigram and trigram level

Next, we move from unigrams to bigrams to understand how readers transition between Wikipedia articles. We study events aggregated by user identifier and sorted by time to investigate the properties of consecutive pageloads and their inter-event time. Here it is important to note that the Wikipedia server instructs the browser to disable the cache, such that the server logs contain essentially all pageloads events, including cases when the readers reloaded an article, e.g., by using the back button.

Bigrams. The logs contain 3.95B bigrams (i.e., two subsequent pageloads by the same user with less than one hour in between [68]). The emerging patterns, described next, are summarized in Table 3.2.

The most frequent bigram pattern ("AB" in Table 3.2) corresponds to transitions between two different articles. It can happen both through internal and external navigation (cf. Fig. 3.2). This pattern represents around 89% of all bigrams. The other possible bigram pattern ("AA"



(a) Distrib. of inter-event times

(b) Prob. of internal transition

(c) External transitions with link

(d) Distance between pageloads

Figure 3.1: Statistics of bigrams as a function of the inter-event time between two pageloads. Dashed curves represent the distributions with AA patterns included.

in Table 3.2), corresponds to the consecutive reload of the same article. Representing 11% of all bigrams, it is rather common (84% share the same referrer). This pattern appears at least once in 37% of the navigation histories of readers with at least two pageloads in the month of data collection. This pattern in the log can be generated by different client behaviors (cf. Fig. 3.2), including repeated consumption as described in previous work [18, 181], user activities involving external navigation, or artificial reloads by the browser when a tab unloaded from memory is restored.

Trigrams. Finally, we also briefly consider the 2.98B trigrams present in the logs. The most common trigram pattern (73%, "ABC" in Table 3.2) represents transitions between three different articles. A variety of behaviors can generate this pattern, including sequential clicks or multitab behavior (cf. Fig. 3.2). The second most common trigram pattern (13%, "ABA" in Table 3.2) can be generated by intentionally revisiting the same page or by clicking the back button (cf. Fig. 3.2). In 89% of ABA instances, the first and last event also share the same referrer. The remaining trigram patterns (ABB, AAB, AAA) are combinations of the bigrams described above.

Device	AB	AA	ABC	ABA	ABB	AAB	AAA
Desktop	0.900	0.099	0.749	0.121	0.047	0.049	0.031
Mobile	0.880	0.119	0.719	0.143	0.055	0.053	0.027
Total	0.888	0.111	0.732	0.134	0.052	0.052	0.029

Table 3.2: Frequencies of bigram and trigram patterns.

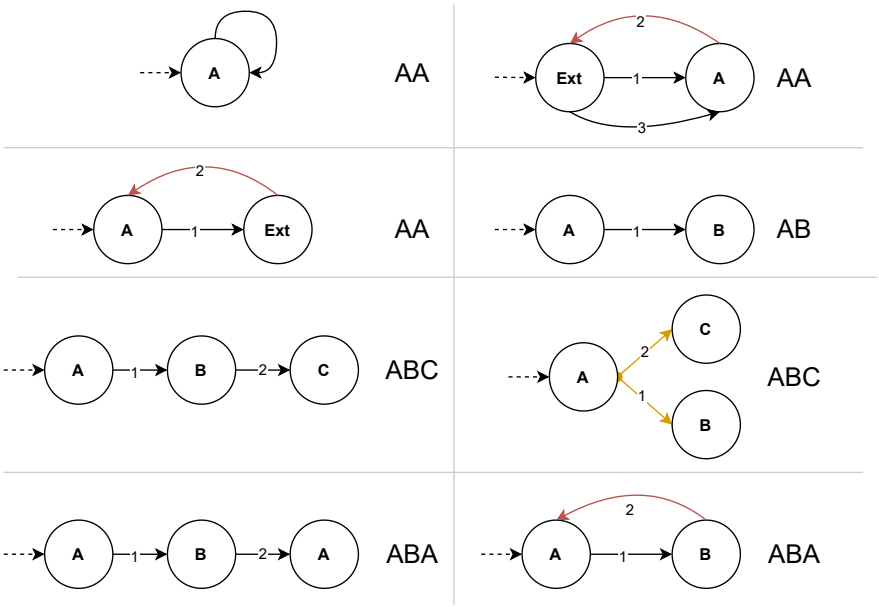


Figure 3.2: Examples of patterns in the logs and the multitude of client-side behaviors that can leave these digital traces. Black arrows represent click, red arrow are back button, yellow are multitab clicks

Dynamics of transitions. In order to understand the dynamics of these transitions, we investigate the inter-event time between the two pageloads in each bigram. The interval between two consecutive pageloads peaks at very short times, with a median of 74 seconds (63 and 93 seconds for mobile and desktop devices, respectively). However, as Fig. 3.1a shows, the distribution is long-tailed, with 22% of pairs separated by more than one hour.

Investigating the referrer of the second page of the bigrams reveals that readers frequently do not use internal links to transition between two articles, but external pages by leaving and re-entering Wikipedia. These external transitions are not rare: in 40.1% (or 35.2% when excluding AA patterns) of the bigrams with less than one hour between the two events, the second page was reached through external navigation. This observation is corroborated by Fig. 3.1b, which shows that for pairs with an inter-event time greater than 3 minutes and 48 seconds, transitions via internal links are even less common than transitions via external navigation. External transitions tend to be semantically coherent: considering all 1.4B AB-type bigrams where the second page is reached via search, in 15% of the cases, the first page explicitly contained the link. This proportion increases to 30% when considering pairs with an inter-event time of less than one hour and 60–70% considering less than 10 seconds (Fig. 3.1c).

The topical coherence of these transitions is also visible by observing the jump size in topic space. Fig. 3.1d plots the average topical distance (measured by the cosine of WikiPDA vectors, cf. Sec. 3.2) as a function of inter-event time, showing that external navigation recorded within a few minutes from the previous pageload shows topical distance comparable to internal navigation.

3.5 Session level

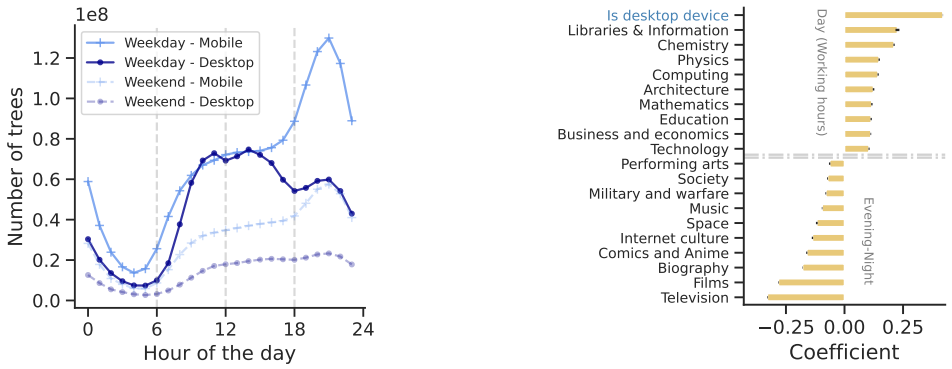
Using our insights about navigation at the unigram, bigram, and trigram levels, we can now characterize entire navigation sessions. We start by introducing two different approaches to conceptualize navigation sessions (Sec. 3.5.1) and discuss how each captures different aspects of readers' navigation. We then describe the properties of reader navigation by focusing on three aspects of the resulting sessions: contextual features defining when and how sessions start (Sec. 3.5.2), structural features of sessions (Sec. 3.5.3), and finally, the evolution of various article properties over the course of navigation sessions (Sec. 3.5.4).

3.5.1 Conceptualizing reader sessions

We introduce two notions of a user session, each capturing different aspects of navigation pathways (details below): (1) *navigation trees* connect pageloads hierarchically based on referrer information, whereas (2) *reading sequences* order pageloads linearly based on temporal information. From the original 6.52B pageloads, we obtain 3.7B navigation trees and 2.51B reading sequences.

Navigation trees [138] describe how readers traverse Wikipedia by following internal links. We generate a tree by connecting pageloads via the referrer contained in HTTP headers. Pages reached through internal transitions are added as children of the most recent load of the referrer, while pageloads with external or *Main_Page* referrers generate a new tree. If a page is loaded multiple times from the same referrer, the parent node retains only the first instance as a child. This method has the advantage of representing coherent sessions created through clicks on internal links and of capturing multitab behavior. The downside is the difficulty to capture content consumption over time for subsequent pages not reached through internal clicks, even if close in time (a common pattern, cf. Sec. 3.4).

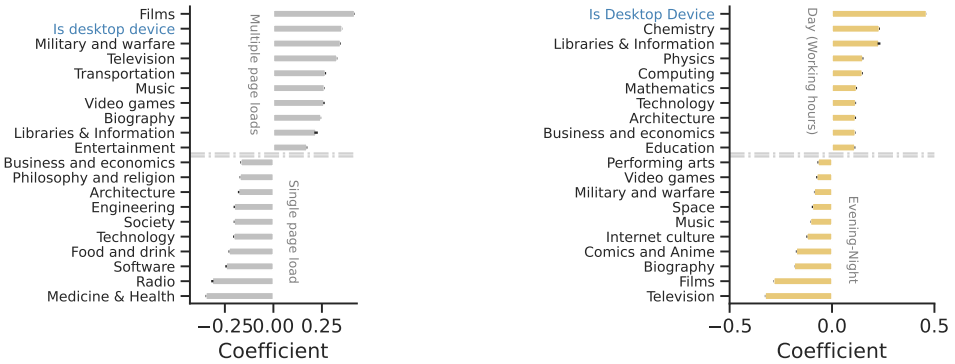
Reading sequences describe how readers consume content in temporal order. They are defined as linear sequences of all pageloads by the same user ordered by time. Sequences are split if the inter-event time between two consecutive pageloads separated by external navigation exceeds a threshold value of one hour, following recommendations from previous studies [68] and common practice [108, 171]. Within such sessions, we keep only the first pageload of each article, in order to only capture the first exposure of the respective content. This method generates topically less coherent sessions, capturing the temporal and linear



(a) Number of trees

(b) Features from regression

Figure 3.3: Statistics about time of day of sessions



(a) Multiple-pageloads

(b) Daytime

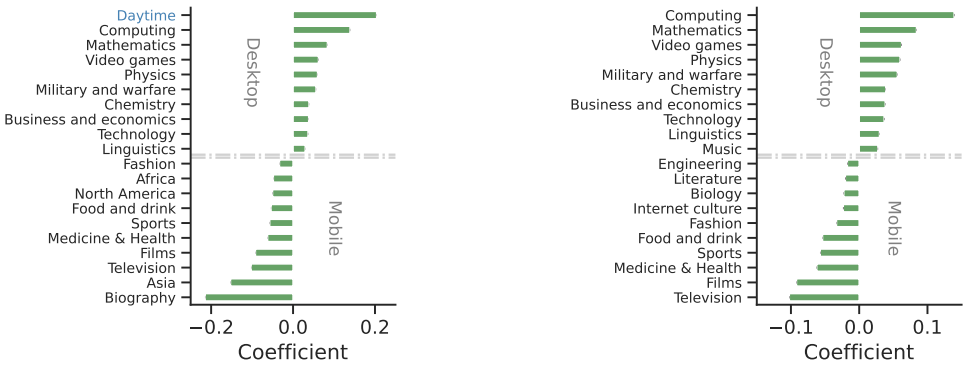
Figure 3.4: Feature contributions to the logistic model predicting if the reading sequence is composed by more than one pageload (Fig. 3.4a), and if the reading sequence started during daytime (Fig. 3.4b).

sequence of pageloads of a reader within a defined period of time, both via internal and external transitions (e.g., multiple external searches). This method has the disadvantage of being a simplification of how readers explore the link network, and a fixed threshold of one hour may not be ideal in every context.

3.5.2 Session context: time and device

We study the context of a session by focusing on the time of the first pageload and the device used to access Wikipedia. This section focuses on navigation trees, but reading sequences give qualitatively similar results (cf. Fig. 3.4b, Fig. 3.5a).

Time. To remove confounding via different timezones, we use the geolocation information to normalize the time of all pageloads to local time. The distribution of session starting times



(a) Navigation trees

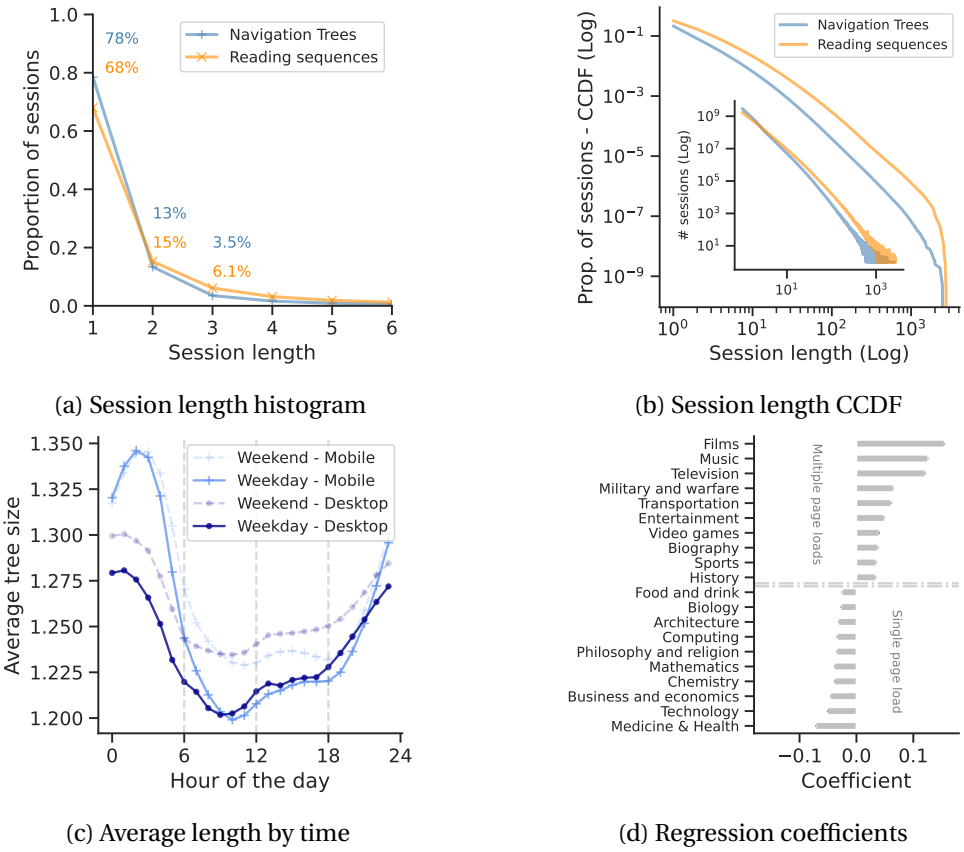
(b) Reading sequences

Figure 3.5: Feature contributions to a logistic model predicting if the session is started from a mobile or desktop device.

follows a regular circadian rhythm (Fig. 3.3a and Fig. 3.7a). Both access methods (desktop and mobile) show a similar pattern during the day, with a substantial increase of mobile sessions in the evening. Wikipedia has fewer sessions during weekends, but with similar temporal distributions as working days. The desktop distribution shows dents at 12:00 and 18:00, mirroring work rhythms with a lunch break around noon and the end of work in the evening (and possibly commuting).

In order to understand which features are associated with requests at different times of the day, we fitted a logistic regression to predict if a pageload was observed during the day or evening/night. We represent each pageload by its topic probabilities (obtained from ORES, cf. Sec. 3.2) and the type of device (desktop or mobile). Binarizing the target variable by representing daytime (9:00–18:00) as the positive class, we obtain an AUC of 0.586. Inspecting feature importance (Fig. 3.3b) shows that desktop devices and articles associated with STEM and education are associated with sessions starting during the day, whereas topics about entertainment are predictors of sessions starting during the evening or night.

Device. Fig. 3.3a indicates that people prefer different devices at different times of day. Next, we study whether specific topics are associated with device types by representing each pageload with the vector of topic probabilities (obtained from ORES) and a feature indicating if the page was loaded during the daytime. We again fit a logistic regression to predict the device used, with an AUC of 0.639. Inspecting feature importance shows that people tend to access STEM and business content from desktop devices, and biographies, entertainment, and medicine from mobile (Fig. 3.5).



(a) Session length histogram

(b) Session length CCDF

(c) Average length by time

(d) Regression coefficients

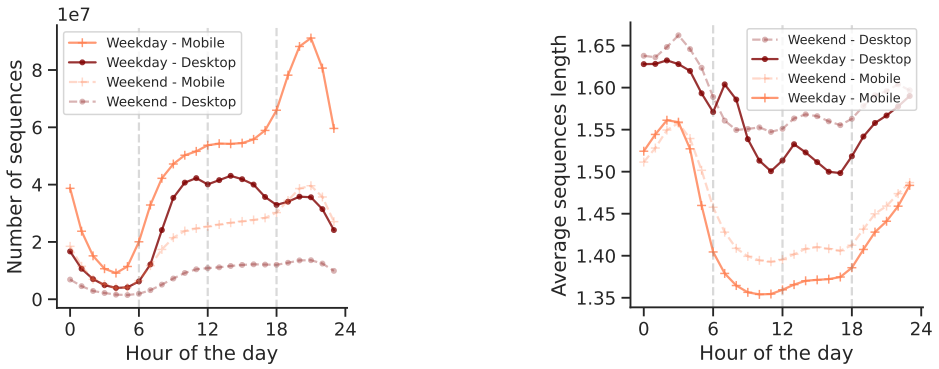
Figure 3.6: Session-length statistics.

3.5.3 Structure of sessions

Session length. We measure session length as the number of pageloads in the navigation tree or the reading sequence, respectively. Most sessions consist of a single pageload (Fig. 3.6a), but the length distribution also exposes a long tail (Fig. 3.6b). Therefore, we summarize session lengths via the geometric mean (arithmetic mean in parentheses). By construction, reading sequences tend to be longer because, unlike navigation trees, they merge both external and internal transitions.

In the case of reading sequences, the average session length shows differences with respect to the access method, with an average length of 1.41 (1.99) for mobile, and 1.54 (2.40) for desktop. This difference is less pronounced for navigation trees, where mobile sessions contain on average 1.23 (1.5) articles, vs. 1.24 (1.5) for desktop. The average session length varies during the day, with readers engaging in longer sessions during the evening and night, for both navigation trees and reading sequences (Fig. 3.6c and Fig. 3.7b).

To understand what properties are associated with short sessions consisting of a single pageload, we fitted a logistic regression to predict if the reader will continue after loading



(a) Number of reading sequences

(b) Session length by time

Figure 3.7: Total count (Fig. 3.7a), and average length (Fig. 3.7b) of the reading sequences started at different time of the day.

Tree size			
Top 10 (larger trees)		**Bottom 10 (smaller trees)**	
1.377	Films	*1.152*	Earth and environment
1.373	Entertainment	*1.148*	Food and drink
1.340	Television	*1.145*	Biology
1.327	Military and warfare	*1.138*	Technology
1.324	Music	*1.128*	Physics
1.295	Comics and Anime	*1.122*	Software
1.284	History	*1.114*	Medicine & Health
1.272	Biography	*1.112*	Computing
1.269	Sports	*1.104*	Mathematics
1.264	Transportation	*1.100*	Chemistry

Table 3.3: Top and bottom 10 topics with respect to (geometric) average tree size (geographical topics excluded).

the first page in a navigation tree (results are qualitatively identical for reading sequences; Fig. 3.4a), representing each first pageload with its topic probabilities (obtained from ORES), device type, and time of day, and obtaining a model with an AUC of 0.606. Inspecting the coefficients of the regression (Fig. 3.6d), we find that longer [shorter] sessions are associated with topical content around entertainment [STEM and medicine]. This observation is corroborated by the substantial difference in average navigation tree size across topics (Table 3.3).

Shape of navigation trees. In order to better understand how readers navigate the link network, we analyze the shape of navigation trees (in contrast, the shape of reading sequences is, by construction, always a linear chain). The three most common patterns (Fig. 3.8, left) are described as follows, in order of decreasing frequency: (1) a linear chain of pageloads; (2) fanning out from one page to several different pages, e.g., by opening multiple tabs or rolling back and selecting a different path; (3) a combination of the two (one-step chain followed by fanning out). These three patterns remain the most frequent for all tree sizes (Fig. 3.8, right).

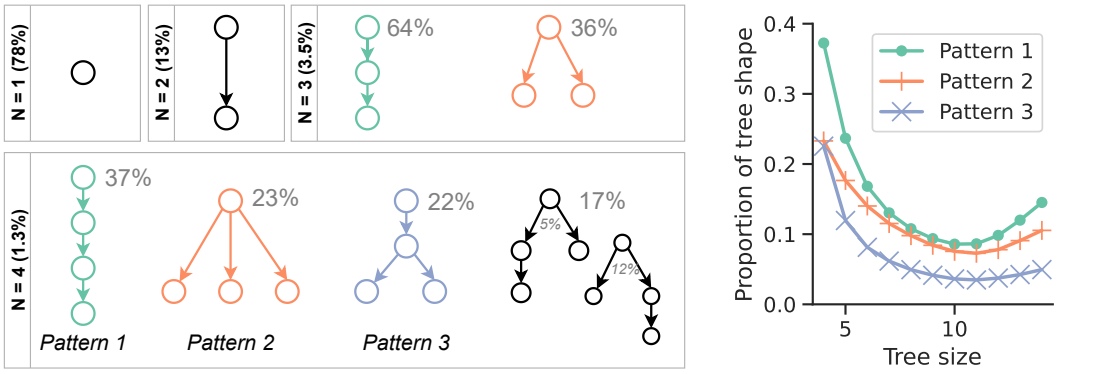


Figure 3.8: Shape of navigation trees. Frequency of patterns for trees size $N \leq 4$ (left panel). Dominance of top three patterns (see main text) for larger trees (right panel).

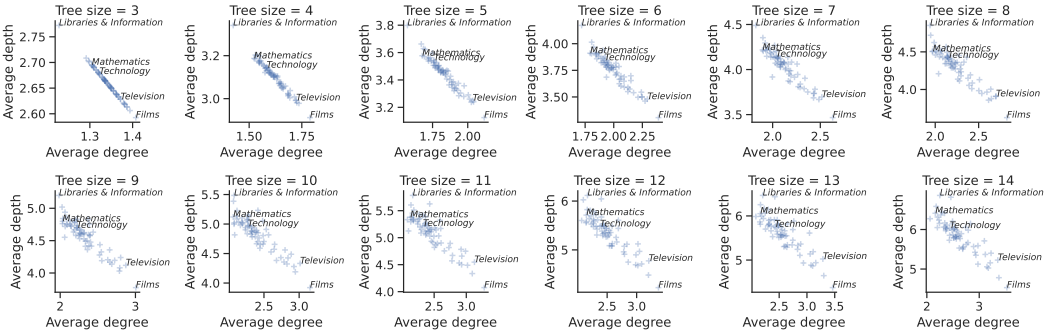


Figure 3.9: Relation between the average depth and average degree for navigation trees of different sizes.

We further characterize the different strategies associated with navigation trees in terms of tree depth (i.e., average length of paths from the root to the leaves) and breadth (i.e., average out-degree of non-leaves in the tree) for trees starting with different topics. Noting that the two metrics are almost perfectly anti-correlated and that the relative ordering of topics is stable across all tree sizes (Fig. 3.9), we define an aggregate tree-breadth ranking for each topic based on the average rank across tree sizes (Table 3.4). This shows that entertainment topics are associated with wider trees with higher branching, and STEM topics are characterized by deeper trees with a more chain-like structure.

3.5.4 Within-session article-property evolution

To shed light on navigation dynamics, we track the evolution of different article properties within sessions. Our evolution analysis revolves around three domains: topic space (distance from the first and previous articles), quality, and network centrality (out-degree and PageRank). Here, reading sequences are represented as defined above, whereas a navigation tree is represented by the linear path from the root to the temporally last leaf, from where the reader ceased to click further via internal links.

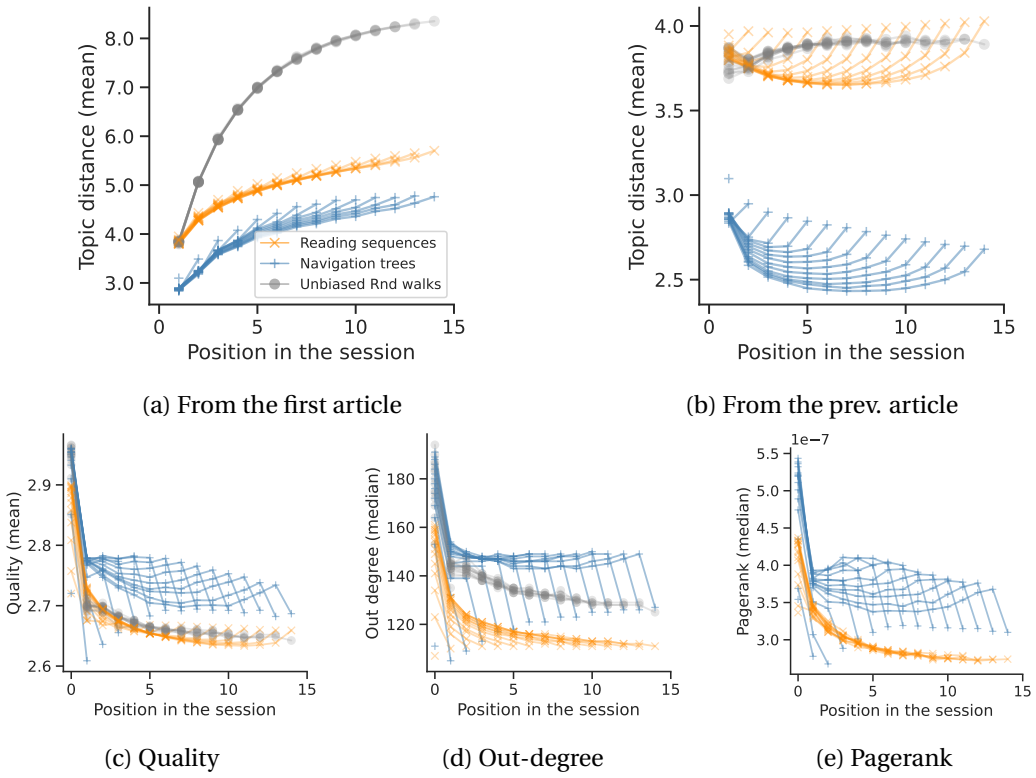
Top 10 (wider trees)			Bottom 10 (deeper trees)		
Rank (mean)	**Std**	**Starting Topic**	**Rank (mean)**	**Std**	**Starting Topic**
1.00	0.00	Films	27.42	2.72	Linguistics
2.50	0.87	Television	29.42	0.95	Earth and environment
3.58	0.76	Entertainment	29.50	1.19	Space
4.50	1.85	Comics and Anime	30.08	2.78	History
4.67	1.31	Education	31.92	1.11	Computing
6.58	1.98	Video games	32.92	1.55	Software
7.92	2.43	Literature	34.67	1.75	Chemistry
8.50	2.36	Fashion	34.75	1.30	Physics
8.83	1.07	Performing arts	35.50	1.26	Mathematics
10.42	2.29	Internet culture	35.67	1.65	Libraries & Information

Table 3.4: Rank with respect to average degree of navigation trees, by topic (geographical topics excluded). A separate rank was computed per tree size (3–15), and arithmetic means over tree sizes are reported, alongside standard deviations.

It is important to note that these two approaches can produce different sequences of pageloads: e.g., a pageload in position 1 of a navigation tree could be in position 4 of a reading sequence (as in Fig. 3.11). Also, the last pageload of each sequence can have different interpretations: for navigation trees, the reader stopped link-based navigation on that page, whereas for reading sequences, the reader left Wikipedia for at least one hour.

In order to better interpret our observations, we compare them with three null models corresponding to different random walkers. We randomly sample 120M paths from the navigation trees, and run (from the tree's starting article) (1) an *unbiased random walker* that selects the next steps with uniform probability from the available links and generates a sequence of the same length as the original path; (2) a *extrinsic-stop biased random walker* that selects the next step based on the pairwise transition probabilities obtained from the public clickstream and generates a sequence of the same length as the original path; (3) an *intrinsic-stop biased random walker* that selects the next step—or stops—based on the pairwise transition probabilities from the public clickstream [158]. We consider sessions up to length 15, stratifying by session length.

Topic space. We measure the topical distance between articles via the KL divergence of their respective WikiPDA topic distribution vectors (Sec. 3.2). For robustness, we tried different topic models (WikiPDA and ORES) and different distance metrics (KL divergence, Euclidean, cosine, and Wasserstein), obtaining qualitatively similar results. First, we study how readers diffuse in topic space starting from the first article, which plays a special role, as it represents the entry point to Wikipedia. On average, readers diffuse in topic space, moving further from the first article with every step (Fig. 3.10a). Reading sequences and navigation trees exhibit the same trend, with a shift due to the tendency of reading sequences to ignore external navigation. All the random walkers show similar increasing trajectories (Fig. 3.12a), diffusing faster than natural navigation when the random walker is unbiased, or biased but extrinsically stopped.



(a) From the first article

(b) From the prev. article

(c) Quality

(d) Out-degree

(e) Pagerank

Figure 3.10: Within-session evolution of 5 article properties. Each curve represents sessions of different lengths.

Second, we measure the semantic step size in topic space by tracking how the topical distance to the previous article evolves. Both navigation trees and reading sequences exhibit a U-shape, suggesting that readers tend to first reduce their semantic step size, before diverging and finally abandoning (Fig. 3.10b). The discrepancy between navigation trees and reading sequences is consistent with the previous observation on diffusion from the first article. Interestingly, this U-shape is similar to the trajectories generated by the intrinsic-stop biased random walker (Fig. 3.12b), as also reported in previous work [158]. In contrast, the other two random walk models show that by selecting a random link or stopping at predefined lengths, the average distance from the previous article tends to stabilize to an equilibrium value.

Quality. The evolution of article quality shows a sharp drop at the beginning, for both reading sequences and navigation trees (Fig. 3.10c). This behavior can be interpreted as a form of regression to the mean, since many sessions start from popular pages with high quality, which thus contribute more to the distribution. By moving one step in the link network, readers naturally reach a page that is, on average, of lower quality. The intuition is confirmed by the behavior of the unbiased random walker, which shows the same drop with the first step (Fig. 3.12c).

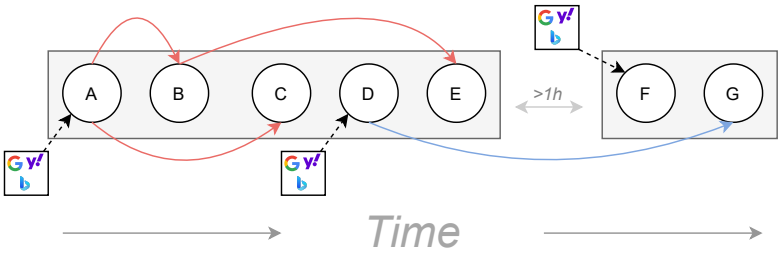


Figure 3.11: This set of log events yields three navigation trees, represented by arrows and composed of ACE, DG, and F. The reading sequences method creates two sessions represented as gray boxes: ABCDE and FG. Square boxes are clicks from external origins

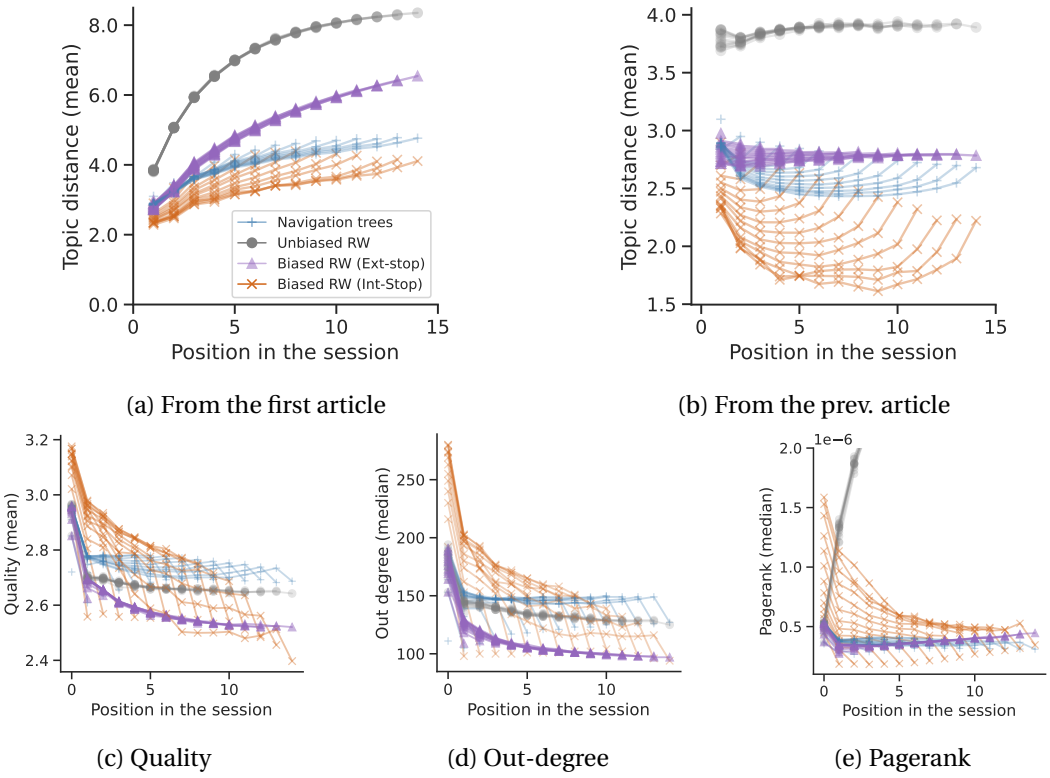
In contrast to reading sequences, navigation trees show a sharp drop in quality with the *last* pageload. This indicates that readers have a higher chance to stop Wikipedia-internal navigation when reaching a low-quality page, and as a result, continue navigating in a different branch of the tree or via an external transition.

Compared to the random walkers (Fig. 3.12c), readers tend to navigate across pages with less variance in quality. The random walkers' traces support the hypothesis that there are articles with a higher chance of terminating the navigation: while the unbiased and extrinsic-stop biased walkers show no termination pattern, the intrinsic-stop biased walker shows a final drop as in human navigation. The organic stopping of this random walker, mirroring readers' behavior more closely, increases the chances to abandon the navigation on pages of low quality that, according to the clickstream data, relay less traffic.

Network centrality. Finally, we are interested in how reader sessions evolve in the network with respect to different centrality measures. We start with out-degree (the number of outgoing links in article bodies). Similar to article quality, the out-degree shows a sharp drop with the first step (Fig. 3.10d) for navigation trees and reading sequences, likely caused by the presence of many sessions starting from pages with a high out-degree. We also find a sharp drop for the last pageload in the sequence of the navigation trees, suggesting that readers have a higher chance of stopping Wikipedia-internal navigation upon reaching a page with low out-degree.

In the case of the random walkers, we draw similar conclusions as for article quality. Whereas unbiased random walks and extrinsic-stop biased random walks show a decrease and stabilization of out-degree, the intrinsic-stop random walker, as humans, terminates on pages of lower degree (Fig. 3.12d). Compared to random walkers, human navigation is more stable: after the initial drop, they have a higher chance to stay on pages with around 150 links.

Finally, we characterize how the PageRank of visited articles changes during sessions. We observe that the PageRank mirrors the evolution of quality and out-degree with regard to the initial drop (Fig. 3.10e). Readers tend to enter more frequently on popular pages with high centrality and naturally move to a less central node in one step. Also for this case, a drop is visible in the last step of the navigation trees, indicating that, when the readers reach an article



(a) From the first article

(b) From the prev. article

(c) Quality

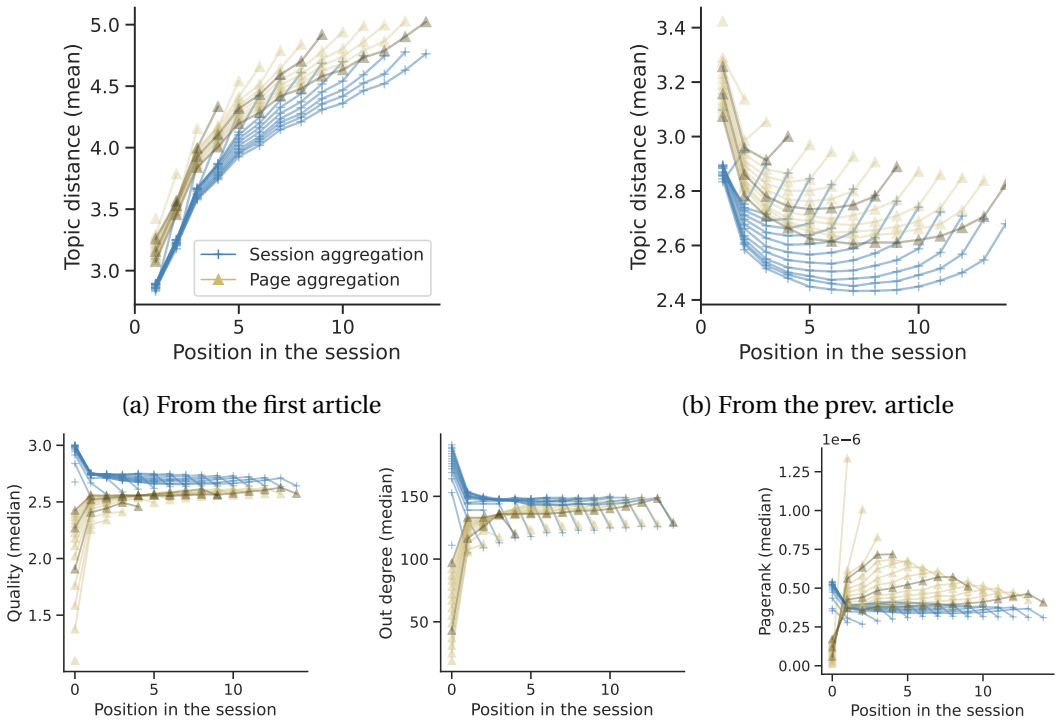
(d) Out-degree

(e) Pagerank

Figure 3.12: Properties evolution of the trajectories generated by the three random walk models compared with the natural navigation based on navigation trees. Each curve represents sessions of different lengths.

leading to the network periphery, they have higher chances to stop the Wikipedia-internal navigation. The random walkers (Fig. 3.12e) show that unbiased walks naturally converge in a few steps to the most central pages with very high PageRank. The extrinsic-stop biased walker, on the contrary, after an initial drop, tends to move to central nodes at a much lower speed. Finally, the intrinsic-stop biased walker, again, shows a final drop from a stable value before abandoning the navigation, similar to human readers.

Aggregation by page. The quantities in Fig. 3.10 correspond to a micro-average over all sessions, where the average behavior can be dominated by sessions starting from the most popular pages. Therefore, we also calculate a macro-average by aggregating on a starting-page level to make each first article contribute equally. The diffusion in topic space is qualitatively similar in both aggregation methods (Fig. 3.13a and Fig. 3.13b). In contrast, for quality, out-degree, and PageRank, the overall trend is inverted, i.e., instead of a sharp drop, we observe a sharp increase in these metrics after the first step (Fig. 3.13c, Fig. 3.13d, and Fig. 3.13e). This discrepancy could be caused by the presence of many low-quality [140] and low-degree articles, such that readers at the first step tend to move to better articles in search of information. Interestingly, the drop towards the last pageload in a session appears across both aggregation methods.



(a) From the first article

(b) From the prev. article

(c) Quality

(d) Out-degree

(e) Pagerank

Figure 3.13: Comparison of the evolution of five different properties when aggregating navigation trees by sequence and by page. Gray trajectories added from readability. Each curve represents sessions of different lengths.

3.6 Discussion

We have provided a systematic characterization of the navigation pathways of readers in Wikipedia through a large-scale study of the site's server logs. Our results provide a rigorous framework to capture and describe information seeking on one of the largest platforms for open knowledge.

3.6.1 Summary of findings

Starting from the raw logs, we developed a systematic pre-processing pipeline which allowed us to quantify how readers reach, and transition between, pages. First, the most common way to reach a page is through an external search engine, followed in frequency by internal navigation from other Wikipedia articles; other sources, such as external websites (mostly social media sites) and other Wikipedia content (such as categories or special pages), are much less frequent, but still substantial in absolute numbers. Second, readers frequently transition between pages via external search engines instead of using a direct Wikipedia link. These external transitions are characterized by larger topic jumps and larger inter-event times between pageloads; they must, however, still be considered semantically meaningful, for, in

many cases, a link for internal navigation would be available. Third, by analyzing sequential patterns, we find that consecutive reloads and revisits to a previous article are rather common (10% or more each).

We continued by characterizing how readers combine the above patterns into extended navigation sequences. First, we introduced two approaches to capture paths of readers: *navigation trees* based only on internal navigation, and *reading sequences* based on the time-ordered pageloads, including internal and external transitions. Second, we described how sessions are affected by their context in terms of device type and time of day. We find that topics related to STEM [entertainment] are more associated with working [evening and night] hours. Third, we measured the size and structure of sessions. While most sessions consist of a single pageload (68–78% depending on the aggregation method), the size distribution shows a long tail with tens of millions of sessions consisting of 10 or more pageloads. The topic not only affects the size but also the shape of trees: while sessions starting from articles on entertainment generally consist of more pageloads, such trees are also broader (higher branching factor) than sessions starting, e.g., from STEM topics, which are smaller and deeper. Fourth, we investigated the within-session evolution of article properties. In topic space, longer sessions diffuse far from the origin with a characteristic U-shape pattern suggesting that readers reduce their steps first, before diverging and finally abandoning the session. The first and last pageload of a session show special behavior regarding the evolution of article quality and network properties. Popular pages are naturally more common as first articles, thus engendering a form of regression to the mean with the second step. An inverted effect appears when sessions are aggregated at the page level, so every starting article is represented equally. The articles at the end of the navigation are typically lower-quality pages, suggesting that readers stop following the internal navigation when they reach these pages acting as network sinks.

3.6.2 Implications

Complexity of data. Our results show that the dataset of navigation paths of readers in Wikipedia extracted from server logs constitutes a non-trivial dataset requiring extreme care in order to avoid drawing spurious conclusions. First, in contrast to existing pre-processing pipelines for sequence analysis (e.g., tokenization steps in NLP), we still lack an understanding of best practices for navigation paths, and as a result had to develop a set of domain-specific heuristics. Second, operationalizing navigation paths makes strong assumptions: while navigation trees from pure internal navigation are more topically coherent with a more complex structure, reading sequences from temporally ordering all of the user's pageloads are less coherent but provide a linear sequence that is not broken by external searching (which is common). The latter typically introduces an additional cutoff for sessions if consecutive pageloads are separated by more than one hour [68]; however, our analysis suggests other potential data-informed choices such as the time separation of internal and external transitions at approximately 4 minutes. Naturally, the suitable choice depends on the question of interest. Third, our analysis

shows that the data can exhibit Simpson's paradoxes; e.g., the inversion of the within-session evolution of page properties such as PageRank depends on the aggregation level. Fourth, the prevalence of trivial patterns (e.g., reload or revisit) points to potential caveats when applying prediction models to session-based recommendation [198].

Diversity. There is extraordinary diversity in the ways readers browse Wikipedia, modulated by topic, device, time of day, etc. This reflects the diversity found in previous studies on the different motivations and information needs of readers across the globe [86, 108, 171]. This heterogeneity indicates caution against simplistic models aiming to capture a single average behavior.

Online ecosystem. The usage of Wikipedia is embedded in a larger online ecosystem. Multiple studies have shown the importance of Wikipedia to search engines [120, 191], as a gateway to the Web [142, 143], and as a main educational resource for online learning more generally [96]. Our results show that this interplay between external and internal (with respect to Wikipedia) also plays a crucial part on an intra-session level when navigating for knowledge.

Navigation in the wild. The navigation of readers on Wikipedia differs from targeted navigation in lab-based settings [73, 199, 200]. We do not observe typical strategies characterized by, e.g., navigation via hubs (an initial increase followed by a drop in out-degree) or gradually decreasing steps in semantic space towards the target. Instead, we find a range of other patterns, such as a U-shape for the step-size in semantic space and an immediate sharp drop followed by largely constant centrality measures (out-degree, PageRank). This highlights conceptual limitations of targeted-navigation experiments to generalize their findings to how humans seek knowledge more generally.

Furthermore, our results provide a more nuanced picture of the conclusions derived from publicly available data, most notably the Wikipedia clickstream [210], which provides aggregate data on the number of times a link was clicked. For example, we can observe that the overall tendency to navigate towards peripheral nodes [46] is mainly driven by the first step after reaching Wikipedia, with subsequent steps showing much smaller differences in centrality measures (with the exception of the last step, see below). One possible interpretation is a regression-to-the-mean effect as popular pages (the starting points of navigation) are generally skewed towards higher centrality and quality.

Content affects navigation. Our results contribute to describing the relation between contents and navigation, expanding the prior understanding of how readership and popularity are influenced by visual position [46] or quality [218]. Our results go beyond the population level, suggesting that low-quality pages lead readers to stop navigating along a specific branch in the navigation tree (and continuing along a different branch or stopping altogether). This is specifically important in the context of knowledge gaps in Wikipedia [154], in order to address the uneven representation of, e.g., articles on women, where a better understanding

of the interaction between content, readers, and editors [52, 167] is crucial to allow for more informed decision-making in designing interventions.

3.6.3 Limitations and future work

Limitations. In terms of limitations, we capture navigation paths only via events in the server logs. Moving forward, how people engage with content could be more accurately observed via client-side instrumentation. The aggregation based on IP and user-agent information also has limitations; e.g., we had to discard the sessions of large organizations with shared IP addresses. Finally, we only focused on a single language, English. While this already revealed a rich spectrum of phenomena, additional variation can be expected from a comparison across languages [108].

Future work. To overcome these limitations, future work should capture the variation in navigation across Wikipedia's over 300 languages. Moreover, in order to better serve the different information needs of readers, a better understanding is needed of how patterns in navigation correspond to underlying motivations [171] or other traits such as curiosity [112]. Finally, in order to capture knowledge seeking more generally, researchers should capture navigation beyond individual platforms to take into account the interdependence of Wikipedia with the rest of the Web.

Conclusion. This chapter offers an overview of how readers reach and consume content on Wikipedia. We characterized the first two stages of the readers' navigation. We presented a pragmatic framework to analyze the navigation patterns and described the properties of the sessions generated. The next chapters will focus on the next stage of the navigation: how readers leave the platform.

4 How Readers Engage with Citations on Wikipedia

4.1 Introduction

Wikipedia is the largest encyclopedia ever built, established through the collaborative effort of a large editor base, self-governed through agreed policies and guidelines [19, 51]. Thanks to the tenacious work of the editor community, Wikipedia's content is generally up to date and of high quality [89, 147], and is relied upon as a source of neutral, unbiased information [121].

Wikipedia's inline references, or citations,[I] are a key mechanism for monitoring and maintaining its high quality [49, 153]. Wikipedia's core content policies require that "people using the encyclopedia can check that the information comes from a reliable source",[II] and citations are the main way to connect a statement to its sources. A clearly distinctive feature of Wikipedia is the fact that many citations are actionable: they are often equipped with hyperlinks to the cited material available on the Web.

As a result, Wikipedia's role on the Web has been defined as the "gateway through which millions of people now seek access to knowledge" [37, 142] and the "bridge to the next layer of academic resources" [62]. A sizable portion of citations on Wikipedia refer to scientific literature [133]. Consequently, Wikipedia is a fundamental gateway to scientific results and enables the public understanding of science [109, 116, 132, 165, 173, 186, 188, 215]. The chance of a scientific reference being cited on Wikipedia varies with the impact factor of the publication venue and its open-access availability [184]. Being cited on Wikipedia can thus be considered an indicator of impact [95]. Nevertheless, a question remains open: to which extent do Wikipedia readers actually cross the bridge and access the broader knowledge referenced in the encyclopedia?

Given the collaborative and open nature of Wikipedia, being able to quantify readers' engagement with the content and its supporting sources is of crucial importance for the constant

[I] We use the terms "reference" and "citation" largely interchangeably.

[II] https://en.wikipedia.org/wiki/Wikipedia:Verifiability, https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources

betterment of the encyclopedia and its role in fostering a self-critical society. By understanding readers' interactions with citations, we can better assess the role of Wikipedia editors and policies in maintaining a high quality of information, measure public demand for secondary sources, and provide insights and potential recommendations to increase the public's interest in references.

The study presented in this chapter takes a step in this direction, by addressing, for the first time, the problem of quantifying and studying Wikipedia readers' engagement with citations. More specifically, we ask the following research questions,

RQ1 To what extent do users engage with citations when reading Wikipedia? (Sec. 4.3)

RQ2 What features of a page predict whether a reader will interact with a citation on the page? (Sec. 4.4)

RQ3 What features of a citation predict whether a reader will interact with it? (Sec. 4.5)

In order to answer these questions, we collect a large dataset comprising all citation-related events (96M) on the English Wikipedia for two months (October 2018, April 2019), including reference clicks, reference hovers, and downwards and upwards footnote click, as visualized in Fig. 4.1. By analyzing this dataset,[III] we make the following main contributions:

- We quantify users' engagement with citations and find that it is a relatively rare event (RQ1, Sec. 4.3): 93% of the links in citations are never clicked over a one-month period, and the fraction of page views that involve a click on a citation link is 0.29%.
- We gain insights into factors associated with seeking additional information via citation interactions, both at the page level (RQ2, Sec. 4.4) and at the link level (RQ3, Sec. 4.5). Through matched observational studies, we show that articles that are of higher quality, and thus also longer and more popular, are associated with a lower propensity of users to interact with citations. Using a logistic regression model trained on linguistic features, we show that more frequently clicked citation links tend to relate to social or life events.

We thus conclude that readers are more likely to use Wikipedia as a *gateway* on topics where Wikipedia is still wanting and where articles are of low quality and not sufficiently informative; and that Wikipedia tends to be the *final destination* in the large majority of cases where the information it contains is of sufficiently high quality.

Our work provides the first study aimed at understanding if and how users engage with citations on Wikipedia, thus paving the way for a broader and deeper understanding of Wikipedia's role in the global information ecosystem.

[III] Notebooks with code at https://github.com/epfl-dlab/wikipedia-citation-engagement

4.2 Data

To study readers' engagement with citations, we collected data capturing where readers navigate and how they interact with citations in English Wikipedia.

4.2.1 Background: Citations in Wikipedia

Articles in Wikipedia are written by editors in wikicode, a markup language that is then translated to HTML by MediaWiki, the software that powers the website. There are different ways to add citations to sources in the text, summarized below. In all cases, the full reference descriptions are rendered as footnotes at the bottom of the page (in a dedicated section called *References*) with an automatically assigned footnote number that is added as a link anchor (e.g., "[1]") in the text of the article wherever the reference is cited (Fig. 4.1). Most references in the *References* section consist of text including the title of the source, the authors' names, the year of publication, and the source's publisher. For 80% of Wikipedia references, the source title is actionable via a clickable link to the source. Also, when reading a page, hovering over a reference's footnote number with the mouse cursor will display a *reference tooltip*,[IV] a pop-up containing the reference text and a clickable link (when present), e.g.,

> Daniel Nasaw (July 24, 2012). "Meet the 'bots' that edit Wikipedia". BBC News.

When readers click on the reference's footnote number, they are sent to the reference description at the page bottom, from where they can jump back to the locations where the reference is cited by clicking on a small icon (e.g., ^).

The most common method to add a reference to an article, also recommended by the Wikipedia guidelines, is via an inline citation using a `<ref/>` tag directly in the context where the reference is first cited. In the tag, the editors can specify the reference details (text and links) by using a predefined template or plain wikicode. In addition to this standard method, some references are added automatically by templates included in the page, such us the geolocations present in the infobox. It is worth noting that a reference can be cited multiple times by assigning it a name and appending the tag to every sentence that should link to it. Given the numerous ways to use the `<ref/>` tag, and in order to have an accurate view of the article, we parsed pages from wikicode to HTML and extracted the information from the HTML code.

4.2.2 Logging citation and page load events

We make use of Wikimedia's *EventLogging* tool,[V] an extension of the MediaWiki software that performs client-side logging of specific types of events. We detect 5 main types of citation-

[IV] https://www.mediawiki.org/wiki/Reference_Tooltips
[V] https://www.mediawiki.org/wiki/Extension:EventLogging/Guide

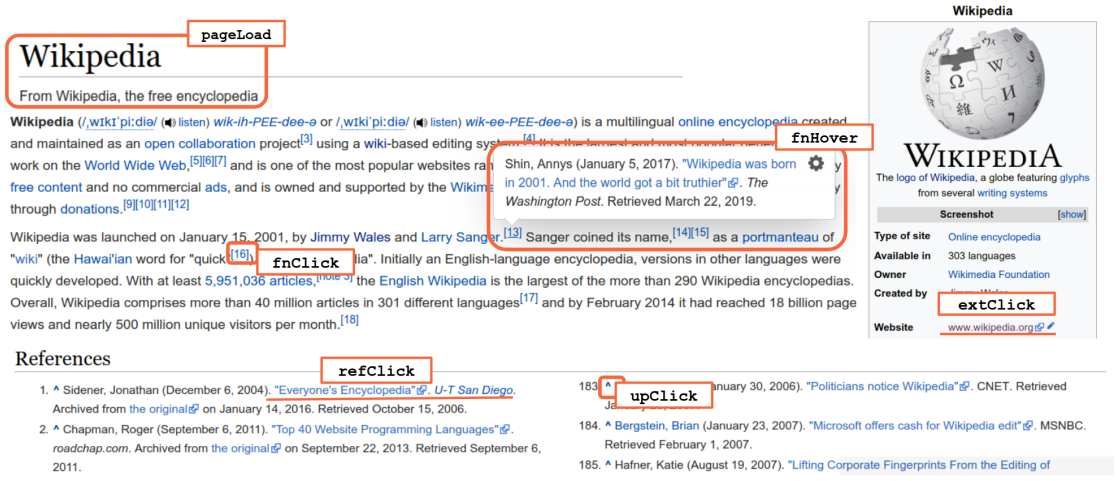


Figure 4.1: Examples of the 6 types of interactions with pages and citations that we record on English Wikipedia using Wikimedia's EventLogging tool.

related events and 1 page load event. In terms of citations, we capture the mouse events that involve any kind of reader interaction with the references (see Fig. 4.1 for a visual explanation):

refClick: a click on a hyperlink in an article's reference section.

extClick: a click on an external link outside the reference section.

fnHover: a hover over a footnote number in the text, logged when the reference tooltip is visible for more than 1 second.

fnClick: a click on a footnote number, which takes the user to the reference section at the bottom of the page.

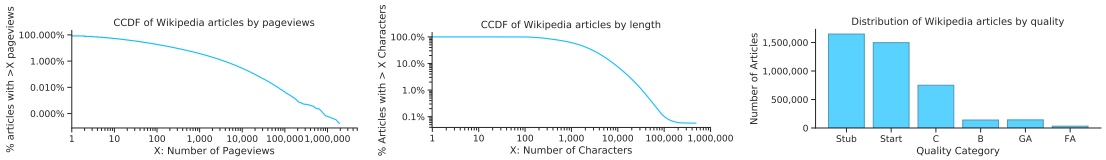
upClick: the inverse of *fnClick:* a click on a reference's up arrow icon that takes the reader back to the part of text where the reference is cited.

pageLoad: in addition to the above citation-related events, this event is triggered whenever a Wikipedia article is loaded.

The EventLogging platform manages a so-called *session token,* a cookie-based identifier that allows us to group events that happened within the same browser tab. We henceforth refer to event sequences that occur with the same session token as **sessions**. Please note that the definition of sessions in this study is not the same used in Chapter 3, but it refers to the client activities recorded in a single browser tab.

We collected 4 contiguous weeks[VI] of Wikipedia mobile and desktop traffic data of citation-related events. We repeated the 4-week data collection over two periods: from September

[VI] We collected exactly 4 weeks to reduce potential seasonal effects due to uneven day-of-the-week frequencies.



(a) (b) (c)

Figure 4.2: Distribution of Wikipedia articles by (a) popularity (number of pageviews), (b) page length (number of characters in wikicode), and (c) quality (increasing from left to right; "GA" for "Good Article", "FA" for "Featured Article") (Sec. 4.2.5).

26 to October 25, 2018, and from March 24 to April 21, 2019. In both cases, we collected all citation-related events (*extClick*, *refClick*, *fnHover*, *fnClick*, *upClick*) and (due to computational infrastructure constraints) sampled *pageLoad* events at the session level at a rate of 33%.

To ensure that the logs reflect reader, rather than editor, behavior, we exclusively retained data from users who in the 4 weeks of data collection acted only as anonymous readers, discarding all events generated by Wikipedia editors (logged-in users or users with anonymous edits) and by bots (which can be filtered out using a detector provided by the EventLogging tool).

Throughout the chapter, we will mostly focus on the data from the second data collection period (April 2019) and only use the October 2018 data for a longitudinal study measuring the impact of article quality on readers' engagement with citations.

4.2.3 Definition of engagement metrics

Two key metrics in our analysis will be the citation click-through rate (CTR) and the footnote hover rate.

For each page p and each session s, let $C(p,s)$ be the indicator function that is 1 if at least one reference was clicked on page p during session s by the respective user (*refClick* event), and 0 otherwise. Analogously, let $H(p,s)$ indicate if the user hovered over at least one footnote (*fnHover* event). Furthermore, let $N(p)$ be the number of sessions during which p was loaded (*pageLoad* event)

Global click-through rate. The global CTR measures overall reader engagement via reference clicks across Wikipedia. It is defined as the fraction of page views on which at least one reference click occurred (treating all views of the same page in the same session as one single event):

$$\text{gCTR} = \frac{\sum_p \sum_s C(p,s)}{\sum_p N(p)}, \tag{4.1}$$

where p ranges over the set of pages that contain at least one reference with a hyperlink.

Page-specific click-through rate. The page-specific CTR for page p is defined as the probability of observing at least one click on a reference in p during a session in which p was viewed:

$$\text{pCTR}(p) = \frac{\sum_s C(p,s)}{N(p)}. \tag{4.2}$$

Finally, we denote the average page-specific CTR over a set P of pages by

$$\text{pCTR}(P) = \frac{1}{|P|} \sum_{p \in P} \text{pCTR}(p). \tag{4.3}$$

Note that pCTR(P) corresponds to a macro average where every page gets the same weight, whereas gCTR corresponds to a micro average where pages are weighted in proportion to the number of sessions in which they were viewed.

Footnote hover rates. In analogy to the above definitions, but when replacing the click indicator $C(p,s)$ with the hover indicator $H(p,s)$, we obtain the global and page-specific footnote hover rates:

$$\text{gHR} = \frac{\sum_p \sum_s H(p,s)}{\sum_p N(p)}, \qquad \text{pHR}(p) = \frac{\sum_s H(p,s)}{N(p)}. \tag{4.4}$$

4.2.4 Capturing event context

Each event is characterized by a set of features that capture information about three aspects of the event: the session in which the event happened, the page, and the reference.

Session: We collect the unique *session token* (cf. Sec. 4.2.2) that identifies the browser tab in which the event occurred.

Pages: At the article level, we store *title, page id, text length of wikicode in characters, number of references,* and *popularity* (number of *pageLoad* events during the data collection period). We also use the ORES *drafttopic* classifier [9] to label each Wikipedia article with a vector of *topics,* whose elements reflect the probability of the page to belong to one the 44 topics from the highest level of the WikiProjects taxonomy.[VII] We further use the ORES *articlequality* model [66] to label articles with a *quality* level, which can take the following values (from low to high quality): “Stub”, “Start”, “C-class”, “B-class”, “Good Article”, “Featured Article”.

References: For each reference clicked or hovered, we record its *URL,* the *text in the reference,* the *text of the sentence* in which the reference is cited, and the *relative position* (character

[VII] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Directory

offset from the start in plain text, divided by page length) in the page where the reference is cited. Since we associate references to their contexts, references to the same source appearing on different pages are treated as distinct.

Wikipedia is dynamic by nature: articles are continuously updated, and their changes are tracked through revisions. To account for the evolution of articles over the 4 weeks of data collection, we aggregate individual revision-level metrics at the article level. To compute article-specific characteristics such as article length or number of references, we calculate their average over all revisions from the logging period. To quantify the amount of reader engagement with a given article (e.g., page loads, reference clicks), we sum all events recorded at each revision of the article.

4.2.5 General statistics of English Wikipedia

By the end of the data collection, English Wikipedia contained 5.8M articles, 5.4M (95%) of which were loaded at least once in our data sample, in a total of 7.4M revisions. Out of these articles, 3.9M (73%) contain at least one citation, linking to a total of 24M distinct URLs.

Over the 4 weeks of data collection, we collected (at a 33% sampling rate) 1.5B *pageLoad* events (62% from the mobile site and the rest from the desktop site). In Fig. 4.2a we report the (complementary cumulative) popularity distribution for the Wikipedia pages that were viewed at least once during the data collection period. The distribution is heavily skewed, with approximately 83% of the articles loaded fewer than 100 times in the 33% random sample (cf. Sec. 4.2.2), or fewer than 300 times when extrapolating to all data.

We observe a similar uneven distribution of page length (Fig. 4.2b), with the majority of articles being very short.

Fig. 4.2c shows that the distribution of article quality levels is also heavily skewed toward low quality levels: most articles are identified as “Stub” or “Start”, and fewer than 300K articles are marked as “Good” or “Featured” articles.

4.3 Prevalence of citation interactions

After these preliminaries, we are now ready to address our first research question, which asks to what extent Wikipedia readers engage with citations.

4.3.1 Distribution of interaction types

We start by analyzing the relative frequency of the different citation events, as defined in Sec. 4.2.2. Over the month of data collection, we captured a total of 96M citation events. Fig. 4.3 shows how these events distribute over the 5 event types, broken down by device type

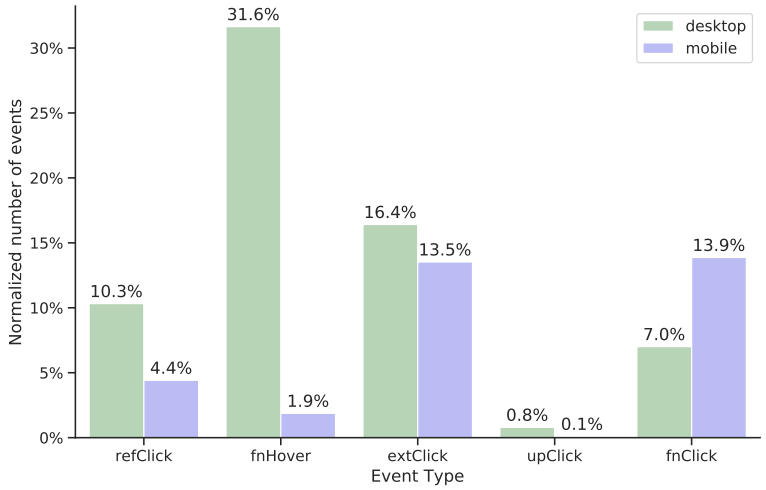


Figure 4.3: Relative frequency of citation-related events (Sec. 4.2.2), split into desktop (green, left bars) and mobile (blue, right bars) in April 2019 (Sec. 4.3.1).

(mobile vs. desktop). We observe that most interactions with citations happen on desktop rather than mobile devices, despite the fact that the majority of page loads (62%) are made from mobile.

The interactions also distribute differently across types for mobile vs. desktop. The by far prevailing event on desktop is hovering over a footnote (*fnHover*) in order to display the reference text. Hovering requires a mouse, which is not available on most mobile devices, which in turn explains the low incidence of *fnHover* on mobile. In order to reveal the reference text behind a footnote, mobile users instead need to click on the footnote, which presumably explains why *fnClick* is the most common event on mobile.

Clicking external links outside of the *References* section at the bottom of the page (*extClick*) is the second most common event on both desktop and mobile, followed by clicks on citations from the *References* section (*refClick*). Finally, the *upClick* action, which lets users jump back from the *References* section to the locations where the citation is used in the main text, is almost never used.

4.3.2 Citation click-through rates

We now focus on the two prevalent interactions with citations, hovering over footnotes (*fnHover*) and leaving Wikipedia by clicking on citation links (*refClick*). (We do not dwell on *extClick* events, as they do not concern citations but other external links; cf. Sec. 4.2.2.)

First, we observe that, out of the 24M distinct URLs that are cited across all articles in English Wikipedia, 93% of the URLs are never clicked during our month of data collection.

Next, we note that the global click-through rate (CTR) across all pages with at least one citation (gCTR, Eq. 4.1) is 0.29%; i.e., clicks on references happen on fewer than 1 in 300 page loads. Breaking the analysis up by device type, we observe again substantial differences between desktop and mobile: on desktop the global CTR is 0.56%, over 4 times as high as on mobile, where it is only 0.13%.

The average page-specific CTR (pCTR, Eq. 4.3) is higher, at 1.1% for desktop and 0.52% for mobile. This is due to the fact that there are many rarely viewed pages (cf. Fig. 4.2a) with a noisy, high CTR. After excluding pages with fewer than 100 page views, the pCTR is 0.67% on desktop, and 0.21% on mobile – 0.49% considering both devices.

Engagement via footnote hovering is slightly higher, at a global footnote hover rate (gHR, Eq. 4.4) of 1.4%. The average page-specific footnote hover rate (pHR, Eq. 4.4) is 0.68% when including all pages with at least one clickable reference, and 1.1% when excluding pages with fewer than 100 page views.[VIII]

Given these numbers, we conclude that readers' engagement with citations is overall low.

4.3.3 Positional bias

Previous work has shown that users are more likely to click Wikipedia-internal links that appear at the top of a page [138]. To verify whether this also holds true for references, we sample one random page load with citation interactions per session and randomly sample one clicked and one unclicked reference for this page load. We then compute each reference's relative position in the page as the offset from the top of the page divided by the page length (in characters). Fig. 4.4, which shows the distribution of the relative position for clicked and unclicked references, reveals that users are more likely to click on references toward the top and (less extremely so) the bottom of the page.

4.3.4 Top clicked domains

Next, we investigate what are the most frequent domains at which users arrive upon clicking a citation.

Initially, we found that the most frequently clicked domain is `archive.org` (Internet Archive), with 882K *refClick* events. Such URLs are usually snapshots of old Web pages archived by the Internet Archive's Wayback Machine. To handle such cases, we extract the original source domains from wrapping `archive.org` URLs.

In Fig. 4.6 we report the top 15 domains by number of *refClick* events. The most clicked domain is `google.com`. Drilling deeper, we checked the main subdomains contributing to this

[VIII] As mentioned in Sec. 4.3.1, hovering is not available on most mobile devices, so the hovering numbers pertain to desktop devices only.

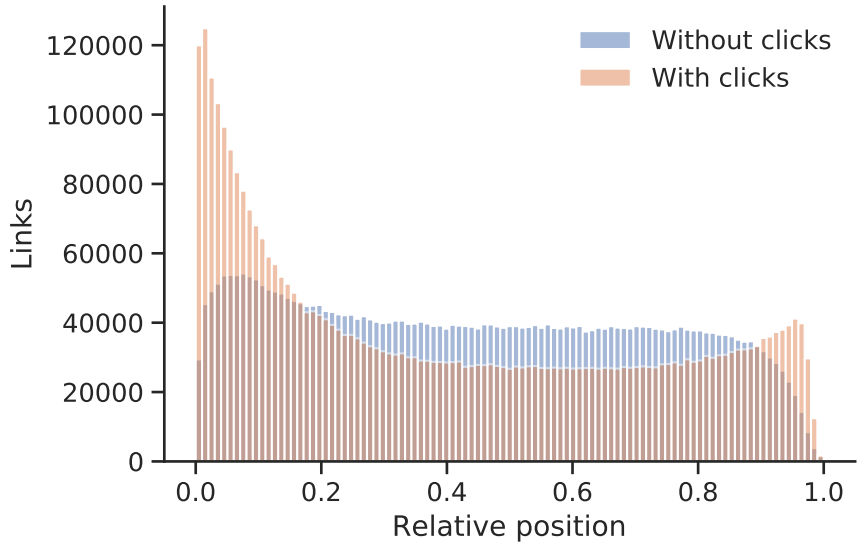


Figure 4.4: Relative position in page of clicked vs. unclicked references, for references with hyperlinks (Sec. 4.3.3).

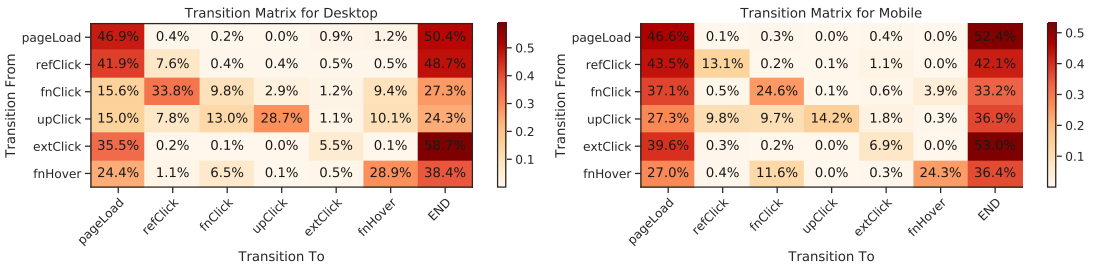
statistic, finding that a significant proportion of clicks goes to `books.google.com`, which is providing partial access to printed sources. The second most clicked domain is `doi.org`, the domain for all scholarly articles, reports, and datasets recorded with a Digital Object Identifier (DOI), followed by (mostly liberal) newspapers (*The New York Times*, *The Guardian*, etc.) and broadcasting channels (BBC).

4.3.5 Markovian analysis of citation interactions

Whereas the above analyses involved individual events, we now begin to look at *sessions:* sequences of events that occurred in the same browser tab (as indicated by the session token; Sec. 4.2.2). Every session starts with a *pageLoad* event, and we append a special *END* event after the last actual event in each session. Note that we cannot directly compare this event with the end of navigation described in Chapter 3 because, in this case, a session is defined in the scope of a browser tab.

By counting event transitions within sessions, we construct the first-order Markov chain that specifies the probability $P(j|i)$ of observing event j right after event i, where i and j can take values from the event set introduced in Sec. 4.2.2 (*pageLoad*, *refClick*, *extClick*, *fnClick*, *upClick*, *fnHover*) plus the special *END* event.

The transition probabilities are reported in Fig. 4.5. We observe that most reading sessions are made up of page views only: on both desktop and mobile, after loading a page, readers tend to end the session (with a probability of around 50%) or load another page in the same tab (47%).



(a) Reader behavior on desktop devices.

(b) Reader behavior on mobile devices.

Figure 4.5: Transition matrices of first-order Markov chains for (a) desktop devices and (b) mobile devices, aggregating reader behavior with respect to citation events when navigating a Wikipedia article with references (Sec. 4.3.5).

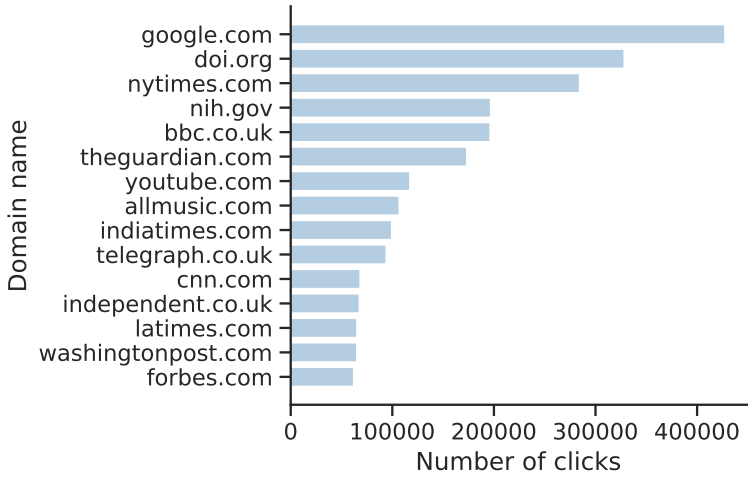


Figure 4.6: Top 15 domain names appearing in English Wikipedia references (Sec. 4.3.4), sorted by number of clicks received during April 2019.

All citation-related events have a very low probability (at most 1.2%) of occurring right after loading a page.

On desktop, reference clicks become much more likely after footnote clicks (34%), and footnote clicks in turn become much more likely after footnote hovers (6.5%), hinting at a common 3-step motif (*fnHover*, *fnClick*, *refClick*), where the reader engages ever more deeply with the citation. Note, however, that this is not true for mobile devices, where, even after readers clicked on a footnote, the probability of also clicking on the citation stays low (0.5%).

Finally, reference clicks (*refClick*) are also common immediately after other reference clicks (8% on desktop, 13% on mobile). Note that for external links outside of the *References* section (*extClick*) we see a different picture: such external clicks are only rarely followed by interactions with citations (*fnHover*, *fnClick*, *refClick*), and in the majority of cases (59% on desktop, 53%

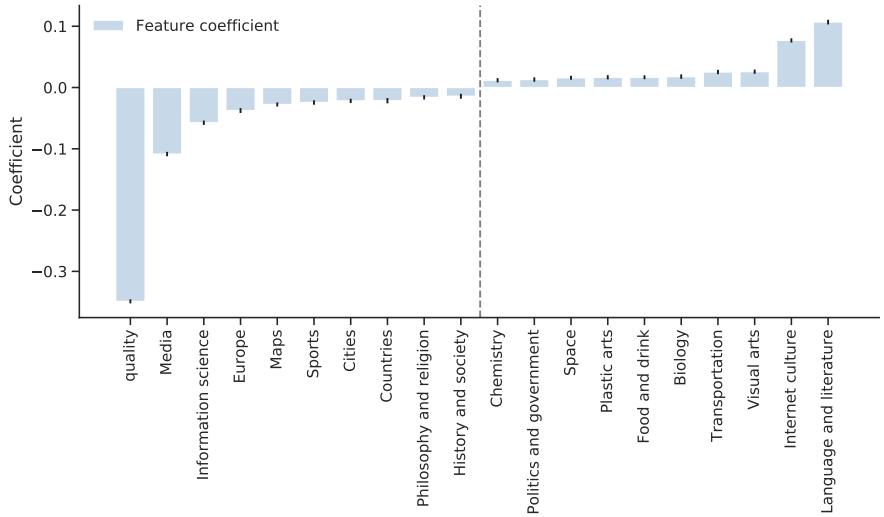


Figure 4.7: Contribution of features to logistic regression model predicting if *refClick* event will eventually occur after page load (Sec. 4.4.1). Top 10 positive and negative coefficients shown, with 95% CIs.

on mobile) they conclude the session, suggesting that Wikipedia is in these cases commonly used as a gateway to external websites.

4.4 Page-level analysis of citation interactions

We now proceed to our second research question, which asks what features of a Wikipedia page predict whether readers will engage with the references it contains.

4.4.1 Predictors of reference clicks

As a first step, we perform a regression analysis. We train a logistic regression classifier for predicting whether a given *pageLoad* event will eventually be followed by a *refClick* event. To assemble the training set, we first find sessions with at least one (positive) *pageLoad* followed by a *refClick* and at least one (negative) *pageLoad* not followed by a *refClick*, and make sure to include at most one such pair per session in order to avoid over-representing power users with extensive sessions. The dataset totals 938K pairs, which we split into 80% for training and 20% for testing.

As predictors we use the article's *topic* vector (with entries from $[0,1]$; Sec. 4.2.4) and the *quality* label (Sec. 4.2.4), which we also normalize to a score in the range $[0,1]$ using the mapping from a previous study [66]. We did not use the number of references and the length of the page, as they are important features in the quality model and would cause collinearity issues due to their high correlation with quality (Pearson's correlation 0.81 and 0.75, respectively).

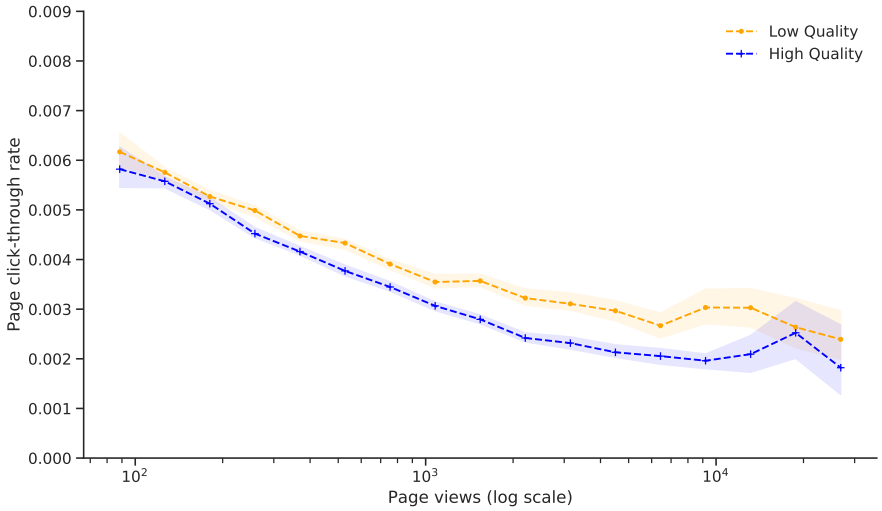


Figure 4.8: Comparison of page-specific click-through rate for low- (yellow) vs. high-quality (blue) articles, as function of popularity (Sec. 4.4.2). Error bands: bootstrapped 95% CIs.

The resulting regression model has an area under the ROC curve (AUC) of 0.6 on the testing set. A summary of the 10 most predictive positive and negative coefficients is given in Fig. 4.7. By far the most important predictor—with a large negative weight—is the article's quality. Moreover, some topics are positive predictors (e.g., "Language and literature", which also includes all biographies, as well as "Internet culture"), while others are negative predictors (e.g., "Media", "Information science").

Given the importance of the quality feature in this first analysis, we now move to investigating its role in a more controlled study.

4.4.2 Effects of page quality

To come closer to a causal understanding of the impact of an article's quality on readers' clicking citations in the article, we perform a matched observational study. The ideal goal would be to compare the page-specific CTR (Eq. 4.2) for pairs of articles—one of high, the other of low quality—that are identical in all other aspects.

Propensity score. Finding such exact matches is unrealistic in practice, so we resort to propensity score matching [12], which provides a viable solution. The propensity score specifies the probability of being treated as a function of the observed (pre-treatment) covariates. Crucially, data points with equal propensity scores have the same distribution over the observed covariates, so matching treated to untreated points based on propensity scores will balance the distribution of observed covariates across treatment groups.

In our setting, we define being of high quality as the treatment and estimate propensity scores via a logistic regression that uses topics, length, number of citations, and popularity as

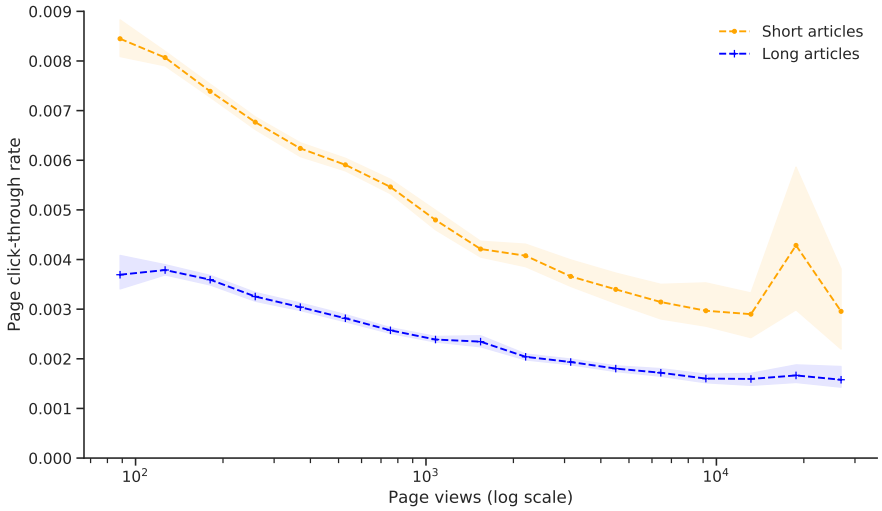


Figure 4.9: Comparison of page-specific click-through rate for short (yellow) vs. long (blue) articles, as function of popularity (Sec. 4.4.3). Error bands: bootstrapped 95% CIs.

observed covariates in order to predict quality as the binary treatment variable. We consider as low-quality all articles tagged as *Stub* or *Start* (74% of the total; Fig. 4.2c), and as high-quality the rest. Articles without a *refClick* or fewer than 100 *pageLoad* events are discarded in order to avoid noisy estimates of the page-specific CTR. This leaves us with 854K articles.

Matching. We compute a matching (comprising 198K pairs) that minimizes the total absolute difference of within-pair propensity scores, under the constraint that the length of matched pages should not differ by more than 10%. This constraint is necessary to ascertain balance on the page length feature because page length is so highly correlated with quality (Pearson correlation 0.81; cf. Sec. 4.4.1). After matching, we manually verify that all observed covariates, including page length, are balanced across groups.

Results. Fig. 4.8 visualizes the average page-specific CTR for articles of low (yellow) and high (blue) quality as a function of article popularity. We can observe that the CTR of low-quality articles significantly surpasses that of high-quality articles across all levels of popularity. In interpreting this result, it is important to recall that page length is one of the most important features in ORES [66], the quality-scoring model we use here. As we control for page length, the gap observed in Fig. 4.8 may be attributed to the remaining features used by ORES, such as the presence of an infobox, the number of images, and the number of sections and subsections.

We hence dedicate our next, final page-level analysis to estimating the impact of page length alone on page-specific CTR.

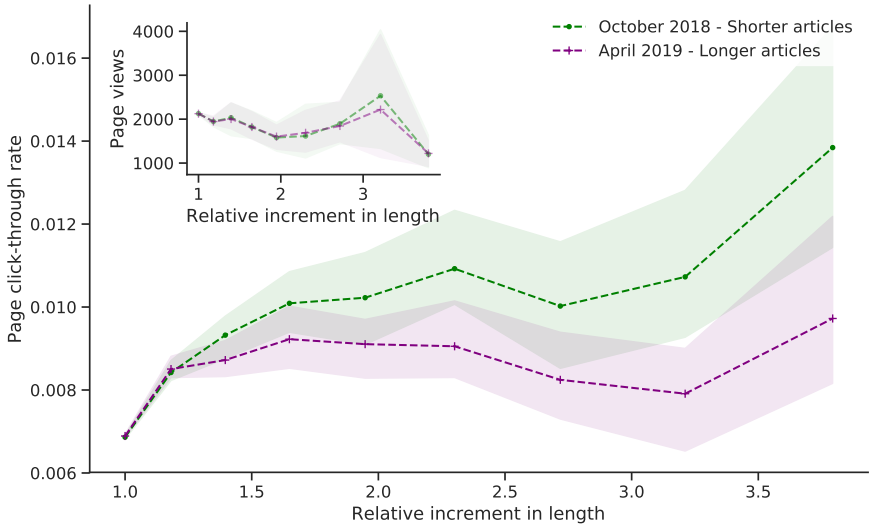


Figure 4.10: Comparison of page-specific click-through rate of shorter (green) vs. longer (purple) revisions of identical articles, as function of length ratio (Sec. 4.4.3). Inset: popularity as function of length ratio. Error bands: bootstrapped 95% CIs.

4.4.3 Effects of page length

In order to measure the effect of page length on CTR, we take a two-pronged approach, first via a cross-sectional study using propensity scores, and second via a longitudinal study.

Cross-sectional study. First, we conduct a matched study based on propensity scores analogous to Sec. 4.4.2, but now with page length as the treatment variable (using the longest and the shortest 40% of articles as treatment groups), and all other features (except quality) as observed covariates. Matching yields 683K pairs, and we again manually verify covariate balance across treatment groups.

The average page-specific CTR of short articles (0.68%) is more than double that of long articles (0.27%; $p \ll 0.001$ in a two-tailed Mann–Whitney U test). Moreover, as seen in Fig. 4.9, this relative difference obtains across all levels of article popularity.

Longitudinal study. While in the above cross-sectional study propensity score matching ensures that the covariates of long vs. short articles are indistinguishable at the aggregate treatment group level, it does not necessarily do so at the pair level. Also, we did not include as observed covariates features describing the users who read the respective articles, and it might indeed be the case that users with a liking for short, niche articles also have a higher probability of clicking citations. In order to mitigate the danger of such remaining potential confounds and achieve even finer control, we now conduct a longitudinal study to assess how a variation in length of the *same* article impacts its CTR.

To do so, we select all articles that grew in length between October 2018 and April 2019, our two data collection periods (Sec. 4.2.2). To control for the effect of page popularity, which was observed to negatively correlate with CTR (Fig. 4.8 and 4.9), we assign a popularity level to each article by binning page view counts into deciles and discard articles whose popularity level has changed between the two periods. This way, we obtain a set of 120K articles with matched long and short revisions.

By grouping these articles by the length ratio of their two revisions and plotting this ratio against the CTR for the long (purple) vs. short (green) versions (Fig. 4.10), we provide a further strong indicator that page length causally decreases the prevalence of citation clicking. According to a Mann–Whitney U test, the CTR difference between long and short revisions is statistically significant with $p < 0.05$ starting from a length increase of 17%, and with $p < 0.01$ from 31%. In addition, to verify that the effect is not confounded by a concomitant change in article popularity, the inset plot in Fig. 4.10 shows that the popularity indeed stays constant between revisions.

4.5 Link-level analysis of citation interactions

Our final research question asks which features of a specific reference predict if readers will engage with it. Note that this is different from RQ2 (Sec. 4.4), where we operated at the page level and did not differentiate between different references on the same page.

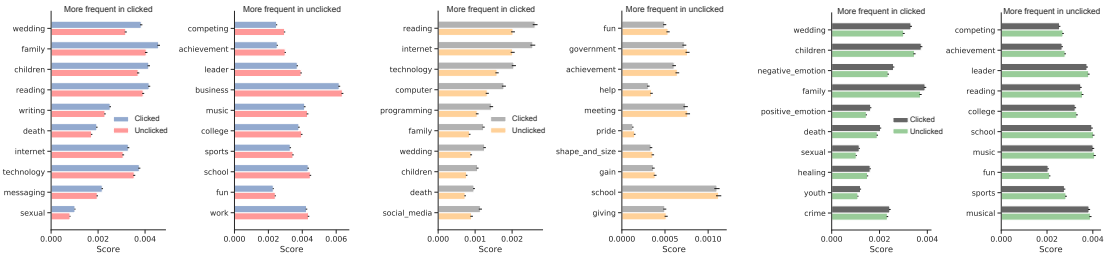
4.5.1 Predictors of reference clicks

We begin with a regression analysis to detect which features predict whether a reference will be clicked. We selected all the references with external links, and we carefully rule out a host of confounds by sampling pairs of clicked and unclicked references from the same page view, thus controlling for situational features such as the page, user, information need, etc. As we saw in Fig. 4.4, references at the top and bottom of pages are *a priori* more likely to be clicked. Thus, to exclude position as a confound and maximize the probability that the user saw both references in a pair, we pick as the unclicked reference in a pair the one that appears closest in the page to the clicked reference. To make sure we sample references associated with a sentence, we discard all footnotes in tables, infoboxes, and images, and keep only those within the article text. Finally, we again sample only one pair per session in order to avoid over-representing readers who are more prone to click on references. This process yields 1.8M reference pairs.

As predictors we use the words in the sentence that cites the respective reference, as well as the words in the reference text (cf. Sec. 4.2.1), represented as binary indicators specifying for

	Positive contribution				Negative contribution			
	In sentence		In reference		In sentence		In reference	
	Word	Coeff.	Word	Coeff.	Word	Coeff.	Word	Coeff.
All topics	greatest	0.36	know	0.25	debut	-0.25	awards	-0.33
	born	0.28	pmc	0.24	moved	-0.16	deadline	-0.32
	died	0.23	2019	0.21	worked	-0.16	billboard	-0.17
	website	0.23	website	0.21	awarded	-0.16	register	-0.17
	ranked	0.23	dies	0.20	joined	-0.13	link	-0.16
	known	0.20	former	0.19	began	-0.13	isbn	-0.15
	professional	0.19	family	0.16	appeared	-0.12	board	-0.14
	relationship	0.19	behind	0.15	score	-0.11	variety	-0.14
	rating	0.18	allmusic	0.15	festival	-0.11	next	-0.14
	article	0.18	story	0.15	attended	-0.11	archive	-0.13
STEM	online	0.25	definition	0.30	requirements	-0.17	oclc	-0.26
	tests	0.23	2019	0.24	run	-0.17	best	-0.23
	2019	0.23	free	0.22	rather	-0.16	jstor	-0.22
	short	0.17	pmc	0.21	another	-0.15	evaluation	-0.16
	known	0.17	website	0.20	said	-0.15	wiley	-0.16
	algorithms	0.16	pdf	0.19	launched	-0.15	london	-0.15
	published	0.16	overview	0.17	less	-0.14	isbn	-0.14
	defined	0.15	methods	0.15	make	-0.12	internet	-0.14
	programming	0.15	introduction	0.14	better	-0.12	industrial	-0.14
	digital	0.15	years	0.13	popular	-0.12	source	-0.14
Culture	article	0.30	daughter	0.36	indicating	-0.42	awards	-0.36
	born	0.28	obituary	0.31	premiered	-0.28	award	-0.33
	greatest	0.27	know	0.31	chart	-0.21	deadline	-0.28
	professional	0.27	instagram	0.29	debut	-0.21	cast	-0.22
	died	0.26	boy	0.28	moved	-0.20	global	-0.21
	known	0.25	sex	0.25	began	-0.17	next	-0.19
	ranked	0.24	wife	0.24	earned	-0.16	isbn	-0.18
	relationship	0.23	former	0.24	recorded	-0.16	drama	-0.18
	website	0.23	historic	0.24	alongside	-0.16	standard	-0.18
	sexual	0.23	2019	0.23	worked	-0.16	tour	-0.18
History and Society	born	0.29	definition	0.43	came	-0.20	jstor	-0.25
	website	0.21	overview	0.22	award	-0.16	record	-0.21
	2019	0.21	best	0.19	transportation	-0.13	link	-0.20
	died	0.20	2019	0.19	protection	-0.12	2002	-0.17
	currently	0.19	website	0.19	member	-0.12	election	-0.16
	known	0.17	statistics	0.17	began	-0.11	1998	-0.15
	referred	0.17	death	0.16	originally	-0.11	ed	-0.15
	customers	0.16	last	0.16	specific	-0.11	isbn	-0.15
	study	0.16	ship	0.15	awarded	-0.10	announces	-0.14
	activities	0.15	top	0.15	addition	-0.10	board	-0.12
Geography	politician	0.50	woman	0.34	debut	-0.45	crime	-0.28
	born	0.26	know	0.27	missing	-0.22	awards	-0.28
	magazine	0.25	dies	0.26	career	-0.21	register	-0.24
	believed	0.23	family	0.23	timmothy	-0.20	link	-0.24
	married	0.23	website	0.20	executive	-0.19	interview	-0.19
	ranked	0.22	mail	0.19	episode	-0.17	2000	-0.17
	video	0.22	father	0.18	months	-0.17	culture	-0.17
	directed	0.18	son	0.18	close	-0.15	htm	-0.16
	crime	0.18	boy	0.18	case	-0.15	music	-0.15
	natural	0.18	biography	0.17	appointed	-0.15	paris	-0.15

Table 4.1: Top positive and negative predictors (words) of reference clicks (Sec. 4.5.1), for different article topics. Words are organized based on where they appear: in the sentence annotated by the reference, or in the reference text.



(a) Click event (sentence text) (b) Click event (reference text) (c) Hover event (sentence text)

Figure 4.11: *Empath* [48] topics most strongly (anti-)associated with citation events (cf. Sec. 4.5.2 for description). Reference text not studied for hover event (Sec. 4.5.3) because unlikely to be visible to user before hovering.

each of the 1K most frequent words whether the word appears in the sentence.[IX] Using these features as predictors, we train a logistic regression to predict the binary click indicator.

We perform this analysis on the full above-described dataset, as well as on subsets consisting only of page views from each of 4 broad categories (derived by aggregating the 44 WikiProjects categories from Sec. 4.2.4): "Culture" (1.3M pairs), "STEM" (436K), "Geography" (530K), and "History and Society" (467K). The model achieves a testing AUC of around 0.55 across these 5 settings.

The words with the largest and smallest coefficients are displayed in Table 4.1, where we observe that, for all article topics except for "STEM", many positive features are related to social and life events and relationships ("dies", "obituary", "married", "wife", "relationship", "sex", "daughter", "family", etc.). Another common pattern across topics is that "2019" is strongly related with clicking, and that career-related references ("awards", "debut", etc.) are less likely to be clicked. We shall further discuss these observations in Sec. 4.6.

On STEM-related pages, open-access references seem to receive more clicks than others, with words like "free" and "pdf" among the top predictors, whereas words related to traditionally closed-access libraries such as JSTOR appear among the negative predictors, in line with previous findings [184].

4.5.2 Topical correlates of reference clicks

For a higher-level view, we perform a topical analysis of citing sentences and reference texts, separately for the clicked vs. the unclicked references from the paired dataset of Sec. 4.5.1.

To extract topics, we use *Empath* [48], which comes with a pre-trained model for labeling input text with a distribution over 200 wide-ranging topics. After applying the model to each data point, we compute the average topic distribution for clicked and unclicked references,

[IX] Stop words were removed, and numbers (except for 4-digit numbers that potentially represent years) were converted to a special number token.

Positive		Negative	
Word	Coeff.	Word	Coeff.
killer	0.16	oclc	-0.22
greatest	0.16	jason	-0.16
critic	0.15	episode	-0.15
things	0.15	die	-0.15
daughter	0.15	dictionary	-0.13
reveals	0.14	spanish	-0.12
baby	0.14	isbn	-0.12
instagram	0.13	le	-0.11
wife	0.13	board	-0.11
sheet	0.13	channel	-0.11

Table 4.2: Top 10 positive and negative predictors (words) of reference click following footnote hover (Sec. 4.5.4).

respectively, and sort topics by the signed difference between their probability for clicked vs. unclicked references.

The topics with the largest positive and negative differences are listed in Fig. 4.11a and 4.11b for citing sentences and reference texts, respectively. The results corroborate those from Sec. 4.5.1, with human factors (wedding, family, sex, death) being more prominent among clicked references, whereas career-related topics such as competitions or achievements receive less attention. Among the most prominent topics for reference texts (Fig. 4.11b), topics related to technology and the Internet also emerge.

4.5.3 Predictors of footnote hovering

The analyses of Sec. 4.5.1 and 4.5.2 considered engagement via reference clicks. As we observed in Fig. 4.3, on desktop devices, hovering over a footnote to reveal the reference text in a tooltip is an even more common way to interact with references. We hence replicated the above analyses with the *fnHover* instead of the *refClick* event (8.7M reference pairs), with the only difference that we excluded words from reference texts as features, since the user is unlikely to have seen those words before hovering over the footnote.

The results echo those of Sec. 4.5.1 and 4.5.2, so for space reasons we do not discuss the regression analysis for footnote hovering (cf. Sec. 4.5.1) and focus on the topical analysis instead (cf. Sec. 4.5.2). Inspecting Fig. 4.11c, we observe that we see a stronger tendency of *fnHover* events, compared to *refClick* events, to be elicited by words that are related to both positive and negative emotions.

4.5.4 Predictors of reference clicks after hovering

Once a user hovers over a (*fnHover*), the text of the corresponding reference is revealed in a so-called reference tooltip (Fig. 4.1). At this point, the user has the choice to either click through to the citation URL (*refClick*) or to stay on the article page. As final analysis, we are interested in understanding what words in the reference text influence the user when making this decision.

We create a dataset by selecting the page loads with at least two footnote hover events, where one converted to a *refClick* (positive), whereas the other did not (negative). As in the previous studies, we selected at most one random pair per session, giving rise to a dataset of 440K pairs of hover events.

Similar to the study in Sec. 4.5.1, we represent reference texts as 1K-dimensional word indicator vectors and use them as predictors in a logistic regression to predict *refClick* events (testing AUC 0.54).

The strongest coefficients are summarized in Table 4.2, painting a picture consistent with the previous analyses: readers, after seeing a reference preview via the tooltip, are more likely to click on the cited link when the reference text mentions social and life aspects ("wife", "baby", "instagram", etc.). The strongest negative coefficients suggest that readers tend to not click through to dictionary entries, book catalogs (ISBN, OCLC), and information in languages other than English: manual inspection revealed that "spanish" is mainly due to the note "In Spanish", "le" is the French article common in French newspaper names (e.g., *Le Monde*), and "die" is a German article.

4.6 Discussion

Our analysis provides important insights regarding the role of Wikipedia as a gateway to information on the Web. We found that fewer than 1 in 300 page views lead to a citation click. In our analysis, we focused on the fraction of users who engage with references, and characterized **how Wikipedia is used as a gateway to external knowledge.** Our findings suggest the following.

- **We engage with citations in Wikipedia when articles do not satisfy our information need.** Sec. 4.4 showed that readers are more likely to click citations on shorter and lower-quality articles. Although this result seemed counter-intuitive at first, since higher-quality articles actually contain *more* references that could potentially be clicked, it is in line with the finding that citations to sources reporting atomic facts that are typically available in Wikipedia articles (e.g., awards, career paths), are also generally less engaging (Sec. 4.5). Collectively, these results suggest that readers are inclined to seek content beyond Wikipedia when the encyclopedia itself does not satisfy their information needs.

- **Citations on less engaging articles are more engaging.** In all of Sec. 4.4 we found that citation click-through rates decrease with the popularity of an article. While this may follow from the previous point because long, high-quality articles tend to be more popular, it may also suggest that less popular articles are visited with a specific information need in mind. Previous work indeed suggests that popular articles are more likely to be viewed by users who are randomly exploring the encyclopedia [171].

- **We engage with content about people's lives.** We clearly saw that readers' interest is particularly high in references about people and their social and private lives (Sec. 4.5). This is especially true for hovers, a less cognitively demanding form of engagement with citations. Hover events are also more likely to be elicited by words that are related to emotions, both positive and negative.

- **Recent content is more engaging.** We found that references about recent events (whose text includes "2019") are more engaging, both in terms of hovering and clicking.

- **Open content is more engaging.** Finally, we saw that references in Wikipedia pages about science and technology, especially if they point to a open-access sources (e.g., having "free" or "pdf" in the reference text), are also more likely to be clicked.

Theoretical implications. Our findings furnish novel insights about Web users and their information needs through the lens of the largest online encyclopedia. For the first time, by characterizing Wikipedia citation engagement, we are able to quantify the value of Wikipedia as a gateway to the broader Web. Our findings enable researchers to develop novel theories about readers' information needs and the possible barriers separating knowledge within and outside of the encyclopedia. Our research can also guide the broader community of Web contributors in prioritizing efforts towards improving information reliability: we found that people especially rely on cited sources when seeking information about recent events and biographies, which suggests that Web content in these areas should be especially well curated and verified. Finally, the fact that readers engage more with freely accessible sources highlights the importance of open access and open science initiatives.

Practical implications. Quantifying Wikipedia article completeness has proven to be a non-trivial task [147]. The notion that article completeness is highly related to readers' engagement with Wikipedia references opens up ideas for novel applications to help satisfy Web users' information needs, including models that quantify lack of information in an article by incorporating signals related to reference click-through rate. Our findings will also help prioritize areas of content to be checked for citation quality by Wikipedia editors: in areas of content where Wikipedia acts as a major gateway, the quality and reliability of sources that readers visit become even more crucial. Finally, the data we collected could empower a model that, given a sentence missing a citation (i.e., with a *citation needed* tag), could quantify how likely readers are to be interested in accessing the corresponding information and thereby help Wikipedia editors prioritize the backlog of unsolved missing-reference cases.

Limitations and future work. The overall low AUC (0.54 to 0.6) of the regression models (Sec. 4.4–4.5) emphasizes the inherent unpredictability of reader behavior. While the significantly above-chance performance renders the models useful for analyzing the impact of various predictors, their performance is currently too low to make them useful as practical predictive tools. Future work should hence invest in more powerful sequence models to improve accuracy.

By focusing on English Wikipedia only, the present analysis provides a limited view of the broader Wikipedia project, which is available in more than 300 languages and accessed by users all over the world. In our future work, we therefore plan to replicate this study for other language editions. So far, we also omitted any user characteristics from our study, such as more global behavioral traits beyond the page-view level, as well as geographic information, which are known to play an important role in user behavior [108, 182]. Future work should incorporate such signals.

We will also investigate reader intents more closely. While click and hover logs reflect the extent to which readers are interested in knowing more about a given topic, they cannot tell us about the specific circumstances that led the user to engage by clicking or hovering, nor about the level of satisfaction achieved by following up on a reference. In the future, we plan to better understand these aspects via qualitative methods such as surveys and interviews.

Further, whereas this analysis focused on links in the *References* section of articles, Chapter 5 introduces our study on other types of external links (cf. Fig. 4.1) in satisfying readers' information needs.

Finally, as exogenous events strongly affect Wikipedia users' information needs [171], future work should go beyond studying Wikipedia as an isolated platform and analyze how citation interaction patterns are warped by breaking news and events with uncertain information. This will sharpen our picture of Wikipedia as a gateway to global information.

Conclusion. This chapter characterized how readers engage with the citations in the articles. Using a large-scale dataset collected from the clients over one month, we describe the properties of pages and references associated with more clicks. The next chapter will continue the characterization of how readers leave Wikipedia by focusing on the other types of external links, with a special emphasis on official links.

5 On the Value of Wikipedia as a Gateway to the Web

5.1 Introduction

Thanks to the collaborative effort of a community of volunteer editors, Wikipedia is the world's largest encyclopedia and an important source of information for millions of people. Wikipedia serves its content as a regular website, allowing editors to add hyperlinks in order to enable readers to more easily find additional content, both internal and external. Internal links help readers locate relevant encyclopedic content by navigating from article to article. In contrast, external links enrich articles with additional content that should not or cannot be included in Wikipedia itself. There are various reasons to add external links,[I] with linked content ranging from official websites, to news articles used as references,[II] to copyrighted material.

Figure 5.1: Example of an official link, in infobox of Wikipedia article about Internet Archive's Wayback Machine.

In this study, we are interested in quantifying and characterizing the outgoing traffic generated by Wikipedia towards external content. Given Wikipedia's crucial societal role and global reach, it is essential to understand how it interacts with the broader Web by driving traffic to external websites. The resulting insights can inform the platform's future design and thus allow it to better cater to readers' information needs around external content. As Web traffic has monetary value—in particular when the traffic goes to commercial websites—an investigation of the external traffic generated by Wikipedia also sheds new light on the poorly understood role it has as a provider not only of information, but also of economic wealth.

[I] https://en.wikipedia.org/wiki/Wikipedia:External_links
[II] https://en.wikipedia.org/wiki/Wikipedia:Citing_sources

Research questions. We approach the question of Wikipedia's value as a gateway to the Web from two angles: informational and economic. Concretely, we pose three research questions:

RQ1 **Level of engagement with external links:** What total volume of traffic does Wikipedia drive to third-party websites? What is the click-through rate of external links, and how does it vary across types of linked content? (Sec. 5.3)

RQ2 **Patterns of engagement with external links:** How do users interact with external links? Do they click through fast or slow, and how does this vary across types of linked content? In what navigational situations do clicks to external websites occur? (Sec. 5.4)

RQ3 **Economic value of external links:** What is the monetary value of the traffic from Wikipedia to external websites? If website owners had to pay for an equivalent amount of traffic via sponsored search, how much would this cost? (Sec. 5.5)

Summary of findings. Based on usage logs gathered over a one-month period from English Wikipedia users' client devices (Sec. 5.2), we quantified the level of engagement with external links **(RQ1)**, determining that English Wikipedia generated 43 million clicks to external websites during the month we studied, despite the fact that, on average, the click-through rate (CTR) of external links was only 0.08%. While most external links (95.5%) occurred in article bodies and cited references (accounting for about two thirds of the external traffic), a disproportionately large fraction (23%) of the total traffic came from a relatively small fraction (0.8%) of all external links, namely from *official links* to the website of the entity covered in the respective article. Such official links are regularly listed in so-called *infoboxes,* short tabular summaries of key facts about the covered entity (see Fig. 5.1 for an example). Since official links witnessed a vastly increased CTR of 2.47% (vs. 0.08% over all external links), we focused our analysis on official links. In a topical analysis, we found that official links associated with articles about businesses, educational institutions, and websites had the highest CTR (a first indicator of the economic value of Wikipedia's external links), whereas official links associated with articles about geographical content, television, and music had the lowest CTR.

By analyzing patterns of engagement with external links **(RQ2),** we observed that Wikipedia frequently serves as a stepping stone between search engines and third-party websites. We captured this effect quantitatively as well as in a manual analysis, where we found that URLs that are down-ranked or censored by search engines, and thus not retrievable via search, can often be found in Wikipedia infoboxes, which leads search users to take a detour via Wikipedia. We conclude that Wikipedia regularly and systematically meets information needs that search engines do not meet, which further confirms Wikipedia's central role in the Web ecosystem.

Finally, we aimed to quantify the hypothetical economic value of the clicks received by external websites from English Wikipedia **(RQ3).** Wikipedia is, of course, free, and it runs thanks to the donations of thousands of people. We thus cannot ask how much money Wikipedia could earn by charging a fee for external clicks—this hypothetical scenario is simply too far from

Table 5.1: Click statistics for external links embedded in Wikipedia articles.

Link location	**Total links**		**Clicks**		**Total**	**Mean**	**Median**
	Number	Perc. of total	Number	Perc. of total	**articles**	**CTR ± SD***	**click time†**
Infobox	2.8M	4.5%	12.5M	29.1%	1.3M	0.90% ± 2.2%	18.6s (45.6s)
Official links	*506K*	*0.8%*	*9.8M*	*22.7%*	*506K*	*2.47% ± 3.0%*	*20.7s (47.8s)*
Body	24.9M	39.5%	16.2M	37.8%	4.0M	0.14% ± 0.7%	35.4s (90.9s)
References	35.3M	56.0%	14.2M	33.1%	3.9M	0.03% ± 0.2%	51.8s (131.4s)
All links	63.1M	100%	43.1M	100%		0.08% ± 0.5%	32.9s (87.1s)

*CTR = click-through rate; considering only links with at least 300 impressions during the one-month study period. †Inter-quartile range in parentheses.

reality—but we may approach the question from a different angle, asking how much money external-website owners would have to pay in order to obtain an equivalent number of clicks by other means, such as paid ads. In this spirit, we applied the Google Ads API to the content of official websites linked from Wikipedia in order to generate keywords for sponsored search and estimated their cost per click at market price. We conclude that the owners of external websites linked from English Wikipedia's infoboxes would need to collectively pay a total of around $7–13 million per month (or $84–156 million per year) for sponsored search in order to obtain the same volume of traffic that they receive from Wikipedia for free.

These numbers exceed even the ballpark guess given in a bullish 2013 analysis [87] that, unlike ours, was not based on real click logs, but on generic rates commonly assumed in the online ad industry, and estimated that Wikipedia could earn $2.5 million monthly via affiliate links. Although our analysis of monetary value should mostly be taken as an indicative "back-of-the-envelope" calculation, it highlights the importance of Wikipedia not only as a source of information, but also as a gratuitous provider of economic wealth.

Economic value of Wikipedia. The value of Wikipedia to the world is not only high but also difficult, if not impossible, to quantify in purely economic terms. It has been shown that Wikipedia is essential—or has the potential to be—in a variety of spillovers with substantial economic impact. For instance, it is of critical importance to Web search engines, such as Google [120], and has also been shown to be useful to improve, or even predict, financial markets [126, 127, 211]. Wikipedia can be used to inform economic development policies [168], improve the visibility of places, with direct positive consequences on tourism [77], and even predict and monitor global health and diseases [57, 76]. Furthermore, Wikipedia has been shown to influence the very development of science [186]. Nevertheless, and perhaps surprisingly, to the best of our knowledge Wikipedia's economic value as an information gateway to the Web has rarely been discussed in previous work. In a rare exception, researchers have considered the value of Wikipedia in providing traffic to Reddit and Stack Overflow [192].

5.2 Data

5.2.1 Wikipedia client-side logs

In order to analyze user engagement with Wikipedia's external links, we made use of the dataset described in details in Sec. 4.2. This dataset consists of logs of all reader interactions with external links in English Wikipedia articles over the one-month period from 24 March to 21 April 2019. The data was captured by the browser on the client side and includes all clicks on external links and a uniformly random sample (33%) of all pageview events, organized into *sessions,* i.e., sequences of events from the same user in the same browser tab. This article reports results using the full dataset when describing external-click events. Whenever pageview counts are involved, we extrapolate from the 33% sample.

The data was collected in accordance with Wikimedia's privacy policy[III] and processed exclusively on Wikimedia computing machinery. Although the data does not contain personally indentifiable information beyond what is implicit in browsing behavior, it cannot be shared publicly. For transparency, we publish our data analysis code at https://github.com/epfl-dlab/WikipediaAsWebGateway.

5.2.2 Article characteristics

At the time of data collection, Wikipedia had around 5.8M articles, which were loaded by readers more than 4.5 billion times during the month studied. We characterized each article by popularity (pageviews during the month studied) and length (number of characters). The popularity distribution was very skewed: 50% of the articles had fewer than 42 views, 90% had fewer than 894 views. In contrast, the average number of pageviews in one month was 700. The most visited 1,550 articles, which represented 0.02% of all articles, accounted for 10% of all pageviews. The most visited pages were articles about topics that were trending in April 2019, such as NIPSEY HUSSLE (5.7M views), NOTRE-DAME DE PARIS (4.7M), BONNIE AND CLYDE (3.5M), or GAME OF THRONES (SEASON 8) (2.6M). Most of the articles were short, and similar to popularity, the number of characters showed a skewed distribution, with a median of 3,888, and an average of 7,793 characters.

ORES topics. ORES[IV] is a toolkit offered by Wikimedia that, among other things, includes functionality for labeling articles with topics, based on a manually curated taxonomy of 64 topics [67] derived from WikiProjects.[V] Based on this categorization, ORES offers a classifier that predicts, for a given article, its probability of belonging to each of the 64 topics. Unlike the model used in Chapter 4, this updated version returns a more fine-grained taxonomy with topics such as BIOGRAPHY. Since a single article may belong to multiple topics, the 64 probabilities generally do not sum to 1. We used topic labels in binarized form, considering

[III] https://foundation.wikimedia.org/wiki/Privacy_policy
[IV] https://www.mediawiki.org/wiki/ORES
[V] https://en.wikipedia.org/wiki/Wikipedia:WikiProject

an article to belong to a topic if the corresponding probability is greater than 50%. Note that, although the taxonomy is hierarchically organized in two levels, in this work we only considered the 57 lower-level topics (listed along the x-axis of Fig. 5.3). Having run the classifier on all articles in the dataset, we observed that overall the most common topic was BIOGRAPHY (1.7M articles), followed by SPORTS (1.4M) and NORTH AMERICA (950K). The least common topics were EASTERN AFRICA (11K) and LIBRARIES & INFORMATION (14K).

5.2.3 External links

External links form our central object of study, so we extracted detailed information about them from Wikipedia articles. Since parsing content from articles in wikitext format might result in missing hyperlinks [125], we extracted the external links from the articles in HTML format instead. As this study focuses specifically on those Wikipedia links that lead to external websites, we adopt the convention that, whenever we simply say "link", we implicitly mean "external link".

We partitioned the external links in the dataset into three classes, according to their position on the page: *infobox,* article *body,* and *references*. *Infoboxes* are tables (typically rendered by the browser on the right-hand side of the page on desktop devices, or at the top of the page on mobile devices) that summarize key information by adding semi-structured content (see Fig. 5.1). In addition to images and textual properties, this area can contain—potentially many—external links pointing to external geolocation services, official registries, or official websites. The links in an article *body* appear inline within the main textual content of the page or in dedicated sections such as "External links" or "See also". Links in article bodies are more heterogeneous, including links to social media pages, PDF documents, or related external material. Finally, we considered as *references* all links used to cite external content in support of a statement. Typically they appear at the bottom of the page, reachable from the article body via numbered link anchors.

During the period considered, Wikipedia had 5.3M articles that contained at least one of 63.1M external links (totaling 49.8M unique target URLs). Table 5.1 (column "Total links") summarizes these values. In total, 35.3M (56.0%) of these links appeared in references, 24.9M (39.5%) in article bodies, and 2.8M (4.5%) in infoboxes. Around 1.3M articles in English Wikipedia had an infobox with links, and the average number of links per infobox in these articles was 2.08. Links spanned from official company links (e.g., schlenkerla.de) to geocoordinates on geohack.toolforge.org to institutional registries (e.g., National Register of Historic Places).

5.2.4 Official links

To further qualify the infobox links, we designed a binary classifier that can distinguish between *official links* and other types of link. It was trained on a random sample of 2,000 infobox links,

manually annotated as "official" or "other". This resulted in a training set of 387 official links and 1,613 other links. To characterize each link, we then computed the following features:

- **URL length:** Number of characters in the URL path (guided by the intuition that official-link URLs tend to be short).
- **Similarity of URL with article title:** Motivated by the usefulness of character n-grams for URL-based topic classification [17], we computed the character n-grams ($n = 1, \ldots, 4$) of the title of the article where the link was placed, the link's anchor text (if non-empty), the domain of the link URL, and the path of the link URL. We then computed, and used as features, the Jaccard similarity of sets of n-grams for three pairs: title/URL-domain, anchor-text/URL-domain, title/URL-path.
- **Similarity of context with marker words:** Jaccard similarities between the character n-gram sets ($n = 1, \ldots, 4$) of high-precision marker words ("official", "website", "homepage", "URL") and of the link's anchor text and context (i.e., the text within the same <TR> tag as the link).

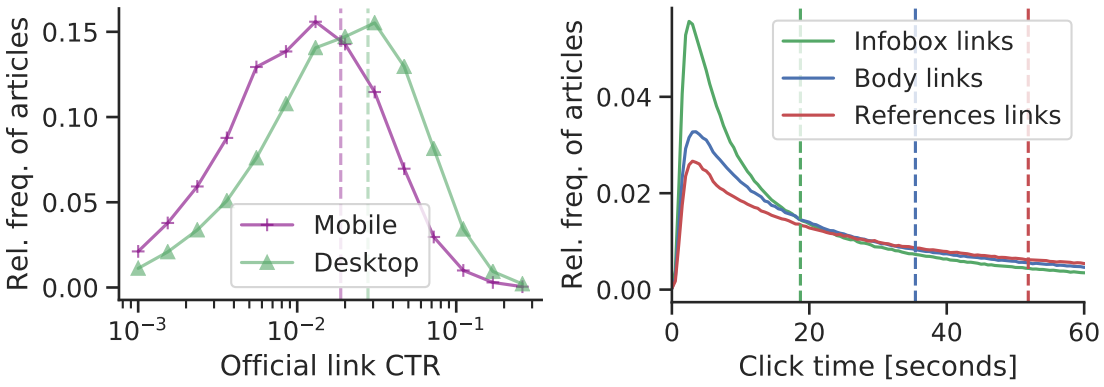
We used these 10 features and the manual labels in a random forest classifier, achieving 5-fold cross-validated precision 0.980 (SD 0.009), recall 0.983 (SD 0.005), and F1 score 0.982 (SD 0.007).

Applying this classifier to all the links in the dataset, we found that 506K of the 63.1M external links corresponded to the official website of the entity described in the respective article. Broken down by article topic, the largest number of official links was associated with NORTH AMERICA (27.2%), EUROPE (25.0%), MEDIA (20.2%), BIOGRAPHY (18.9%), ASIA (17.1%), EDUCATION (10.5%), and BUSINESS AND ECONOMICS (9.5%), with the latter containing mainly articles about companies. (The percentages sum to more than 100% because articles may belong to multiple topics.)

As expected, according to the classifier, the vast majority (98.1%) of articles with an official link had exactly one official link, and conversely, the vast majority (97.1%) of official links appeared as such in exactly one article's infobox. This has the added advantage that official links can be characterized by features derived from their corresponding Wikipedia articles (e.g., topics, content words).

5.2.5 Definitions

Definition: click-through rate (CTR). Our main metric for measuring engagement with external links is the *click-through rate (CTR)*, which, intuitively, is simply the number of times a link was clicked, divided by the number of times the link was displayed (by virtue of being contained in an article that contained the link). In practice, care must be taken, as it frequently happens that the same article is viewed multiple times in the same session, e.g., because the



(a) Click-through rate

(b) Click time

Figure 5.2: Usage of external links. (a) Distribution of click-through rate of official links by device type (vertical lines: means). (b) Distribution of click time by link type (vertical lines: medians).

user refreshes the page or clicks the back button. To guard against overcounting such multiple views, we grouped all pageviews of the same article a that occurred during the same session s and call the unique pair (a, s) one visit of a.

With N_l as the number of visits (i.e., distinct (a, s) pairs) upon which link l was displayed, and C_l as the number of visits upon which link l was clicked, we define the CTR of link l as C_l/N_l, i.e., the fraction of visits upon which l was clicked, out of all visits upon which l was displayed. Since each official link belongs to exactly one article (with extremely rare exceptions; cf. Sec. 5.2.4), we may also, in a slight abuse of terminology, speak of the "CTR of an article", implying the CTR of the official link associated with the article.

In order to reliably estimate CTRs, we need to avoid small denominators, so we restricted our analyses to links that were displayed upon at least 300 pageview events. In the case of official links, this resulted in a set of 160K links (and their corresponding articles).

Definition: click time. In order to capture how long users dwell on an article before they click an external link, we define the notion of *click time,* which measures the number of seconds between the pageview event on which the link was displayed and the click on the link itself. If the same external link was clicked multiple times in the same session, we only considered the first pageview that was accompanied by an external click.

Since click times are unbounded above and follow a heavy-tailed distribution, we used medians, rather than means, for aggregation.

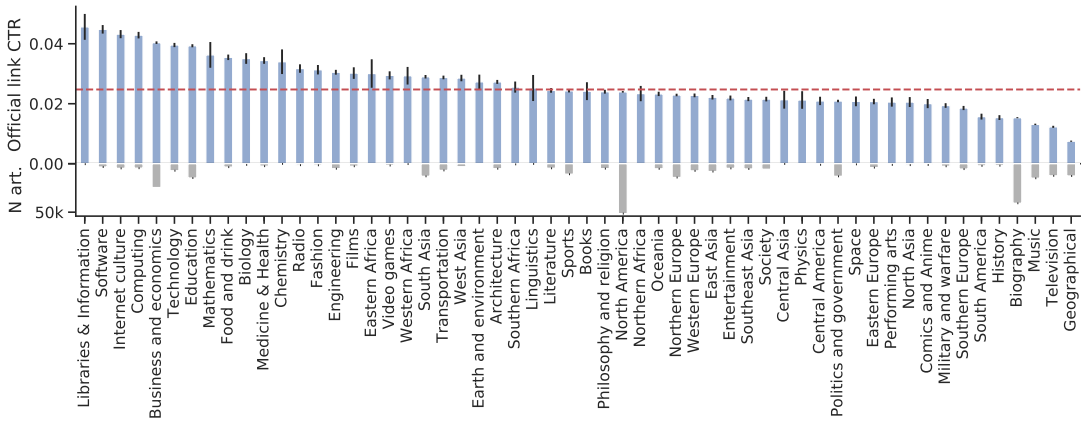


Figure 5.3: Official-link click-through rate by article topic. *Blue bars:* means with bootstrapped 95% confidence intervals. *Gray bars:* number of articles with official links. *Red dashed line:* global mean.

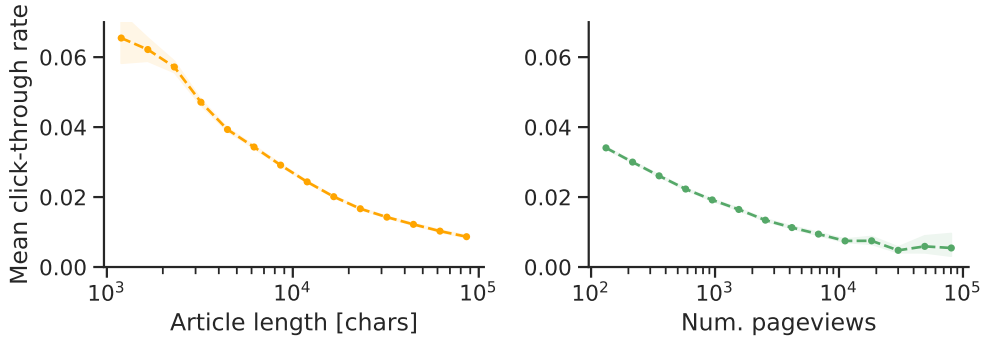
5.3 Level of engagement with external links

We start our analysis by quantifying the level of engagement with external links, both overall (Sec. 5.3.1) and by article topic (Sec. 5.3.2).

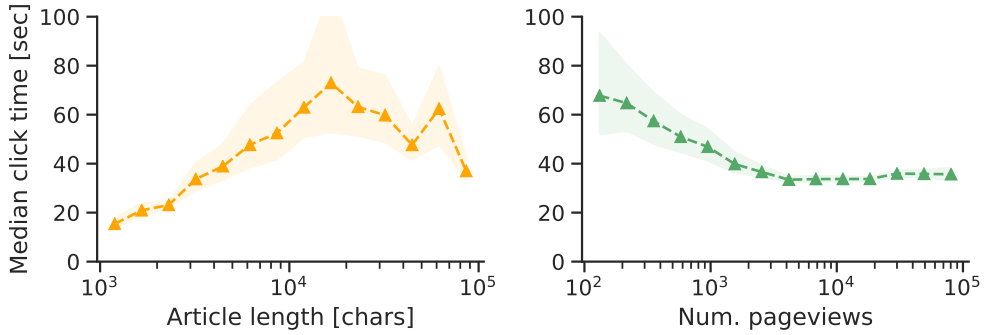
5.3.1 Overall click statistics

Overall click statistics are summarized in Table 5.1. During our one-month data collection period, there was a total of around 4.5 billion Wikipedia pageviews, which led to around 43.1M clicks on external links. The total volume of external clicks was distributed roughly evenly over the three classes of external links: those in infoboxes (12.5M), those in references (14.2M), and those in article bodies (16.2M). As the vast majority of external links was located in references (56.0%) and article bodies (39.5%), the CTR of infobox links (0.90%) vastly exceeded that of links in references (0.03%) and article bodies (0.14%). To ascertain that this was not simply caused by the fact that infobox links appear higher up on the page, we also computed the CTR of article-body links appearing in the top 20% of the page. This yielded a CTR of 0.20%, much closer to the 0.14% of article-body links overall than to the 0.90% of infobox links.

Official links. Official links play a key role. Although they constituted only 18% of the 2.8M infobox links, they accounted for 78% of the 12.5M clicks on infobox links, with a CTR of 2.47%, nearly 3 times as high as that of infobox links overall. The average CTR was even higher on desktop devices, where it reached 2.78%, vs. 1.87% on mobile (Fig. 5.2a). Given their prominence, we shall focus mostly on official links from here on, and unless stated otherwise, we henceforth refer to official links when simply writing “links”.



(a) Click-through rate



(b) Click time

Figure 5.4: (a) Click-through rate and (b) click time of official links as functions of article length (left) and popularity (right), with 95% CIs. Official links on longer pages are clicked more rarely and more slowly; those on more popular pages are clicked more rarely and more quickly.

Geographical differences. The top 5 countries by pageview volume were, in this order, the United States, the United Kingdom, India, Canada, and Australia. They generated 71.6% of the total traffic. Among these countries, the CTR on official links was highest in the U.S. (2.36%), followed by India (2.14%), the U.K. (1.53%), Canada (1.38%), and Australia (1.11%).

5.3.2 Click-through rates by topic

Next, we aim to understand how the click-through rates of official links vary by topic as defined by the ORES classifier introduced in Sec. 5.2.2. Since, in the vast majority of cases, official links are associated with exactly one article, we may label each official link with the topics of that article. (Throughout the discussion that follows, keep in mind that each article, and thus each official link, may be labeled with multiple topics.)

Fig. 5.3 visualizes the mean CTR (in blue), as well as the number of articles with official links (in gray), by topic. We see that official links relating to LIBRARIES & INFORMATION, SOFTWARE, and INTERNET CULTURE had the highest click-through rates, whereas geographical content, media-related content, and biographies on average saw the lowest engagement.

Controlling for article length and popularity. The length and popularity of a Wikipedia article correlated strongly and negatively with the CTR of the official link contained in its infobox (Fig. 5.4a), possibly because longer articles, by offering more information, reduce the user's need to gather additional information from external links, and because more popular articles are more likely to appear in shallower information-seeking sessions [143]. Since length and popularity also vary by topic, they might act as confounds that could potentially explain an observed variation of CTR by topic.

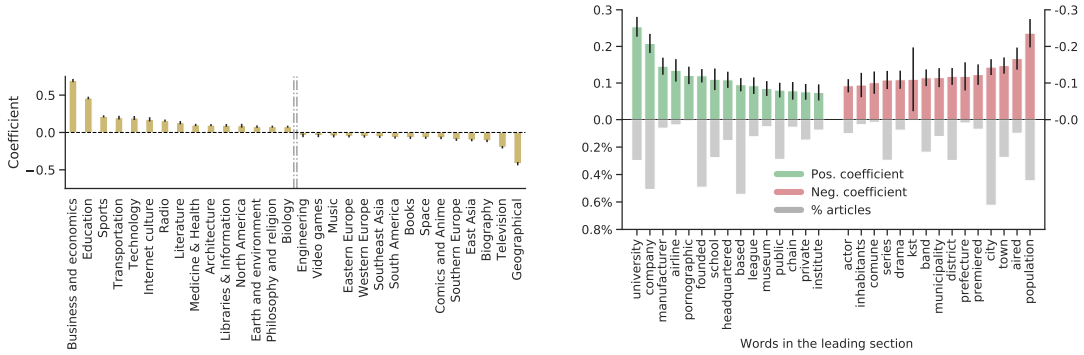
To tease these two confounds apart from the impact of topics alone, we controlled for length and popularity in a matched analysis, as follows. We split the set of articles at the median CTR into high- vs. low-CTR articles, and we split the length and popularity ranges into 1,000 equally sized bins each. We then defined a bipartite graph with edges between articles that fell in different CTR halves, but in the same bins with respect to length and popularity. Using the Euclidean distance in the space defined by the logarithmic length and popularity as edge weights, we found a minimum matching and retained only the 112K matched articles (out of originally 160K). This procedure successfully balanced the dataset.[VI]

In the balanced dataset, we binarized the CTR by splitting at the median and fit a logistic regression to model whether an article belonged to the high- or low-CTR group, with topic indicators as predictors (pseudo $R^2 = 0.20$, $p < 10^{-307}$). (The advantage of performing regression modeling rather than a simple comparison of per-topic average CTRs is that topics are correlated, which is accounted for by the regression model.)

The 15 largest positive and negative coefficients, plotted in Fig. 5.5a, revealed a slightly different ranking than Fig. 5.3, with BUSINESS AND ECONOMICS and EDUCATION emerging as the strongest predictors of a high CTR, whereas GEOGRAPHICAL and TELEVISION remained the strongest predictors of a low CTR.

Top of the CTR ranking. While manually screening the data, we realized that, among the articles with the highest official-link CTR, there was a disproportionate fraction of articles about websites (which are generally classified by ORES under the topic INTERNET CULTURE), and in particular websites related to file sharing and pornography, some with CTRs of 40% or more, e.g., Library Genesis (47%), RARBG (45%), or The Pirate Bay (43%). To determine whether official links of Wikipedia articles about websites dominated the top of the CTR ranking in general, we repeated the above regression analysis with a small modification: instead of predicting the top half vs. the bottom half of the article ranking with respect to CTR, we now predicted the top L articles (an absolute, rather than relative number) vs. the same number of samples from the bottom half, matched on length and popularity. This way, plotting the fitted coefficients for a given topic as a function of L reveals whether the topic is particularly over-represented among the highest-CTR official links (manifested in a sharply decreasing curve). The results, presented in Fig. 5.6, clearly show that INTERNET CULTURE—a

[VI] The standardized mean differences in logarithmic length and logarithmic popularity dropped from 0.7 to 0.00017, and from 0.54 to 0.000005, respectively.



(a) Topic-based analysis (b) Word-based analysis

Figure 5.5: Association of click-through rate of official links with article properties, captured via 15 largest positive and negative coefficients (with 95% CIs) from binary logistic regression models that predict above- vs. below-median CTR, using as predictors (a) article topics or (b) words from lead paragraphs (controlling for article length and popularity). Gray bars in (b): percentage of articles whose lead paragraph contains the word.

topic held by most articles about websites—is indeed particularly over-represented among the articles with the very highest official-link CTR. Similar effects were observed for SOCIETY (a loose mix of articles), SPORTS, SOFTWARE, and ENTERTAINMENT, among others. On the contrary, we observed that GEOGRAPHICAL, BIOGRAPHY, and TELEVISION, among others, were particularly under-represented among the highest-CTR official links.

Fine-grained topical analysis. The topics from the ORES classifier used above are rather broad. In order to obtain more fine-grained insights, we conducted a word- rather than topic-level analysis, where we represented an official link by the words contained in the lead paragraph of the article in whose infobox the link appeared as an official link (via z-score-standardized TF-IDF vectors restricted to the 3,000 most frequent words across all articles with official links). Mirroring the above regression analysis, but now using words rather than topics as predictors (pseudo $R^2 = 0.31$, $p < 10^{-307}$), this analysis revealed words associated with high- vs. low-CTR links. Fig. 5.5b, which shows the 15 words with the largest positive and negative coefficients, confirms our previous findings while adding nuance. We see that EDUCATION specifically marks high-CTR links about universities, schools, institutes, and museums; BUSINESS AND ECONOMICS, about companies, manufacturers, chains, and airlines; and INTERNET CULTURE, about adult websites.

Summary. Taking stock of the findings so far, we reiterate that official links play a key role among Wikipedia's external links, with CTRs far above those of other types of external link. We observed a large amount of variation depending on article topics, with official links associated with articles about websites, software, businesses, education, and sports seeing particularly high engagement.

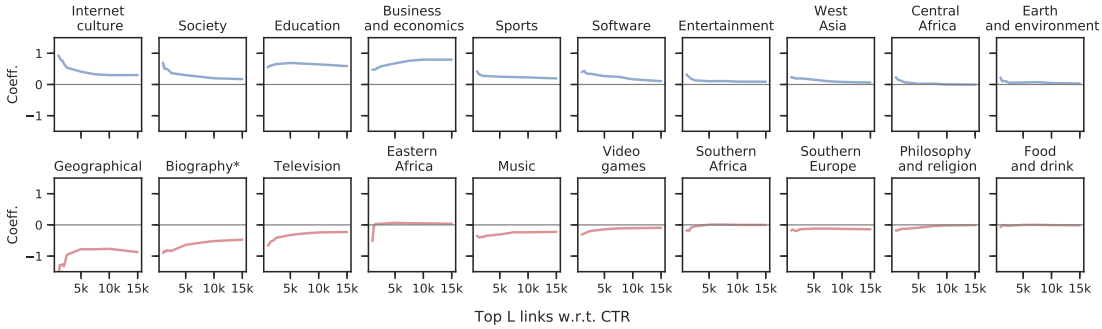


Figure 5.6: Prevalence of topics among most frequently clicked official links. We fitted binary logistic regression models that used article topics as predictors to predict if an article's official link is among the top L highest-CTR links. Plots show regression coefficients for individual topics (predictors) as functions of L (for values of L between 1K and 15K). Topics are sorted by the leftmost values of their curves. Sharply decreasing [increasing] curves correspond to topics that are particularly over-represented [under-represented] among the links with the most extreme CTR. (More details: Sec. 5.3.2, "Top of the CTR ranking".)

5.4 Patterns of engagement with external links

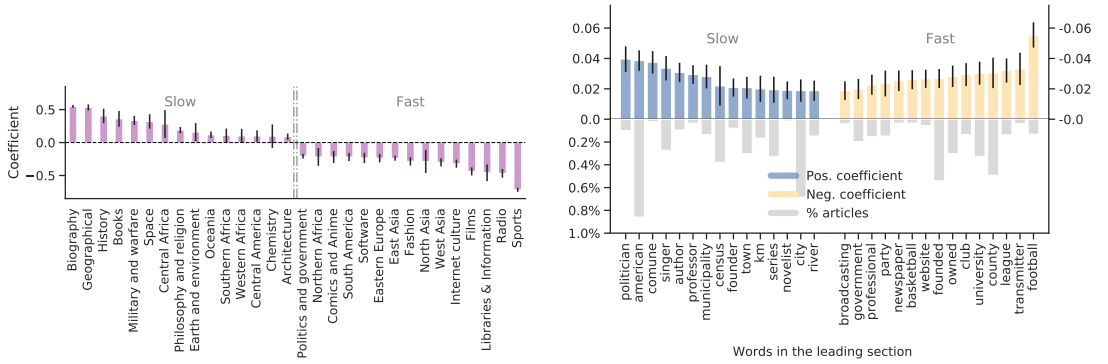
Above, we established which types of external link have a particularly high CTR. Next, we investigate more closely the patterns by which users engage with external links.

5.4.1 Click time

We start by analyzing the click time (cf. Sec. 5.2.5), which captures how long users dwell on an article before leaving it toward an external website via an external link.

Click time statistics are summarized in Table 5.1, and click time distributions are plotted in Fig. 5.2b, for the three types of link: those appearing in infoboxes, article bodies, and references, respectively. (Note that, although we consider only official links in most of this analysis—and indeed in the rest of this section—we nevertheless report this basic statistic for all types of external link beyond official links only.) The global median click time was 32.9 seconds (31.8 seconds for desktop, 34.4 seconds for mobile), with a much lower value for infobox links (18.7 seconds; 20.1 seconds for official links), and larger values for the article-body links (35.4 seconds) and reference links (51.8 seconds). The short click time of infobox links, however, seems to be due to their prominent position within articles: when approximately controlling for position by considering only article-body links in the top 20% of the page, the median click time dropped to 22.2 seconds, only 10% longer than for infobox links.

After this general characterization, we from here on focus on official links in infoboxes. Similar to CTR (Fig. 5.4a), the click time of official links was correlated with article length and popularity (Fig. 5.4b), such that clicks took more time on longer and on less popular articles. When analyzing click time by topic, we thus again controlled for these two factors via matching, as



(a) Topic-based analysis (b) Word-based analysis

Figure 5.7: Association of click time of official links with article properties, captured via 15 largest positive and negative coefficients (with 95% CIs) from linear regression models that predict logarithmic click time, using as predictors (a) article topics or (b) words from lead paragraphs (controlling for article length and popularity). Gray bars in (b): percentage of articles whose lead paragraph contains the word.

they could otherwise confound the analysis. Analogously to the setup of Sec. 5.3.2, we split the set of articles at the median click time into articles with "slow" vs. "fast" clicks on official links, and subsequently found a bipartite matching in order to ensure that the distribution of length and popularity was nearly identical in articles with "slow" vs. "fast" official-link clicks.[VII] We then fitted a linear regression on the resulting balanced dataset in order to model the logarithmic click time as the outcome variable, using (as in Sec. 5.3.2) topic indicators as predictors ($R^2 = 0.090$, $p < 10^{-307}$).

The 15 largest positive and negative coefficients are plotted in Fig. 5.7a, revealing that clicks on official links to entertainment-related websites occurred faster, whereas links to websites on more classic encyclopedic topics, such as biographies, geographical content, history, etc., occurred more slowly.

To gain more granular insights than at the coarse topic level, we again ran an analysis at the word level, parallel to the word-level analysis of Sec. 5.3.2, but this time in the linear regression setup just described ($R^2 = 0.17$, $p < 10^{-307}$). The words most indicative of slow and fast clicks (Fig. 5.7b) mirror the findings from the topic-level analysis (Fig. 5.7a), but we now also see that official links on articles about websites and universities were clicked particularly fast.

5.4.2 Wikipedia as a stepping stone

We just noted that official links on articles about websites typically have short click times. Moreover, in Sec. 5.3.2 we had observed that such links are also strongly over-represented

[VII] The standardized mean differences in logarithmic article length and logarithmic popularity dropped from 0.076 to 0.0001, and from 0.30 to 0.00005, respectively. 138K of the original 160K data points were retained after matching.

among the highest-CTR links. This observations led us to hypothesize that the main interest of users interacting with these links might be to find these links to begin with, rather than to find the content that surrounds the links in their respective Wikipedia articles. In other words, we hypothesized that Wikipedia serves as a "stepping stone" toward external websites, whereupon users barely set foot onto Wikipedia before leaving again towards different content that they actually were intent on finding. To further investigate this hypothesis, we considered the relationship between CTR and click time, visualized in Fig. 5.8, which presents a scatter plot of the coefficients obtained from the previously described regressions for CTR (x-axis; cf. Sec. 5.3.2) and click time (y-axis; cf. Sec. 5.4.1). As the plot shows, topics with a high CTR tended to have low click times (lower right quadrant), notably SPORTS, RADIO, and INTERNET CULTURE (a topic that, as mentioned, primarily tags articles about websites). We take this as another indicator of the existence of a class of articles from which users leave Wikipedia frequently and fast.

If indeed a distinct class of articles on Wikipedia are used heavily as stepping stones, then we should be able to identify many articles that were visited primarily from outside of Wikipedia (e.g., from search engines), rather than from other Wikipedia articles (via internal navigation), and that had a high CTR on the external links they contain. With the goal of finding such articles, we define the *external-referrer frequency (ERF)* of an article as the fraction of all visits made to the article via a click from a referral page external to Wikipedia. Note that external referrals almost exclusively stemmed from search engines: about 70% of them specified a search engine referrer URL in the logs, and about 29% did not specify any referrer URL, but it is suspected that a majority of these visits also come from search [119]. Hence, we included empty referrers in our analysis, although the conclusions were identical when excluding them.

The ERF histogram, plotted in Fig. 5.9a, shows that most articles were primarily visited from outside of Wikipedia, but only few articles had a very high ERF close to 1. Although extreme-ERF articles were in total visited less than medium-ERF articles (Fig. 5.9b), they generated a total number of external clicks comparable to that generated by medium-ERF articles (Fig. 5.9c). The most important piece of evidence, however, comes from Fig. 5.9d, which plots, on the y-axis, the mean CTR for situations where the respective article was reached from an external referrer. The plot shows that the articles that were nearly exclusively reached from external referrers (with an ERF close to 1) are precisely those articles that also had the highest CTR after being reached from an external referrer.[VIII]

Taken together, these facts provide evidence of a class of articles that serve as mere "stepping stones", "revolving doors", or "in-and-outs": users come from elsewhere in order to find a particular link and immediately leave Wikipedia by clicking that link.

But why, then, would users go through Wikipedia in the first place, if all they want is to go to a website linked from Wikipedia? We shall discuss potential reasons for this behavior in Sec. 5.6.

[VIII] This analysis only considered articles with at least 100 visits from external referrers, in order to avoid noise due to division by small numbers.

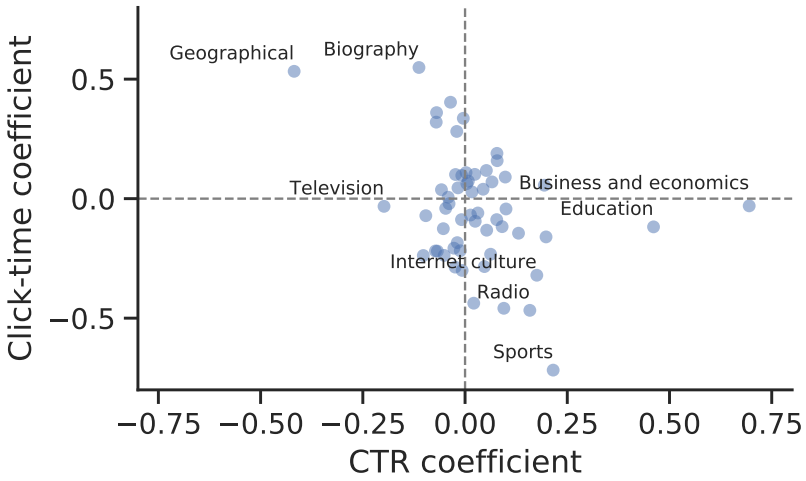


Figure 5.8: Click-through rate (x-axis) vs. click time (y-axis) of official links. Each point represents one topic. CTR and click time of a topic captured in terms of the topic's coefficient in the regressions summarized in Fig. 5.5a and 5.7a, respectively.

5.5 Economic value of external links

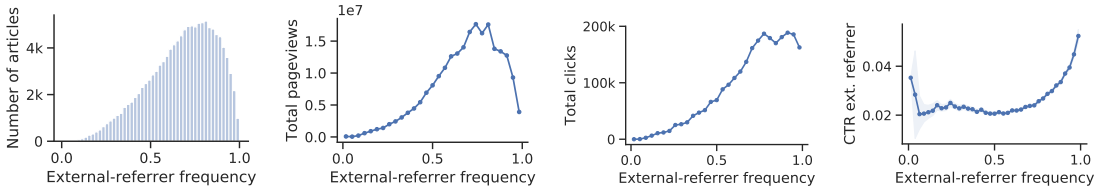
Our final set of analyses aims to estimate the monetary value of the traffic that Wikipedia drives to external websites.

The idea behind our calculations is the following. Search engines generally charge website owners money in exchange for driving clicks to their sites from sponsored search results (e.g., Google Ads), whereas Wikipedia conveys a large volume of traffic to official websites for free. We therefore ask, "How much money would a search engine want from website owners to obtain, via ads, the same number of clicks they obtain from Wikipedia for free?"

While we could be asking, "How much money could Wikipedia earn by charging a fee for external clicks?", we consider this counterfactual scenario too far from reality: Wikipedia is open and free by design, and it functions rather differently from platforms driven mostly by advertising. Moreover, we could not estimate the "price paid" to Wikipedia in this scenario, as we do not know whether website owners would be willing or able to actually pay for any hypothetical click fees. On the contrary, estimating the "price asked"—the cost of online ads—is entirely feasible.

5.5.1 Methodology

Google Ads. To estimate the price of achieving a certain set of URL clicks via online ads, we used the Google Ads API, with "sponsored search" as the ads network. Google Ads is one of the most prolific advertising platforms, and the primary source of revenue for Google's parent company, Alphabet [5]. The Google Ads API allows advertisers to create campaigns for



(a) (b) (c) (d)

Figure 5.9: Quantification of Wikipedia's role as a stepping stone toward external websites. (a) Histogram of *external-referrer frequency (ERF)* of Wikipedia articles, where ERF is defined as the fraction of times the article was visited via a referral page external to Wikipedia. (b) Total number of pageviews of articles within each ERF bin. (c) Total number of official-link clicks of articles within each ERF bin. (d) Official-link CTR upon pageviews with an external referrer (most likely a search engine), with 95% CIs. Articles with an extreme ERF close to 1 are rare (a), but generate a disproportionately large number of official-link clicks (c vs. b), especially when reached from search engines (d).

promoting a URL by placing bids on campaign-related search keywords. The bid expresses the maximum amount that the website owner is willing to pay each time the promoted URL is clicked when shown on the search result page for the respective keyword. When a user searches for a keyword specified by the campaign, an auction system determines which sponsored URL to show among all the candidates competing for the keyword.[IX] When the user clicks a promoted link, the campaign owner pays the auction value. Note that the paid price is not necessarily equal to, but only bounded by, the website owner's bid.

From URLs to keywords. Our intended analysis started from clicks on official links (URLs) observed in the Wikipedia logs and aimed to estimate how much these clicks would cost when obtained via Google Ads instead of Wikipedia. The Google Ads API, on the contrary, requires keywords, not URLs, as input. Thus, in order to leverage Google Ads for our analysis, we had to work our way backwards and determine appropriate keywords that a website owner might use to advertise a given URL. Since the choice of the right keywords is critical to increase a website's discoverability while keeping ad costs down [164], Google Ads offers a tool called *Keyword Planner,* which, given a website URL, generates a set of relevant keywords, alongside information on the historical bidding range and search volume for those keywords. Using the Keyword Planner, we generated 11 keywords for each official link: the title of the corresponding Wikipedia article, plus the top 10 keywords returned by the Keyword Planner. As examples, Table 5.2 summarizes the most relevant keywords generated for two different websites: Coursera (a popular online course platform) and American Airlines.

Cost-per-click (CPC) forecasting. Once the set of keywords and the bids have been set, Google Ads can make a prediction about the cost of the campaign through its forecasting tool. The prediction model uses historical data to simulate the auction system and provides an

[IX] With the "broad search" option, the number of matching searches increases by also considering substrings and permutations of the tokens in campaign keywords.

Table 5.2: Keywords, alongside estimated average cost per click (CPC), for two example websites.

	Coursera	American Airlines
From title	COURSERA	AMERICAN AIRLINES
Keywords recommended by Google Ads Keyword Planner	ONLINE COURSES	AA
	ONLINE COLLEGES	AIRLINE FLIGHTS
	ONLINE CLASSES	AIRLINE TICKETS
	MOOC	AIRLINES
	ONLINE LEARNING	AMERICAN
	FREE ONLINE COURSES	AMERICAN AIRLINES FLIGHTS
	ONLINE DEGREES	CHEAP AIR TICKETS
	OPEN UNIVERSITY COURSES	CHEAP AIRLINE TICKETS
	ONLINE EDUCATION	FLIGHT TICKETS
	ONLINE UNIVERSITIES	US AIRWAYS
Est. avg. CPC	$0.79	$1.10

estimated number of clicks and the average price for every keyword. The tool predicts the campaign's average *cost per click (CPC)* by combining the keyword costs with their expected click-through rates. In practice, the forecasting tool simulates campaigns for specific target countries. We used the top 5 English-speaking countries (U.S., U.K., India, Canada, and Australia; cf. Sec. 5.3.1) as target countries, since they accounted for a large portion (71.6%) of all external-link clicks in the Wikipedia logs studied here.

Estimating the value of official links. Leveraging the above tools, we estimated how much a website owner would need to pay to Google Ads for a single click to their website as follows:

1. Obtain keywords for the website via the Keyword Planner.
2. For each keyword, set the bid to the 80th percentile of the keyword's historical auction price.
3. Estimate the cost (CPC) for one click on the website link by feeding the keywords and their bids to the forecasting tool (using "broad search", cf. footnote IX).

We emphasize that setting a high bid (step 2) does not automatically entail a high CPC. Indeed, as we shall see, the winning price was usually much lower than the bid. A high bid ensures that we are likely to win the simulated auction and that the promoted link is actually displayed to the user, which is required in order to obtain clicks—the event whose cost we are aiming to estimate.

5.5.2 Results

Cost per click (CPC). We applied the above-described procedure to a total of 3,600 official links from Wikipedia infoboxes, obtained by randomly sampling an equal number of articles

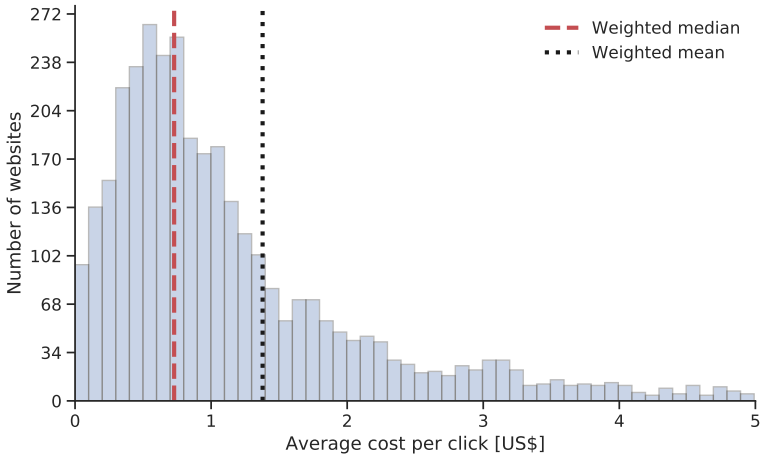


Figure 5.10: Distribution of official-link CPC, estimated via Google Ads API. Vertical lines represent weighted versions of median and mean, respectively, where each link was weighted according to its click volume in Wikipedia logs.

from each of the 57 topics. During the one-month study period, these 3,600 official links were clicked 2.73M times in total.

As mentioned, the bid passed to the forecasting tool is not necessarily the winning price of the simulated auction; it merely caps expenses. In practice, the auction price reached our bid in only 8.9% of cases; on average, the auction price was 58% of the bid.

The CPC distribution is shown in Fig. 5.10. When weighting all links evenly (i.e., macro-averaging), the mean and median CPC is $1.64 and $0.90, respectively. As not all links are equally popular, a more reasonable estimate of the overall CPC may be obtained by weighting each link according to the number of clicks it received (i.e., micro-averaging). The resulting weighted CPC is slightly lower, with a mean and median of $1.37 and $0.73, respectively (vertical lines in Fig. 5.10).

Investigating the weighted mean CPC for individual topics, we found considerable variation, with the highest CPCs for MATHEMATICS, MEDICINE & HEALTH, BOOKS, and ARCHITECTURE, and the lowest CPCs for MUSIC, SPORTS, FASHION, and FILMS (omitting from the list topics that mark geographical regions, such as NORTH AMERICA).

Monthly value of traffic to official websites. Multiplying the weighted CPC with the overall number of 9.8M clicks on official links during the one-month study period, we estimate the total monthly value generated by the traffic from Wikipedia to official websites as **$13.4 million** when using the mean CPC estimate, or as **$7.2 million** when using the median CPC estimate.

Broken down by topic, we obtained the (mean-based) estimates of Fig. 5.11. The topic with the highest total monthly value ($1.9M) is NORTH AMERICA, a macro category assigned to a

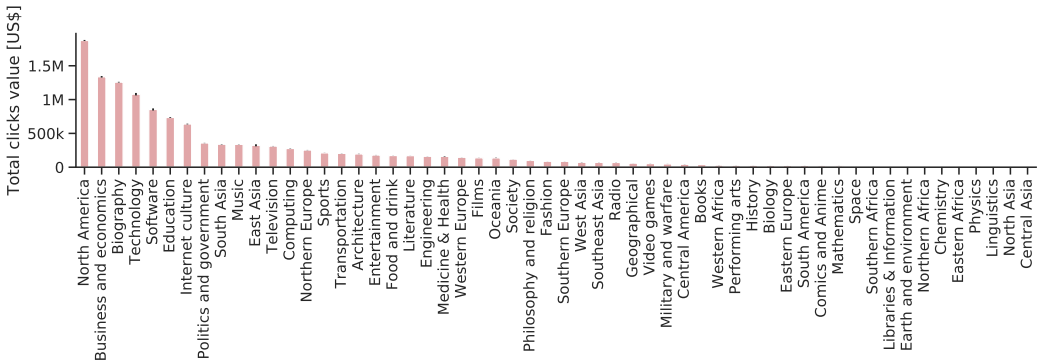


Figure 5.11: Estimated total monthly value of official links in Wikipedia infoboxes by topic, obtained by multiplying the mean cost per click (CPC) of links from the respective topic with the total number of clicks on those links in the Wikipedia logs.

large set of articles, including U.S. companies and people. It is followed by BUSINESS AND ECONOMICS ($1.3M), BIOGRAPHY ($1.3M), TECHNOLOGY ($1.0M), and SOFTWARE ($0.9M).

5.6 Discussion

Wikipedia as a gateway to the Web. While the value of Wikipedia's knowledge is fairly well known, less known was the hidden and significant additional value of Wikipedia as a gateway. Building on top of existing work related to Wikipedia readers' behavior analysis, we have uncovered a new perspective on the role of Wikipedia in the broader Web ecosystem. We offered a description of the value of Wikipedia as a gateway under multiple levels of analysis. Overall, we found that a substantial fraction of Wikipedia readers use the encyclopedia as a gateway to the broader Web: readers engage more and faster with official links in the article's infobox than with links in the article body or in the reference section. We found that Wikipedia's role as a gateway to external content is particularly pronounced when users visit articles about websites, software, businesses, education, and sports, among others, where the click-through rate of official links is the highest.

Wikipedia as a stepping stone. We found an inverse relation between the time it takes to click on an external link and its average click-through rate, showing clusters of topics, such as SPORTS and INTERNET CULTURE where engagement with links was high and fast, or conversely, where engagement with link was low and slow, such as BIOGRAPHY and GEOGRAPHICAL. We also observed that articles that were visited particularly frequently from external referrers (mostly search engines) also had a particularly high probability of an official-link click after being reached from external referrers. These results indicate that a certain distinct set of Wikipedia articles is leveraged by users in the spirit of "stepping stones" or "revolving doors", which are reached nearly exclusively from external referrers (mostly search) and from which the user leaves Wikipedia immediately again by clicking an official external website link. This begs the question: If users visit Wikipedia from a search engine result page only to leave it

immediately towards a third website, why would they not simply use the search engine to locate the third website to begin with?

We conjecture that the reason is that the search engine cannot fulfill the user's information need in such situations, whereas Wikipedia can. When manually screening the data (focusing on popular articles with at least 30K pageviews during the one-month study period), we found that, among the articles with the highest CTR, there was a disproportionate fraction of file-sharing (5 of the top 6) and pornographic (5 of the top 15) websites. Such search results are frequently censored or down-ranked by search engines, depending on the search engine's corporate policy as well as legislation in the user's country. Indeed, manually searching Google for the names of the 15 articles with the highest CTR (and more than 30K monthly pageviews) from two locations (U.S. and Germany) revealed that 5 file-sharing websites and 1 pornographic website were not listed by Google among the top 10 search results in at least one of the two locations. (Additionally, two controversial websites were not online anymore at the time of research, about 18 months after data collection.) While these findings remain small-scale and anecdotal, they suggest that Wikipedia fills a functional gap, as a workaround for content suppressed by search engines (sometimes for valid reasons, e.g., when the linked material is copyrighted or illegal).

Wikipedia's role as a transitory stepping stone towards external content can have important implications for Web user studies. For example, researchers working on disinformation diffusion might want to take into account the function of Wikipedia as a short yet crucial stopover in Web users' information-seeking journeys.

The economic value of traffic generated by Wikipedia. Finally, we set out to estimate the monetary value of Wikipedia as a gateway to the broader Web. The infoboxes contained in English Wikipedia articles collectively list over half a million official-website links, which were clicked 9.8M times during our one-month study period. These clicks were generated by Wikipedia for free, amidst a Web ecosystem that is majorly powered by paid ads. We asked, "If the respective website owners wanted to achieve the same number of clicks via sponsored search results, what would be the price?" We estimate that achieving the 9.8M monthly clicks on official links would cost a total of $7–13 million using Google Ads. Extrapolating to 12 months, the yearly cost would amount to $84–156 million. This is a remarkably high number, considering that the annual revenue of the Wikimedia Foundation, the non-for-profit organization that operates Wikipedia and its sister projects, stands at around $110 million,[X] coming entirely from donations and voluntary contributions. We also emphasize that the estimated economic value of $84–156 million pertains to English Wikipedia only, whereas Wikimedia's annual revenue of $110 million needs to support all Wikimedia projects across languages.

We showed that, when buying clicks from Google Ads instead of obtaining them from Wikipedia for free, the types of businesses that would have to pay the most would be North American

[X] https://wikimediafoundation.org/about/financial-reports

companies, as well as software and technology businesses. While the narrative about tech companies' donations to Wikipedia has often been around their massive usage of the free encyclopedic content for products and algorithms [41, 209], these findings might provide yet another perspective on how these companies benefit from the hard work of hundreds of thousands of volunteer editors.

More broadly, our work expands the small body of literature on measuring the value of Wikipedia to the Web [120, 192]. While previous work focused on the value of content *production,* for example estimating that Wikipedia generates $1.7 million of Reddit and Stack Overflow's revenue, based on the amount of Wikipedia-linked posts on those platforms [192], we focused here on the value of Wikipedia *traffic*. We provided, for the first time, an estimation of the monetary value offered—for free—by Wikipedia to the broader Web ecosystem by means of link navigation.

Limitations and future work. This study should be considered in the light of its limitations. Most notably, it was constrained to data collected during one month from English Wikipedia only, and as such provides a limited view of readers' general behavior. Future work should replicate the study for different time periods and language versions in order to paint a more complete and inclusive picture of Wikipedia readers' engagement with external links.

Besides broadening the scope, future work should also go deeper by more closely investigating why certain types of official link see particularly high or low CTRs (e.g., the CTR of links to geographical and biographical content was particularly low; cf. Fig. 5.5a). Also, considering that official links related to BUSINESS AND ECONOMICS saw the highest CTRs, it will be interesting to analyze which businesses benefit most from the free traffic provided by Wikipedia.

Whereas our investigation of the volume (RQ1) and patterns (RQ2) of engagement with external links on Wikipedia was primarily measurement-based, our estimation of the economic value of Wikipedia as a gateway to the Web (RQ3) was more speculative. On the one hand, we operated under the assumption that our methodology for obtaining costs per click via the Google Ads API is sound and provides accurate estimates, despite the fact that we relied on keyword suggestions of unknown quality from the Google Ads Keyword Planner and on uncertain auction simulations from the Google Ads forecasting tool. On the other hand, and more fundamentally, estimating the economic value of the Web traffic generated by Wikipedia necessarily requires arguing about a hypothetical, counterfactual ("what if") situation, in our case, "What if website owners were to pay for the same number of clicks via Google Ads instead of obtaining them from Wikipedia for free?"

Although similar reasoning has been applied to estimate the value of images from Wikimedia Commons [47], it remains open how realistic that "what if" is: as a matter of fact, Wikipedia *is* providing those clicks for free, so why would website owners ever decide to pay for them instead? As one concrete example, one could imagine a situation where a website owner would want to increase traffic to their site, in which case our estimates indicate how much it would

cost them to double the traffic they already receive from Wikipedia for free. Alternatively, one could imagine scenarios where Wikipedia were to be blocked by censorship or ceded to exist entirely, in which case our estimated economic value of traffic from Wikipedia would correspond to the loss on behalf of website owners due to the lack of that traffic. Finally, and more boldly, one could imagine a setting where Wikipedia decided to introduce a fee for clicks on official website links, in which case our estimate would upper-bound the amount of extra revenue Wikipedia could possibly earn from such a fee. Although the latter setting is highly unrealistic, we consider it a useful thought experiment that can help emphasize Wikipedia's importance as a provider of free traffic.

As a final remark, we argue that all notions of economic value are fundamentally counterfactual at heart, as they always consider "what if" scenarios ("If A were to give X to B, how much money would B give to A in return?"), which is also the reason why business valuations of companies are routinely criticized as absurdly off [150].

Conclusion. This chapter characterized the interaction of the Wikipedia readers with the external links on the platform. To conclude, we hope this work will offer ideas and methods to those interested in exploring Wikipedia's role in the larger Web ecosystem in more depth, and that it will help quantify the true value of the largest encyclopedic knowledge repository on the Web.

Expanding Wikipedia Toolbox Part II

6 Crosslingual Topic Modeling with WikiPDA

6.1 Introduction

With 53 million articles written in 299 languages, Wikipedia is the largest encyclopedia in history. To leverage and analyze individual language editions, researchers have successfully used topic models [171]. Topic models [20] are unsupervised machine learning techniques that represent documents as low-dimensional vectors whose dimensions are interpretable as topics. The goal of this chapter is to move beyond individual language editions and develop a topic model that works for all language editions jointly. Our method, *Wikipedia-based Polyglot Dirichlet Allocation (WikiPDA),* learns to represent articles written in any language in terms of language-independent, interpretable semantic topics. This way, articles that cannot be directly compared in terms of the words they contain (as the words are from different vocabularies) can nevertheless be compared in terms of their topics.

Such a model is tremendously useful in practice. With close to a billion daily page views, Wikipedia plays an important role in everyday life, and it is equally important as a dataset and object of study for researchers across domains: Google Scholar returns about 2 million publications for the query "Wikipedia". Although English is but one of 299 language editions, it is currently by far the most studied by researchers, to an extent that goes well beyond what can be justified by size alone.[I] The scarcity of easy-to-use crosslingual topic models contributes to this skew, affecting even those studies that go beyond English; e.g., in previous work[108], researches compared the usage of 14 language editions via survey data and browsing logs, from quantifying differences in users' topic interests across languages.

Although each language on its own can be readily handled via standard topic models, which are based on bags of words and thus straightforward to apply to any language with minimal preprocessing, such models are insufficient for comparing content across languages because

[I] For instance, with 6 million articles, the English edition is 5 times as large as the Vietnamese one, whereas Google Scholar returns over 300 times as many results for "en.wikipedia.org" (387K) as for "vi.wikipedia.org" (1,250).

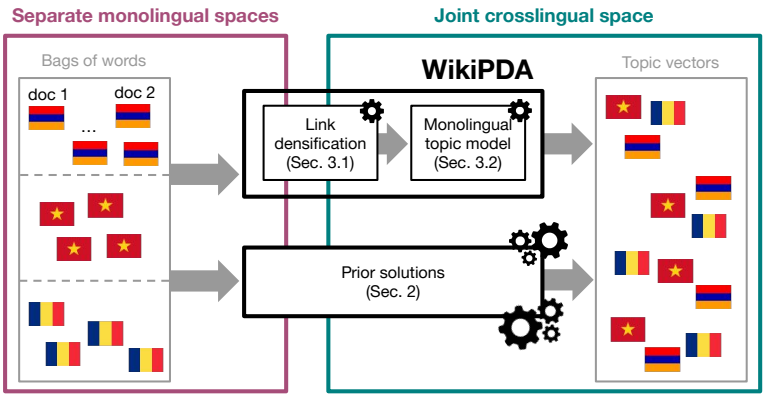


Figure 6.1: Overview of Wikipedia-based Polyglot Dirichlet Allocation (WikiPDA). Whereas prior crosslingual topic models typically learned to directly map from separate monolingual bag-of-word spaces to a joint crosslingual topic-vector space, WikiPDA proceeds in two steps: In the first step, language-specific bags of words are mapped to language-independent bags of out-links, using the fact that each language-specific Wikipedia article (and thus each out-link) corresponds to one language-independent concept in the Wikidata knowledge base. In the second step, the language-independent bags of out-links are fed to a vetted, powerful monolingual topic model such as LDA.

in general the topics learned for one language do not have clearly corresponding topics in the other languages.

Prior solutions. To address this problem, researchers have extended monolingual topic models [70, 195] by mapping documents from separate monolingual spaces into a joint crosslingual topic space. This paradigm has been proposed under various names (e.g., crosslingual, multilingual, polylingual, bilingual, or correspondence topic models;), but the basic idea is identical—namely to enhance the model by enforcing crosslingual alignment at the level of words or documents. For instance, in document-alignment models, a topic is modeled not as a single word distribution, but as a set of word distributions—one per language—, and different language versions of the same document are constrained to the same mix of topics during training. As Wikipedia articles are aligned across languages via the Wikidata knowledge base,[II] Wikipedia has served as a prominent training dataset for models based on document alignment.

Another set of methods does not use document-aligned corpora, but word alignments from bilingual dictionaries, modeling topics as distributions over crosslingual equivalence classes of words [69, 82, 216]. A method introduced by Boyd-Graber *et al.* [22] require neither an aligned corpus nor a dictionary.

[II] For instance, the Wikidata concept Q44 corresponds to the English article BEER, French BIÈRE, Finnish OLU, etc.

Document-alignment-based methods make effective use of large aligned corpora but are hampered by requiring that aligned documents be about the same topics, which is frequently not the case in practice and eliminates an important use case of crosslingual topic models *ab ovo,* namely quantifying how an identical concept is described in different languages (cf. Sec. 6.4.1). Word-alignment-based methods, on the contrary, do not suffer from this shortcoming but are hampered by the scarcity of multilingual dictionaries beyond two languages.

WikiPDA marries the best of both worlds by leveraging the *document alignment* provided by Wikidata in the spirit of *word alignment* methods: representing articles as bags of links, rather than bags of words, may be seen as inducing a common vocabulary spanning 299 languages, without unnaturally forcing corresponding articles in different languages to have identical topic distributions.

Proposed solution: WikiPDA.[III] We leverage Wikipedia's crosslingual article alignment from a different angle, by recognizing that Wikipedia articles are not just plain text, but laced with links to other articles. An article's set of outgoing links (*"bag of links"*) is a concise summary of the article's key content. Crucially, since each linked article is itself associated with a language-independent Wikidata concept, bags of links immediately give rise to a crosslingual input representation "for free". Starting from this representation, WikiPDA works in two steps, by first densifying bags of links using matrix completion and then training a standard monolingual topic model. Whereas in previous methods, translating from mono- to crosslingual space constitutes the core computation, in WikiPDA it constitutes a mere preprocessing step (Fig. 6.1). Put differently, whereas prior work has harnessed Wikipedia's crosslinguality to *increase model complexity,* we leverage it to *decrease data complexity.* This way, WikiPDA can leverage, as its core computation, standard monolingual topic models such as latent Dirichlet allocation (LDA) [21], which have been vetted in practice, come with implementations on all platforms, and scale to massive datasets.

A human evaluation shows that WikiPDA topics extracted jointly from 28 language editions of Wikipedia are more coherent than those from monolingual text-based LDA, thus offering crosslinguality at no cost (Sec. 6.3). We demonstrate WikiPDA's practical utility in two applications (Sec. 6.4): a topic comparison of Wikipedia across 28 languages, and crosslingual supervised document classification. Finally, we show WikiPDA's ability to operate in the challenging zero-shot setting (Sec. 6.4.3), where a model is applied to new languages without any fine-tuning, a powerful capacity not shared by prior crosslingual topic models, to the best of our knowledge.

With WikiPDA, researchers possess a new tool for studying the content of all of Wikipedia's 299 language editions in a unified framework, thus better reflecting Wikipedia's linguistic diversity.

[III] Code, library, models, data: https://github.com/epfl-dlab/WikiPDA

Table 6.1: Statistics of the 28 Wikipedia language editions.

		Articles (in thousands)			**Links (in millions)†**			**Disambiguation evaluation† (Sec. 6.3.1)**				
			Num. w/	% of			Dens.	Ambig.	Accuracy for num. candidates in interval			
	Language	Num.	≥10 links*	total†	Sparse	Densified	ratio‡	anchors	$[1,\infty]$	$[2,\infty]$**	$[1,10]$	$[2,10]$**
ar	Arabic	987	507	2%	14.6	49.1	3.4	48%	0.86	0.70 (0.24)	0.88	0.75 (0.34)
ca	Catalan	611	468	2%	16.2	71.4	4.4	46%	0.88	0.74 (0.23)	0.90	0.78 (0.33)
cs	Czech	410	357	1%	14.3	52.0	3.6	48%	0.86	0.71 (0.26)	0.88	0.74 (0.34)
de	German	2043	1851	7%	70.1	304.1	4.3	44%	0.86	0.68 (0.22)	0.88	0.73 (0.33)
el	Greek	164	117	<1%	4.1	17.1	4.1	46%	0.87	0.71 (0.26)	0.89	0.75 (0.33)
en	English	5571	4511	18%	206.9	594.3	2.9	39%	0.88	0.68 (0.18)	0.90	0.74 (0.33)
es	Spanish	1461	1332	5%	56.8	179.9	3.2	37%	0.86	0.63 (0.19)	0.89	0.70 (0.33)
fa	Persian	674	341	1%	8.7	34.8	4.0	49%	0.86	0.71 (0.22)	0.89	0.78 (0.33)
fi	Finnish	451	348	1%	10.4	31.8	3.1	54%	0.86	0.75 (0.25)	0.89	0.80 (0.35)
fr	French	2013	1684	7%	81.5	247.3	3.0	42%	0.85	0.64 (0.19)	0.89	0.73 (0.32)
he	Hebrew	239	229	1%	12.3	52.7	4.3	47%	0.87	0.72 (0.25)	0.89	0.76 (0.33)
id	Indonesian	495	345	1%	8.9	33.8	3.8	52%	0.85	0.71 (0.23)	0.87	0.76 (0.33)
it	Italian	1458	1093	4%	54.0	193.6	3.6	45%	0.84	0.64 (0.20)	0.88	0.73 (0.33)
ja	Japanese	1097	1030	4%	60.4	80.8	1.3	44%	0.84	0.64 (0.22)	0.87	0.71 (0.33)
ko	Korean	418	307	1%	12.3	28.8	2.3	42%	0.84	0.63 (0.25)	0.89	0.74 (0.35)
nl	Dutch	1889	958	4%	33.2	116.2	3.5	55%	0.84	0.71 (0.22)	0.87	0.76 (0.33)
pl	Polish	1289	986	4%	35.3	105.4	3.0	43%	0.88	0.72 (0.21)	0.90	0.76 (0.31)
pt	Portuguese	964	742	3%	28.0	102.2	3.6	40%	0.86	0.64 (0.22)	0.88	0.70 (0.33)
ro	Romanian	378	240	1%	7.4	29.2	3.9	46%	0.90	0.78 (0.24)	0.90	0.79 (0.32)
ru	Russian	1406	1143	4%	47.7	172.9	3.6	37%	0.87	0.66 (0.21)	0.90	0.72 (0.31)
sq	Albanian	71	19	<1%	0.7	2.6	3.8	53%	0.89	0.79 (0.33)	0.89	0.79 (0.36)
sr	Serbian	579	424	2%	9.7	46.0	4.7	50%	0.87	0.75 (0.27)	0.89	0.79 (0.33)
sv	Swedish	3453	3178	12%	59.3	118.8	2.0	60%	0.91	0.85 (0.26)	0.92	0.87 (0.36)
tr	Turkish	319	227	1%	7.0	20.8	3.0	48%	0.86	0.71 (0.24)	0.89	0.78 (0.34)
uk	Ukrainian	905	742	3%	23.0	80.6	3.5	45%	0.88	0.73 (0.25)	0.90	0.78 (0.33)
vi	Vietnamese	1218	543	2%	15.0	71.8	4.8	59%	0.83	0.72 (0.30)	0.86	0.75 (0.39)
war	Waray	1251	1142	4%	15.6	29.8	1.9	99%	0.46	0.46 (0.37)	0.46	0.46 (0.37)
zh	Chinese	1028	576	2%	23.4	31.8	1.4	54%	0.85	0.72 (0.25)	0.88	0.77 (0.34)
	Average	1173	908		33.4	103.5	3.3	48%	0.85	0.69 (0.24)	0.87	0.74 (0.33)
	Total	32844	25437	100%	936.8	2900.0						

*Links counted after link densification. †Considering only articles with ≥10 links after densification. **Random baseline in parentheses.
‡Densification ratio = Densified/Sparse.

6.2 Method

Existing crosslingual topic models take a monolithic approach, mapping directly from monolingual bags of words to crosslingual topic distributions. WikiPDA, on the contrary, procedes sequentially (Fig. 6.1), first mapping monolingual bags of words to crosslingual bags of links (Sec. 6.2.1) and then mapping crosslingual bags of links to crosslingual topic distributions (Sec. 6.2.2). In what follows, we describe these two stages in turn.

6.2.1 Link densification

Wikipedia's *Manual of Style*[IV] asks authors to add links that aid navigation and understanding. Key concepts are thus linked to their articles, allowing us to use bags of links, in lieu of bags of words, as concise article summaries. Crucially, bag-of-links elements—articles—are associated with language-independent Wikidata concepts, so in principle, the crosslingual article representations to be fed to the downstream topic model may be obtained simply by extracting links from articles.

In practice, however, human editors frequently fail to add all relevant links [204], and they are explicitly instructed to add links parsimoniously (e.g., by linking only the first mention of every concept). For topic modeling, such human-centric factors are of no concern; rather, we prefer semantically complete article summaries with information about the frequency

[IV] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking

of constituent concepts. Hence, the first phase of WikiPDA is link densification, where we link as many plain-text phrases as possible to the corresponding Wikidata concepts (e.g., all occurrences of "beer", "Beer", "beers", etc., in the article about INDIA PALE ALE should be linked to Wikidata concept Q44).

The difficulty arises from ambiguous phrases (e.g., in some contexts, "Beer" should be linked to BEER, DEVON [Q682112], an English village). Disambiguating phrases to the correct Wikidata concept is the so-called "wikification" task, with several existing solutions [122, 124, 135, 204], any of which could be plugged in. Given the scale of our setting, we opted for a lightweight approach based on matrix completion: First, given a Wikipedia language edition, build the adjacency matrix A of the hyperlink graph, where both rows and columns represent Wikidata concepts, and entry a_{ij} is non-zero (details in Sec. 6.2.3) iff the article about concept i contains a link to that about concept j. Then, decompose $A \approx UV^\top$ using alternating least squares [94], such that both U and V are of low rank r. The rows of U are latent representations of articles when serving as link sources, and the rows of V, when serving as link targets, optimized such that, for existing links (i, j), we have $a_{ij} \approx u_i v_j^\top$ (where single subscripts are row indices). For non-existing links (i, j), the dot product $u_i v_j^\top$ provides a score that captures how well the new link (i, j) would be in line with the existing link structure.

Thus, the scores $u_i v_j^\top$ can be used to disambiguate the plain-text phrases p in article i: consider as the set C_p of candidate targets for p all articles j for which p occurs as an anchor at least once in the respective language edition of Wikipedia, and select the candidate with the largest score, i.e., link the phrase p in article i to $\arg\max_{j \in C_p} u_i v_j^\top$.

In principle, a decomposition computed for one language can be used to disambiguate links in any other language. In this case, however, we computed a separate decomposition for each language, in order to be able to model language-specific patterns.

6.2.2 Topic modeling

The bags of links resulting from link densification can be fed to any monolingual topic model based on bags of words, by using a vocabulary consisting of Wikidata concepts rather words, and by using as the document corpus the union of all Wikipedia articles pooled across all languages considered. Concretely, we use LDA as the topic model, but any other model based on bags of words, such as PLSA [79], would be compatible with our method. As usual, the number K of topics is set manually by the user.

6.2.3 Implementation and corpus details

Link densification. We considered as potential anchors for new links all 1- to 4-grams, with preference given to longer n-grams (e.g., "India pale ale" as a whole is linked to INDIA PALE ALE, rather than "India" to INDIA, and "pale ale" to PALE ALE). We did not consider n-grams

whose occurrences are linked with a probability below the threshold of 6.5% [124] (e.g., "a", "the", etc.), since, like stop words, they usually do not represent semantically relevant content.

Decompositions of the adjacency matrix A used rank $r = 150$. Before the decomposition, A's entries were weighted in the spirit of inverse document frequency, giving more weight to links occurring in few articles: if i links to j, we set $a_{ij} = -\log(d_j/N)$, and $a_{ij} = 0$ otherwise, where d_j is the number of articles that link to j, and N is the number of articles in the respective Wikipedia [123].

Topic modeling. Since LDA may perform poorly with short documents [179], we removed articles with fewer than 10 links after densification. Further, we ignored concepts appearing as links in fewer than 500 articles across all languages.

Corpus. We worked with 28 language editions of Wikipedia (details in Table 6.1), in their snapshots of 20 February 2020. We work only with articles from namespace 0 (the main namespace). Links and anchor texts were extracted from wiki markup. Redirects were resolved. After all preprocessing, the corpus encompassed 25M documents across all 28 languages, with a vocabulary of 437K unique Wikidata concepts.

Code and model availability. We release code and pre-trained models (cf. footnote III) for a wide range of K. For $K = 40$ and 100, topics were manually labeled with names. On a single machine (48 cores, 250 GB RAM), the full pipeline for all 28 languages with fixed hyperparameters ran in under 24 hours. As the code uses Apache Spark, parallelizing over many machines is straightforward and would further reduce the runtime.

6.3 Evaluation

Next, we evaluate the two stages of our pipeline, link densification (Sec. 6.3.1) and topic modeling (Sec. 6.3.2).

6.3.1 Link densification

Densification increased the number of links substantially, by a factor of 3.3, to an effective 114 links per article, on average over all 28 languages (details in Table 6.1).

The large fraction of ambiguous anchors (48%) underlines the importance of disambiguation. To evaluate disambiguation accuracy, we masked 5% of the entries of the adjacency matrix A before decomposing it (cf. Sec. 6.2.1). Each masked link is associated with a potentially ambiguous anchor text p. Given p, we generated all candidate targets j and ranked them by their score $u_i v_j^\top$. Disambiguation accuracy is then defined as the fraction of masked matrix entries for which the top-ranked candidate was correct. It is summarized, for all 28 languages,

in the 4 rightmost columns of Table 6.1, where column "$[l,u]$" contains the accuracy for anchors with at least l and at most u candidates.

The column "$[1,\infty]$" shows the overall accuracy for all anchors (85% on average over all 28 languages). Since this column includes trivial, unambiguous anchors, the column "$[2,\infty]$" is more interesting. Although lower, these numbers are still satisfactorily high (69% on average), particularly when compared to the random baseline (24%).

Manual error analysis revealed that anchors with a large number of candidates tend to be inherently hard to disambiguate even for humans (e.g., "self-titled album" has 712 candidates). Hence, preferring precision over recall, our implementation ignores phrases with more than 10 candidates, obtaining an average accuracy of 87% for the remaining anchors ("$[1,10]$").

Moreover, we found that, even when the exact link target was not identified, often a semantically close target was predicted. Since this suffices for our purpose—which is not to disambiguate all links perfectly, but rather to create better document representations for the downstream topic modeling step—the accuracy of 87% should be considered an underestimate of the true, effective utility.

The quality of disambiguated links is confirmed by the superior performance of densified, compared to raw, bags of links, as discussed next.

6.3.2 Topic modeling

We evaluated 4 model classes, each trained on a different corpus:

1. **WikiPDA, dense links, 28 languages:** full model as described in Sec. 6.2.
2. **WikiPDA, sparse links, 28 languages:** the same, but without link densification.
3. **WikiPDA, dense links, English:** trained on English only, rather than on all 28 languages.
4. **Text-based LDA, English:** bag-of-words LDA trained on English text (not links).

For each model class, we trained and evaluated models for 10 values of K, ranging from 20 to 200. In the following, "model" refers to an instance of a model class trained for a specific K.

Comparing model classes 1 and 2 lets us determine the benefits of link densification; comparing model classes 1 and 3, whether including more languages hurts performance, as has been found to be the case in other crosslingual settings [88]; and comparing model classes 3 and 4, whether using bags of links rather than bags of words makes a difference.

Methodology: intruder detection. The evaluation of topic models is challenging. Traditionally, it has been based on automatic scores such as perplexity, capturing how "surprising" documents from a held-out corpus are, given the training set. Unfortunately, perplexity does

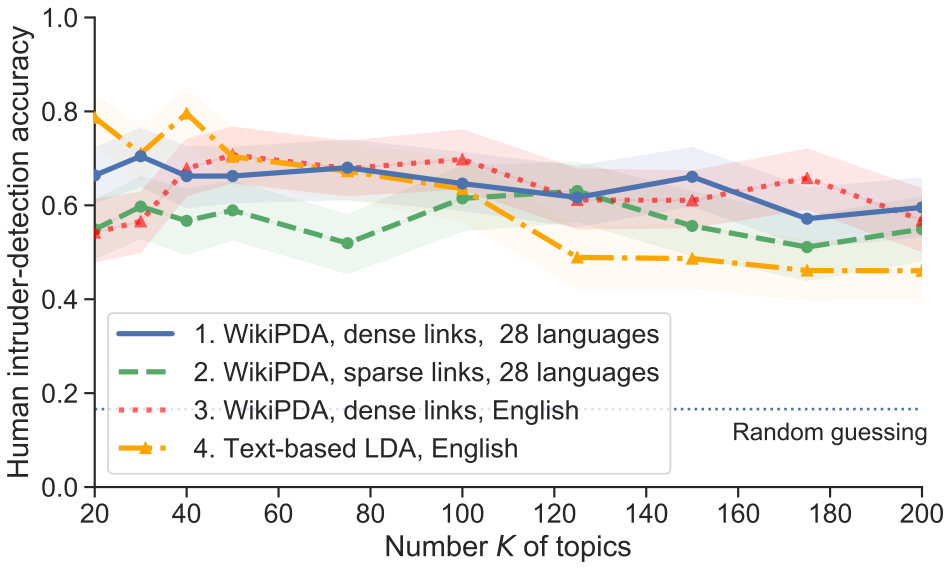


Figure 6.2: Evaluation of topic models. Topic coherence measured in terms of human intruder-detection accuracy (higher is better), with 95% confidence intervals.

not necessarily correlate with human judgment, and in some cases an inverse relation has even been reported [26]. Since we are interested in interpretable models, we quantified the utility of topics in a human, rather than automatic, evaluation, using the *word intruder* framework proposed by [26]. Given a model to evaluate, we randomly selected $n = 20$ of the $K \geq 20$ topics and extracted the top 5 Wikidata concepts per topic. Then we selected an *intruder concept* for each topic: a concept that ranked low for that topic, but high for at least one other topic (in particular, the concept with the largest rank difference was selected). A shuffled list of the 6 concepts (described by their English names) was shown to a human evaluator, who was asked to spot the intruder. The more coherent a topic, the easier it is to spot the intruder, so human accuracy serves as a measure of topic coherence.

Crowdsourcing setup. For each model, human accuracy was estimated based on $12n = 240$ workers' guesses obtained from 12 independent rounds of the above procedure on Amazon Mechanical Turk. Workers were shown a page with instructions and a batch of 17 intruder-detection tasks: 16 regular tasks (4 per model class), each with a different K, and one control task with an obvious answer to assess worker reliability (all workers were found to be reliable). Workers were encouraged to search online in case they did not know the meaning of a concept. To not reveal a pattern, we used each model and each intruder at most once per batch.

Results. Fig. 6.2 shows that, with the full WikiPDA model (model class 1), human intruder-detection accuracy was 60–70%, depending on K, far above random guessing (16.7%).

Comparing model classes 1 and 2, we see that the dense WikiPDA model yielded results consistently above the sparse model (by up to 15 percentage points), showing the utility of link densification.

Comparing model classes 1 and 3, we find that the dense WikiPDA model for 28 languages performed indistinguishably from the dense model for English only; i.e., adding more languages did not make the topics less coherent. This outcome is noteworthy, since on other crosslingual tasks (e.g., document retrieval), performance on a fixed testing language decreased when adding languages to the training set [88].

Comparing model classes 3 and 4 (both English only) shows that, whereas the performance of text-based LDA degrades with growing K, WikiPDA is more stable. While text-based LDA is slightly better for small $K \leq 50$, WikiPDA prevails for $K \geq 75$. This suggests that the link-based models are more customizable to use cases where the problem requires a specific K.

Note that the text-based LDA model is not language-independent and thus not truly a competitor with crosslinguality in mind. Rather, it should *a priori* be considered a strong ceiling: text-based LDA is the de-facto standard for analyzing the content of Wikipedia articles in monolingual settings [108, 171]. Thus, by surpassing the topic coherence of text-based models, WikiPDA offers crosslinguality "for free".

6.4 Applications

WikiPDA enables a wide range of applications, some of which we spotlight next. We emphasize that the purpose of this section is not to take a deep dive into specific directions, but rather to exemplify the utility of WikiPDA as a general tool for analyzing Wikipedia across languages.

6.4.1 Comparing Wikipedia across languages

Wikipedia's different language editions are maintained by independent volunteer communities, each with their own cultural background and with potentially diverging interests. Understanding the differences in content coverage across Wikipedia language editions constitutes a major topic for researchers in multiple domains [15, 24, 71, 72, 103, 118, 160], and WikiPDA will be a useful tool for their endeavors.

Topic bias. Using WikiPDA, we studied the topic bias of 28 language editions (cf. Table 6.1). To obtain a first impression, we pooled the topic vectors from all languages and reduced their dimensionality from $K = 40$ down to 2 dimensions via t-SNE [113]. A heat-map visualization of the reduced, 2-dimensional topic vectors for each language is presented in Fig. 6.3. The visual heterogeneity of the heat maps is a stark indication of the topic heterogeneity of the various language editions.

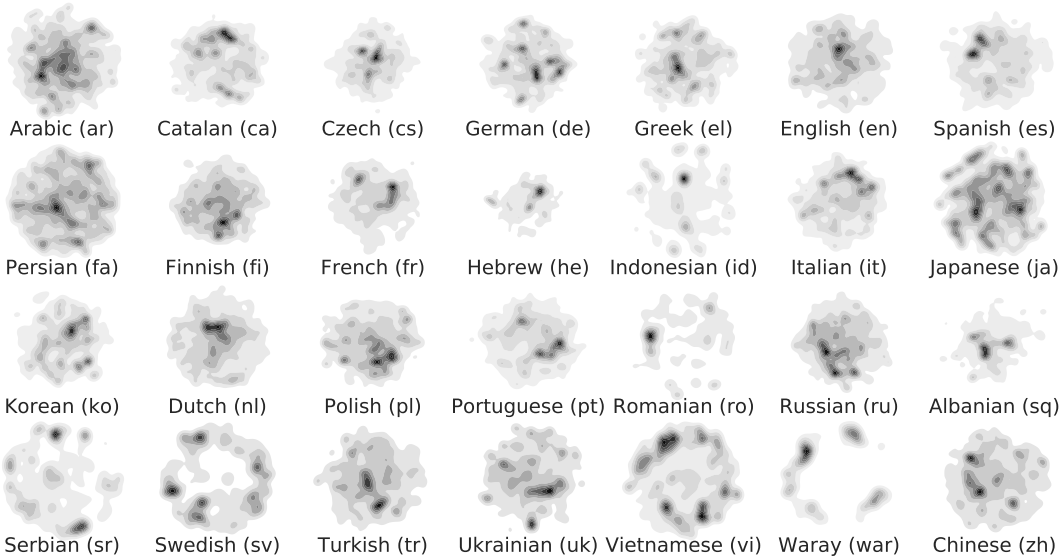


Figure 6.3: Heat-map visualization of the topic distributions of 28 Wikipedia language editions, obtained by reducing the dimensionality of the topic vectors from $K = 40$ down to 2 dimensions via t-SNE [113]. The visual heterogeneity of the heat maps highlights the topic heterogeneity of the various language editions.

Whereas reducing the dimension from 40 to 2 is advantageous for visual inspection, the 2 dimensions resulting from t-SNE—unlike the $K = 40$ original dimensions—do not have a clear interpretation anymore. In order to study differences across languages with respect to individual topics, we therefore conducted a regression analysis. For each language L, we randomly sampled 20K articles as positive examples and 20K/27 = 740 from each of the 27 other languages as negative examples, applied an 80/20 train/test split, and trained a one-vs.-all logistic regression classifier to predict whether an article is from language L, given the article's topic distribution.[V] On average the 28 classifiers achieved an area under the ROC curve (AUC) of 78% for $K = 40$, or 84% for $K = 200$, significantly above the random baseline of 50%, indicating major differences across language editions. Inspecting the fitted coefficients for $K = 40$, depicted in Fig. 6.4, revealed the specificities of individual languages. First and foremost, country- or region-specific topics appeared among the most discriminative topics. Additionally, several more surprising associations emerged: e.g., COMICS is the topic most indicative of English and Dutch, and it is most counter-indicative of Ukrainian and Catalan; GEOPOLITICS is prominently featured in Hebrew; ICE HOCKEY and TENNIS, in Korean; etc.

Note that some topics appear to conflate multiple concepts that one would rather expect to emerge as distinct topics of their own (e.g., the topic POLAND, VIETNAM). Such conflation is expected for the small number of $K = 40$ topics used to produce Fig. 6.4. Although such a small K is convenient as it allows for the manual inspection and naming of topics, it provides too

[V] Since in logistic regression the log odds are modeled as a linear function of the predictors, we also expressed the probabilistic predictors (namely, the K topic probabilities) as log odds, so input and output use the same "units".

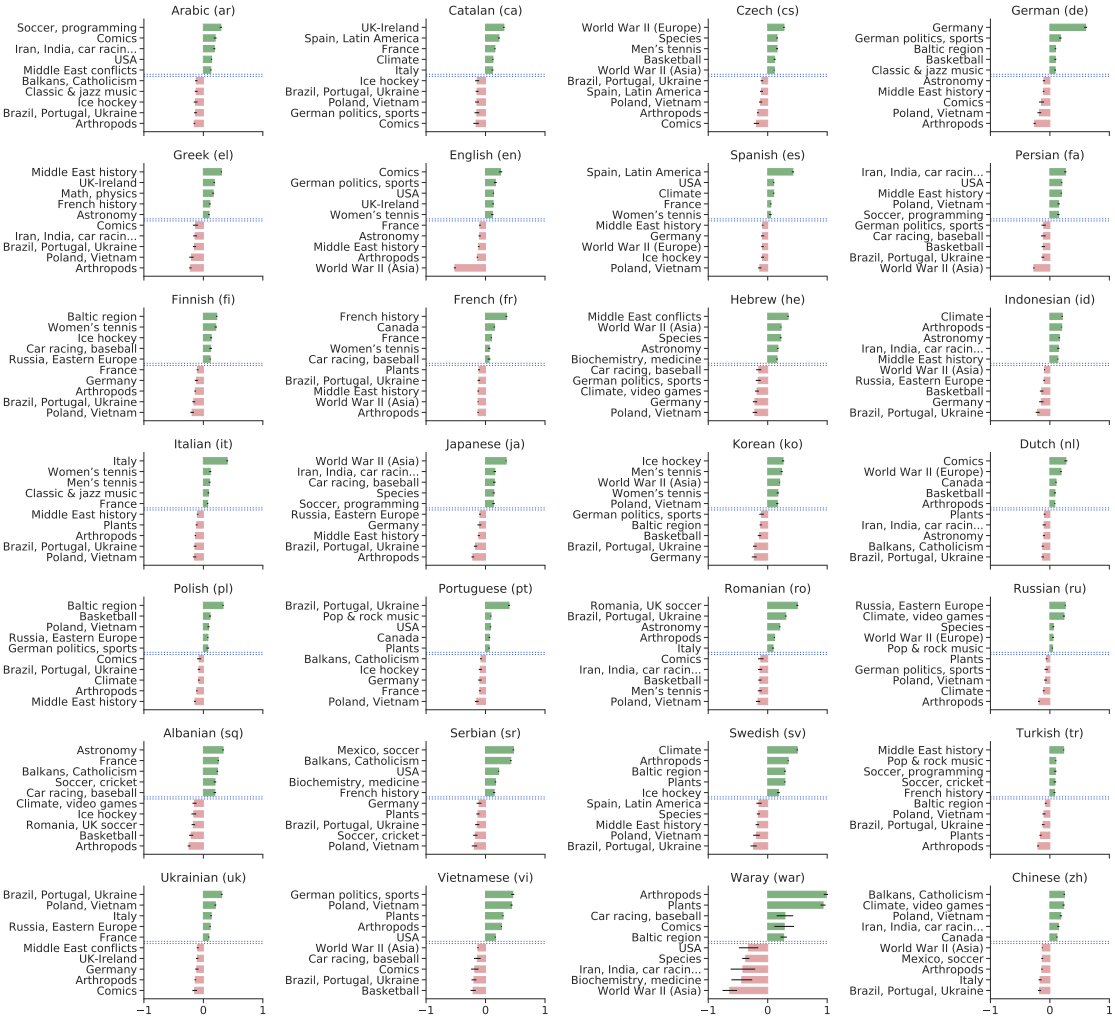
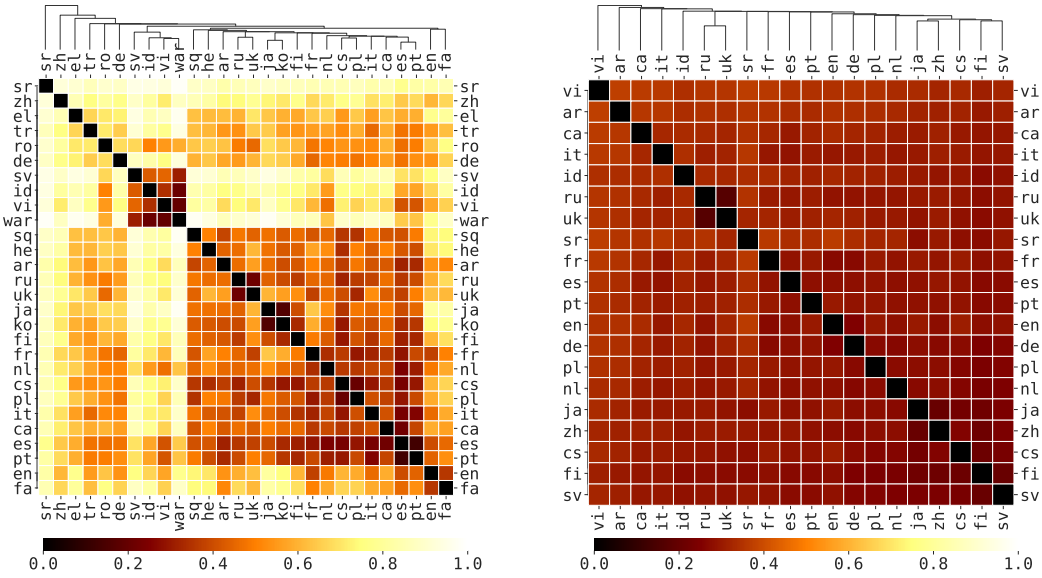


Figure 6.4: Topic bias of 28 Wikipedia language editions. For each language L, a logistic regression was trained to predict if an article was written in language L, using the article's distribution over WikiPDA topics (labeled manually with names) as predictors. Most predictive positive and negative coefficients are shown, with 95% confidence intervals.



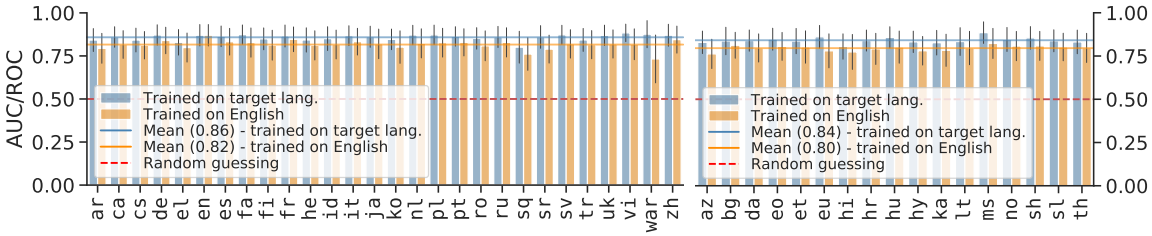
(a) Click-through rate (b) Click time

Figure 6.5: Cosine distance between Wikipedia language editions. **(a)** 28 languages, each represented via average topic vector of all articles. **(b)** 20 top languages, considering only the 16K articles included in all 20 languages.

little capacity to capture all of Wikipedia's diversity. Accordingly, we found that the conflation effect is reduced as K grows. Also note that conflation does not lead to catch-all "garbage" topics, but to topics consisting of distinct subtopics (e.g., POLAND and VIETNAM). The fact that subtopics are dissimilar is in fact desirable: we found that, when allowing for a larger model capacity (larger K), they tend to become distinct, non-redundant topics of their own.

Distance between language editions. Next, we computed pairwise distances for all language editions via the cosine distance between the languages' mean topic vectors. As we do not rely on topic labels here, we use the larger $K = 200$. Fig. 6.5a shows the distance matrix. Rows and columns correspond to languages, sorted by performing agglomerative clustering based on the distances and listing the leaves of the dendrogram in left-to-right order, such that the most similar pairs cluster along the diagonal of the distance matrix. Clear topic similarities (darker colors) emerge for languages of countries with historical or geographical ties, including Japanese/Korean, Russian/Ukrainian, Czech/Polish, and Portuguese/Spanish. Waray (spoken in the Philippines) clusters with Indonesian, Vietnamese, and—more surprisingly—Swedish, a language that, linguistically speaking, could not be more distant. Investigating the reasons, we found that Swedish and Waray are among the three Wikipedias (the third being Cebuano) in which Lsjbot was active, a bot that created 80–99% of the articles in those languages. Fig. 6.4 suggests that Lsjbot created particularly many biology-related articles (with ARTHROPODS and PLANTS appearing as prominent topics in both Swedish and Waray), a finding not even mentioned on the Wikipedia page about Lsjbot itself.[VI] Also, it seems that the bot, which

[VI] https://en.wikipedia.org/w/index.php?title=Lsjbot&oldid=949492392



(a) 28 languages seen during WikiPDA training (b) 17 zero-shot languages

Figure 6.6: Performance on supervised topic classification, using unsupervised WikiPDA topics as features. For each language L, two models were evaluated: trained on L (blue); trained on English (orange). Error bars: standard deviation over 64 binary classification tasks (one per supervised topic label). Similarity of blue and orange shows that classifier works on languages not seen during supervised classifier training. Similarity between (a) and (b) shows that classifier and WikiPDA models work on languages not seen during unsupervised WikiPDA training.

was created by a Swede, gave Waray Wikipedia a Swedish bias, as indicated by Waray's large coefficient for the topic BALTIC REGION in Fig. 6.4. These nuggets exemplify how WikiPDA enables the cross-cultural study of Wikipedia.

The above-noted differences may be due to different language editions containing articles about different concepts. An equally interesting question asks to what extent the languages differ in how they discuss identical concepts. To quantify this, we found the 16K articles in the intersection of the 20 largest language editions and computed, for each language pair and each common article, the cosine distance of the two languages' topic vectors for the article. Averaging the 16K distances yields Fig. 6.5b, which paints a more uniform picture than Fig. 6.5a, with no important differences remaining between languages. Note, however, that Russian/Ukrainian, Finnish/Swedish, and Chinese/Japanese cover the same concepts in particularly similar ways.

6.4.2 Supervised topic classification

WikiPDA is an unsupervised technique. It discovers whatever topics are best suited for summarizing the data. Sometimes, however, researchers may want to exert more control by fixing a set of topics ahead of time and classifying documents into those in a supervised manner. For instance, with the ORES library,[VII] Wikimedia provides a supervised classifier for categorizing English articles into a manually constructed taxonomy of 64 topics, based on features derived from the articles' English text [9]. We explored if WikiPDA topic vectors can be used as features instead, giving rise to a language-independent model, whereas the ORES model is language-specific. More generally, we establish whether WikiPDA is effective as an unsupervised pre-training step for supervised downstream tasks.

[VII] https://ores.wikimedia.org

Setup. For training and testing (following an 80/20 split), we used the same dataset that had been used to train the ORES topic model,[VIII] consisting of 5.1M English articles, each labeled with 64 binary labels that specify for each of the 64 topic classes defined by ORES whether the article belongs to the class. The labels were obtained by the creators of ORES by manually mapping WikiProjects[IX]—and thus implicitly their constituent articles—to the 64 high-level topics. Note that, although the ORES training data consists of English articles only, Wikidata's crosslingual alignment allows us to propagate labels to other languages for any article that has an English version. For this practical reason, our evaluation focuses on such articles.

Each article can belong to multiple classes, so we trained an independent binary logistic regression classifier per class, on a balanced training set where negative examples were sampled evenly from the 63 other classes. Performance was found to increase with K, so we used $K = 200$. For each language L, we performed two evaluations: first, with a model trained on articles from L (after transferring labels from the English dataset via the alignment given by Wikidata) and second, with a model trained on English.

Results. In Fig. 6.6a, we show two AUC values (macro-averages over the 64 classes) for each language L: one when testing the classifier trained on L itself (blue); the other, when testing the classifier trained on English (orange). Performance is high across all languages, with an average AUC of 86% for the language-specific classifiers. The single, fixed classifier trained on English performed only slightly worse when evaluated on the other languages, with an average AUC of 82%.

In English, the easiest class was VIDEO GAMES (AUC 97%); the hardest class was BIOGRAPHIES (AUC 74%).

Note that the primary goal of these experiments was not to achieve maximum classification performance by all means. Indeed, exploratory results showed that simply switching from logistic regression to gradient-boosted trees immediately boosted the AUC by 2–3 percentage points, and we would expect methods that leverage state-of-the-art pre-trained language models such as BERT [40] to perform even better. We emphasize that our main goal is to discover interpretable topics in an unsupervised fashion and that the supervised application presented here is primarily intended as an additional evaluation to assess the usefulness of the learned representations, not so much as an end goal in itself.

In this light, the main take-aways of this section are twofold: (1) WikiPDA's unsupervised topics can be readily translated to a different set of manually defined topics, which demonstrates their utility as a general low-dimensional representation that captures the topic essence of a document. (2) Due to the crosslingual nature of WikiPDA topics, a supervised model trained on one language (here: English) can be transferred to any other language not seen during supervised training, achieving high performance without any fine-tuning.

[VIII] Code: https://github.com/wikimedia/drafttopic
[IX] https://en.wikipedia.org/wiki/Wikipedia:WikiProject

In our final set of experiments, described in the next section, we push the language-transfer paradigm even further, by moving to a setting where the target language was absent not only during training of the supervised classifier, but also during unsupervised training of the WikiPDA topics that the supervised classifier uses as features.

6.4.3 Zero-shot language transfer

The bags of links by which WikiPDA represents input documents are composed of language-independent Wikidata concepts (one per out-link). No matter in what language an article is written, its bag of links can be immediately compared to the bags of links extracted from any other language. This way, a WikiPDA model trained on a certain set of languages can be used to infer topics for articles from any new language not seen during WikiPDA training. In other words, WikiPDA inherently enables *zero-shot language transfer*. This capability is particularly convenient for low-resource languages, where the available data might not suffice to learn meaningful topics, and it sets WikiPDA apart from all the previously proposed crosslingual topic models, which need to be retrained whenever a new language is added.

To showcase WikiPDA's zero-shot capability, we used the model trained on the 28 languages of Table 6.1 to infer topics for all articles in 17 more languages (cf. Fig. 6.6b) and repeated the supervised topic classification experiments (Sec. 6.4.2) for these languages. As in Sec. 6.4.2, we evaluated two supervised topic classifiers for each language: one trained on the respective language, the other trained on English. Note that in neither case were the 17 new languages included in topic model training; rather, the topic vectors that served as input to the supervised classifier were inferred based on a WikiPDA model trained only on the 28 old languages. Despite this, the mean AUC on the 17 new languages (Fig. 6.6b) nearly reached that on the 28 old ones, both for the language-specific classifiers (84% for the new languages, vs. 86% for the old languages) and for the English classifier (80% vs. 82%).

Finally, to further validate the applicability of WikiPDA topics in the zero-shot setting, we repeated the analysis of Fig. 6.4, fitting logistic regression models to predict the language of an article given its topic vector. Classification performance was as high on the 17 new languages as on the 28 languages seen during topic model training (mean AUC 79% for $K = 40$; 84% for $K = 200$), indicating that the topic vectors capture the peculiarities of the 17 new languages well, even though the languages were not seen during topic model training.

6.5 Discussion

We presented Wikipedia-based Polyglot Dirichlet Allocation (WikiPDA), a novel crosslingual topic model for Wikipedia, founded on the fact that Wikipedia articles are laced with language-independent links to other articles. Our human evaluation showed that the topics learned from 28 languages are as coherent as those learned from English alone, and more coherent than those from text-based LDA on English, a noteworthy finding, given that other

crosslingual tasks have suffered by adding languages to the training set [88]. We demonstrated WikiPDA's practical utility in two example applications and highlighted its capability for zero-shot language transfer.

The key insight underpinning WikiPDA is that, when taken as bags of links, Wikipedia articles are crosslingual from the get-go, leading to two big advantages, interpretability and scalability:

Interpretability. By invoking a probabilistic topic model such as LDA as a subroutine, WikiPDA inherits all advantages of that model, including the interpretability of topics as distributions over terms (in our case, Wikidata concepts) and of documents as distributions over topics. Even more important, as WikiPDA's vocabulary consists of Wikidata concepts, which have names in all languages, bags of links and learned topics (distributions over the vocabulary), can be interpreted even without understanding the corpus languages. WikiPDA therefore confers distinct advantages, compared to black-box models such as those trained via deep learning, including recurrent neural networks and transformers. Although such models might perform better at specific downstream tasks, they could not be considered valid alternatives if the user expects to obtain interpretable document representations. On the contrary, it is precisely such users whom WikiPDA is intended to serve.

Scalability. In bag-of-links space, the corpus can be treated as monolingual, such that standard topic models apply, for which highly efficient algorithms exist; e.g., online algorithms for LDA can handle massive amounts of text [78] and have been implemented for high-performance machine learning libraries (e.g., Vowpal Wabbit) and massively parallel big data platforms (e.g., Apache Spark). Scaling WikiPDA to all of Wikipedia's 299 languages is thus fully within reach, whereas previous methods have usually been deployed on small sets of languages only. That said, given WikiPDA's zero-shot capability (Sec. 6.4.3), training on all 299 languages may not even be necessary, as a model trained on a few languages can be immediately applied to unseen languages "for free".

On the use of machine translation. Instead of making documents language-agnostic by mapping them to bags of links, one might be tempted to pursue an alternative approach where all documents would first be machine-translated to a pivot language. While feasible in principle, we consider such an approach to run counter to our design principles. First, translating entire Wikipedia editions would be a costly endeavor that many researchers cannot afford (indeed, much of the literature on crosslingual document representations starts from the very desideratum to circumvent machine translation), whereas extracting and denisfying links from all Wikipedia articles is efficiently feasible. Second, the Wikipedia editions that could benefit most from crosslingual topic models in order to discover knowledge gaps and content skewness are the low-resource languages, for many of which no machine translation models exist to begin with; e.g., we are not aware of translation models for the Waray language, although it has the 11th-largest Wikipedia edition and features prominently in our analysis.

Limitations. Finally, we discuss two potential concerns: language imbalance and link sparsity. First, Wikipedia's language editions vary considerably in size, so the learned topics are dominated by larger languages. Whether this is desirable or not depends on the specific use case. Future work should explore the effects of upweighting smaller languages, e.g., by downsampling large languages, upsampling small languages, or incorporating weights into LDA's objective function. Alternatively, one could aggregate articles at the concept level, e.g., by unioning the bags of links of the articles about the same Wikidata concept in different languages.

Second, compared to text-based models, our link-based model works with sparser inputs, even after link densification. While advantageous computationally, this raises two questions: (1) whether the set of Wikidata concepts is rich enough to capture all semantic facets of an input document; and (2) whether WikiPDA can handle very "short" documents, i.e., articles with very few outgoing links. Regarding the first question, we emphasize that, with over 32M entities, Wikidata is large enough to cover most vocabulary entries. In this light, moving from words to Wikidata entities may be seen as additionally offering some common NLP preprocessing steps for free: the removal of stop words and rare n-grams, plus lemmatization. Regarding the second question, we trained and tested the supervised topic classification model (Sec. 6.4.2) for English again, this time only on articles with fewer than 10 links (19% of all articles). The model still performed well (AUC 85%), only 2 percentage points lower than when using all articles, including those with many links, indicating that WikiPDA is not hampered in important ways by articles with few links.

Going beyond Wikipedia. In the version presented here, WikiPDA is specifically geared toward the analysis of Wikipedia articles as input documents, since the matrix-completion-based link densification method requires at least a few pre-existing Wikipedia links in each input document. Pushing further, it will be interesting to extend WikiPDA to work on documents without any Wikipedia links whatsoever. We believe this is well within reach, by exploiting WikiPDA's modularity: any method that annotates documents with links to a knowledge base can be plugged in instead of matrix completion, including sophisticated entity linking methods for linking mentions in plain-text documents (i.e., without pre-existing links to Wikidata) to a knowledge base [92, 169]. This would widen the scope of applicability to many document types, as long as the documents contain entity mentions that can be linked to the knowledge base. Moreover, past work [203] has shown that plain-text-based entity linkers can be fruitfully combined with matrix completion as used in this chapter (as postprocessing), and we would expect similar outcomes for WikiPDA.

Such extensions would allow us to answer many important questions outside of the realm of Wikipedia. For instance, given a corpus of school textbooks in many languages, we may ask: how do different countries' curricula treat the same subjects (e.g., World War II)? Given a corpus of COVID-related news from all over the world, we may ask: on what aspects of the pandemic have news outlets in different countries focused over time? Given a corpus of search engine results for queries about climate change in different languages, we may ask: is

there a biased view on the issue based on user language? In all of these examples, we require a common, interpretable set of topics across languages, which monolingual topic models cannot provide. As a way forward, off-the-shelf entity linkers can be used as preprocessing to annotate the corpus with links to Wikipedia articles; WikiPDA can take over from there. During training, topics can be learned either on a Wikipedia corpus, or on the corpus of downstream study itself, if preferred.

That said, as laid out in the introduction, Wikipedia on its own constitutes such an important use case that, even without expanding the scope of WikiPDA beyond Wikipedia articles, it still provides a highly valuable tool for a large research community.

Conclusion. To conclude, WikiPDA offers researchers a practical tool for studying the content of all of Wikipedia's 299 language editions in a unified framework, thus better reflecting Wikipedia's real diversity.

7 WikiHist.html: English Wikipedia's Full Revision History in HTML Format

7.1 Introduction

Wikipedia constitutes a dataset of primary importance for researchers across numerous subfields of the computational and social sciences, such as social network analysis, artificial intelligence, linguistics, natural language processing, social psychology, education, anthropology, political science, human–computer interaction, and cognitive science. Among other reasons, this is due to Wikipedia's size, its rich encyclopedic content, its collaborative, self-organizing community of volunteers, and its free availability.

Anyone can edit articles on Wikipedia, and every edit results in a new, distinct revision being stored in the respective article's history. All historical revisions remain accessible via the article's *View history* tab.

Wikitext and HTML. Wikipedia is implemented as an instance of MediaWiki,[I] a content management system written in PHP, built around a backend database that stores all information. The content of articles is written and stored in a markup language called *wikitext* (also known as *wiki markup* or *wikicode*).[II] When an article is requested from Wikipedia's servers by a client, such as a Web browser or the Wikipedia mobile app, MediaWiki translates the article's wikitext source into HTML code that can be displayed by the client. The process of translating wikitext to HTML is referred to as *parsing*. An example is given below, in Fig. 7.1.

Wikitext: '''Niue''' ({{lang-niu|Niuē}}) is an [[island country]].

HTML: Niue (Niuean: Niuē) is an island country.

Figure 7.1: Example of wikitext parsed to HTML.

[I] https://www.mediawiki.org
[II] https://en.wikipedia.org/wiki/Help:Wikitext

Wikitext provides concise constructs for formatting text (e.g., as bold, cf. yellow span in the example of Fig. 7.1), inserting hyperlinks (cf. blue span), tables, lists, images, etc.

Templates and modules. One of the most powerful features of wikitext is the ability to define and invoke so-called *templates*. Templates are macros that are defined once (as wikitext snippets in wiki pages of their own), and when an article that invokes a template is parsed to HTML, the template is expanded, which can result in complex portions of HTML being inserted in the output. For instance, the template *lang-niu,* which can be used to mark text in the Niuean language, is defined in the Wikipedia page TEMPLATE:LANG-NIU, and an example of its usage is marked by the red span in the example of Fig. 7.1. Among many other things, the infoboxes appearing on the top right of many articles are also produced by templates. Another kind of wikitext macro is called *module*. Modules are used in a way similar to templates, but are defined by code in the Lua programming language, rather than wikitext.

Researchers' need for HTML. The presence of templates and modules means that the HTML version of a Wikipedia article typically contains more, oftentimes substantially more, information than the original wikitext source from which the HTML output was produced. For certain kinds of study, this may be acceptable; e.g., when researchers of natural language processing use Wikipedia to train language models, all they need is a large representative text corpus, no matter whether it corresponds to Wikipedia as seen by readers. On the contrary, researchers who study the very question how Wikipedia is consumed by readers cannot rely on wikitext alone. Studying wikitext instead of HTML would be to study something that regular users never saw.

Unfortunately, the official Wikipedia dumps provided by the Wikimedia Foundation contain wikitext only, which has profound implications for the research community: researchers working with the official dumps study a representation of Wikipedia that differs from what is seen by readers. To study what is actually seen by readers, one must study the HTML that is served by Wikipedia. And to study what was seen by readers in the past, one must study the HTML corresponding to historical revisions. Consequently, it is common among researchers of Wikipedia [46, 108, 171] to produce the HTML versions of Wikipedia articles by passing wikitext from the official dumps to the Wikipedia REST API,[III] which offers an endpoint for wikitext-to-HTML parsing.

Challenges. This practice faces two main challenges:

1. Processing time: Parsing even a single snapshot of full English Wikipedia from wikitext to HTML via the Wikipedia API takes about 5 days at maximum speed. Parsing the full history of all revisions (which would, e.g., be required for studying the evolution of Wikipedia) is beyond reach using this approach.

[III] https://en.wikipedia.org/w/api.php

2. Accuracy: MediaWiki (the basis of the Wikipedia API) does not allow for generating the exact HTML of historical article revisions, as it always uses the latest versions of all templates and modules, rather than the versions that were in place in the past. If a template was modified (which happens frequently) between the time of an article revision and the time the API is invoked, the resulting HTML will be different from what readers actually saw.

Given these difficulties, it is not surprising that the research community has frequently requested an HTML version of Wikipedia's dumps from the Wikimedia Foundation.[IV]

Dataset release: WikiHist.html. With the WikiHist.html dataset introduced in this chapter, we address this longstanding need and surmount the two aforementioned hurdles by releasing the complete revision history of English Wikipedia in HTML format. We tackle the challenge of scale (challenge 1 above) by devising a highly optimized, parallel data processing pipeline that leverages locally installed MediaWiki instances, rather than the remote Wikipedia API, to parse nearly 1 TB (bzip2-compressed) of historical wikitext, yielding about 7 TB (gzip-compressed) of HTML.

We also solve the issue of inconsistent templates and modules (challenge 2 above) by amending the default MediaWiki implementation with custom code that uses templates and modules in the exact versions that were active at the time of the article revisions in which they were invoked. This way, we approximate what an article looked like at any given time more closely than what is possible even with the official Wikipedia API.

In addition to the data, we release a set of tools for facilitating bulk-downloading of the data and retrieving revisions for specific articles.

Download location. Both data and code can be accessed via https://doi.org/10.5281/zenodo.3605388.

Chapter structure. In the remainder of this chapter, we first describe the WikiHist.html dataset (Sec. 7.2) and then sketch the system we implemented for producing the data (Sec. 7.3). Next, we provide strong empirical reasons for using WikiHist.html instead of raw wikitext (Sec. 7.4), by showing that over 50% of all links among Wikipedia articles are not present in wikitext but appear only when wikitext is parsed to HTML, and that these HTML-only links play an important role for user navigation, with click frequencies that are on average as high as those of links that also appear in wikitext before parsing to HTML.

[IV] See, e.g., https://phabricator.wikimedia.org/T182351.

7.2 Dataset description

The WikiHist.html dataset comprises three parts: the bulk of the data consists of English Wikipedia's full revision history parsed to HTML (Sec. 7.2.1), which is complemented by two tables that can aid researchers in their analyses, namely a table of the creation dates of all articles (Sec. 7.2.2) and a table that allows for resolving redirects for any point in time (Sec. 7.2.3). All three parts were generated from English Wikipedia's revision history in wikitext format in the version of 1 March 2019. For reproducibility, we archive a copy of the wikitext input[V] alongside the HTML output.

7.2.1 HTML revision history

The main part of the dataset comprises the HTML content of 580M revisions of 5.8M articles generated from the full English Wikipedia history spanning 18 years from 1 January 2001 to 1 March 2019. Boilerplate content such as page headers, footers, and navigation sidebars are not included in the HTML. The dataset is 7 TB in size (gzip-compressed).

Directory structure. The wikitext revision history that we parsed to HTML consists of 558 bzip2-compressed XML files, with naming pattern enwiki-20190301-pages-meta-history$1.xml-p$2p$3.bz2, where $1 ranges from 1 to 27, and p$2p$3 indicates that the file contains revisions for pages with ids between $2 and $3. Our dataset mirrors this structure and contains one directory per original XML file, with the same name. Each directory contains a collection of gzip-compressed JSON files, each containing 1,000 HTML article revisions. Since each original XML file contains on average 1.1M article revisions, there are around 1,100 JSON files in each of the 558 directories.

File format. Each row in the gzipped JSON files represents one article revision. Rows are sorted by page id, and revisions of the same page are sorted by revision id. As in the original wikitext dump, each article revision is stored in full, not merely as a diff from the previous revision. In order to make WikiHist.html a standalone dataset, we include all revision information from the original wikitext dump, the only difference being that we replace the revision's wikitext content with its parsed HTML version (and that we store the data in JSON rather than XML).

The schema therefore mirrors that of the original wikitext XML dumps[VI], but for completeness we also summarize it in Table 7.1a.

Hyperlinks. In live Wikipedia, hyperlinks between articles appear either as blue or as red. Blue links point to articles that already exist (e.g., /wiki/Niue), whereas red links indicate that the target article does not exist yet (e.g., /w/index.php?title=Brdlbrmpft&action=edit&redlink=1). This distinction is not made in the wikitext source, where all links appear in identical format

[V] Downloaded from https://dumps.wikimedia.org/enwiki/.
[VI] https://www.mediawiki.org/w/index.php?title=Help:Export&oldid=3495724

(e.g., [[Niue]], [[Brdlbrmpft]]), but only when the respective article is requested by a client and parsed to HTML. As the existence of articles changes with time, we decided to not distinguish between blue and red links in the raw data and render all links as red by default. In order to enable researchers to determine, for a specific point in time, whether a link appeared as blue or red and what the hyperlink network looked like at that time, we also provide the two complementary datasets described next.

7.2.2 Page creation times

The lookup file page_creation_times.json.gz (schema in Table 7.1b) specifies the creation time of each English Wikipedia page. To determine if a link to a target article A was blue or red at time t (cf. Sec. 7.2.1), it suffices to look up A in this file. If A was created after time t or if it does not appear in the file, the link was red at time t; otherwise it was blue.

7.2.3 Redirect history

Wikipedia contains numerous redirects, i.e., pages without any content of their own whose sole purpose is to forward traffic to a synonymous page. For instance, NIUE ISLAND redirects to NIUE. Link occurrences in the wikitext dumps, as well as our derived HTML dumps, do not specify whether they point to a proper article or to a redirect. Rather, redirects need to be explicitly resolved by researchers themselves, a step that is complicated by the fact that redirect targets may change over time. Since redirect resolution is crucial for analyzing Wikipedia's hyperlink network, we facilitate this step by also releasing the full redirect history as a supplementary dataset: the file redirect_history.json.gz (schema in Table 7.1c) specifies all revisions corresponding to redirects, as well as the target page to which the respective page redirected at the time of the revision.

7.2.4 Limitation: deleted pages, templates, modules

Wikipedia's wikitext dump contains all historical revisions of all pages that still existed at the time the dump was created. It does not, however, contain any information on pages that were deleted before the dump was created. In other words, when a page is deleted, its entire history is purged. Therefore, since WikiHist.html is derived from a wikitext dump, deleted pages are not included in WikiHist.html either.

When using WikiHist.html to reconstruct a past state of Wikipedia, this can lead to subtle inaccuracies. For instance, it follows that the rule of Sec. 7.2.2 for deciding whether a link was blue or red at time t will incorrectly tag a link (u, v) as red if v existed at time t but was deleted before 1 March 2019 (the date of the wikitext dump that we used). Although such inconsistencies are exceedingly rare in practice, researchers using WikiHist.html should be aware of them.

Since MediaWiki handles templates and Lua modules (together referred to as *macros* in the remainder of this section) the same way it treats articles (they are normal wiki pages, marked only by a prefix *Template:* or *Module:*), deleted macros are not available in the revision history either. It follows that a deleted macro cannot be processed, even when parsing a revision created at a time before the macro was deleted. This leads to unparsed wikitext remaining in the HTML output in the case of templates, and to error messages being inserted into the HTML output in the case of Lua modules.

In some cases, we observed that editors deleted a macro and created it again with the same name later. This action introduces the problem of losing the revision history of the macro before its second creation. In such cases, we assume that the oldest macro revision available approximates best how the macro looked before its deletion and use that version when parsing article revisions written before the macro was deleted.

We emphasize that the limitation of deleted pages, templates, and modules is not introduced by our parsing process. Rather, it is inherited from Wikipedia's deliberate policy of permanently deleting the entire history of deleted pages. Neither can the limitation be avoided by using the Wikipedia API to parse old wikitext revisions; the same inconsistencies and error messages would ensue. On the contrary, WikiHist.html produces strictly more accurate approximations of the HTML appearance of historical revisions than the Wikipedia API, for the API always uses the latest revision of all templates and modules, rather than the revision that was actually in use at the time of the article revision by which it was invoked.

7.3 System architecture and configuration

Wikipedia runs on MediaWiki, a content management system built around a backend database that stores all information on pages, revisions, users, templates, modules, etc. In this project we only require one core functionality: parsing article content from wikitext to HTML. In MediaWiki's intended use case, parsing is performed on demand, whenever a page is requested by a Web client. Our use case, on the contrary, consists in bulk-parsing a very large number of revisions. Since MediaWiki was not built for such bulk-parsing, the massive scale of our problem requires a carefully designed system architecture.

System overview. Our solution is schematically summarized in Fig. 7.2. As mentioned in Sec. 7.2.1, the input to the parsing process consists of the hundreds of XML files that make up English Wikipedia's full revision history in wikitext format. Our system processes the XML files in parallel, each in a separate parent process running on a CPU core of its own. Parent processes read the data from disk (in a streaming fashion using a SAX XML parser) and spawn child processes that parse the article contents from wikitext to HTML. Each child process has access to its own dedicated MediaWiki instance. The parent processes collect the HTML results from the child processes and write them back to disk. Although this architecture is straightforward in principle, several subtleties need to be handled, described next.

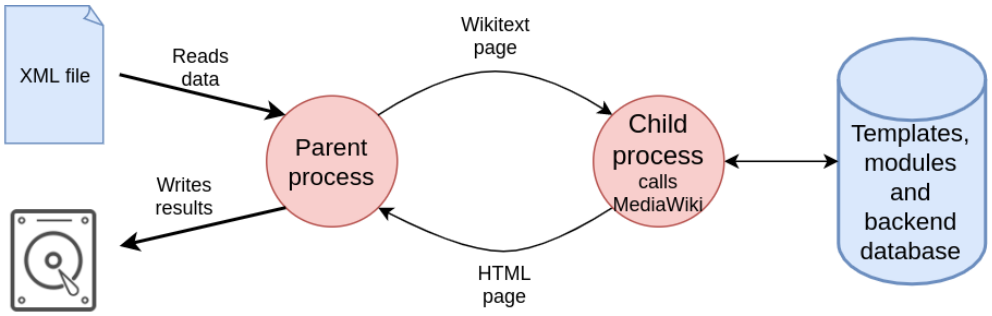


Figure 7.2: Architecture for parsing Wikipedia's revision history from wikitext to HTML.

Template and module expansion. Wikitext frequently invokes macros (templates and modules) that need to be expanded when parsing to HTML. Since macros may (and frequently do) themselves change over time, it is important to use the version that was active at the time of the article revision that is being parsed, given that we aim to reconstruct the HTML as it appeared at the time of the article revision. MediaWiki unfortunately does not provide such a retroactive macro expansion mechanism, but instead always uses the latest available version of each macro. We therefore provide a workaround ourselves, by implementing an interceptor that, every time a macro is expanded, selects the historically correct macro version based on the revision date of the page being parsed, and returns that macro version to the parser instead of the default, most recent version.[VII] More precisely, we select the most recent macro version that is older than the article revision being parsed.

MediaWiki version. Not only templates and modules, but also the MediaWiki software itself has changed over time, so in principle the same wikitext might have resulted in different HTML outputs at different times. To strictly reproduce the exact HTML served by Wikipedia at a given time, one would need to use the MediaWiki version deployed by Wikipedia at that time. Juggling multiple versions of MediaWiki would, however, severely complicate matters, so we started by consulting the Internet Archive Wayback Machine[VIII] in order to compare identical article revisions in different HTML snapshots taken at times between which live Wikipedia's MediaWiki version changed. Screening numerous revisions this way, we found no noticeable differences in the HTML produced by different MediaWiki versions and therefore conclude that it is safe to use one single MediaWiki version for all revisions. In particular, we use the latest long-term support version of MediaWiki, 1.31.[IX]

Parser extensions. MediaWiki offers numerous extensions, but not all extensions used by live Wikipedia are pre-installed in MediaWiki's default configuration. We therefore manually installed all those extensions (including their dependencies) that are necessary to reproduce live

[VII] To support this procedure, the caching mechanisms of MediaWiki must be turned off, which introduces significant latency.
[VIII] https://archive.org/web/
[IX] https://www.mediawiki.org/wiki/MediaWiki_1.31

Wikipedia's parsing behavior. In particular, we mention two crucial parser extensions: *ParserFunctions*,[X] which allows for conditional clauses in wikitext, and *Scribunto*,[XI] the extension that enables the usage of Lua modules in wikitext.

Database connectivity. By design, MediaWiki instances cannot run without a persistent database connection. However, given that (1) wikitext-to-HTML parsing is the only functionality we require, (2) the input to be parsed comes directly from a wikitext dump rather than the database, and (3) we intercept template and module lookups with custom code (see above), we never actually need to touch the MediaWiki database. Hence we need not populate the database with any data (but we still need to create empty dummy tables in order to prevent MediaWiki from throwing errors).

Scaling up. Given the amount of wikitext in the full revision history, parallelization is key when parsing it. We explored multiple common solutions for scaling up, including Spark and Yarn, but none of them satisfied all our requirements. Therefore, we instead settled on a custom, highly-optimized implementation based on Docker[XII] containers: we bundle the modified MediaWiki installation alongside the required MySQL database into a standalone Docker container and ship it to each machine involved in the data processing.

Failure handling. Failures can happen during the parsing process for multiple reasons, including malformed wikitext, memory issues, etc. Detecting such failures is not easy in MediaWiki's PHP implementation: in case of an error it calls the `die` function, which in turn interrupts the process without raising an exception. As a workaround, the parent processes (one per XML file; see above) are also responsible for monitoring the status of the child processes: whenever one of them fails, the event is detected and logged. By using these logs, processing of the failure-causing revisions can be resumed later, after writing custom code for recognizing problematic wikitext and programmatically fixing it before sending it to the parser. Our deployed and released code incorporates all such fixes made during development runs.

Computation cost. We used 4 high-end servers with 48 cores and 256 GB of RAM each. Each core ran one parent and one child process at a time. In this setup, parsing English Wikipedia's full revision history from wikitext to HTML took 42 days and, at a price of CHF 8.70 per server per day, cost a total of CHF 1,462.

7.4 Advantages of HTML over wikitext

Our motivation for taking on the considerable effort of parsing Wikipedia's entire revision history from wikitext to HTML was that raw wikitext can only provide an approximation of the full information available in a Wikipedia article, primarily because the process of parsing

[X] https://www.mediawiki.org/wiki/Extension:ParserFunctions
[XI] https://www.mediawiki.org/wiki/Extension:Scribunto
[XII] https://en.wikipedia.org/w/index.php?title=Docker_(software)&oldid=934492701

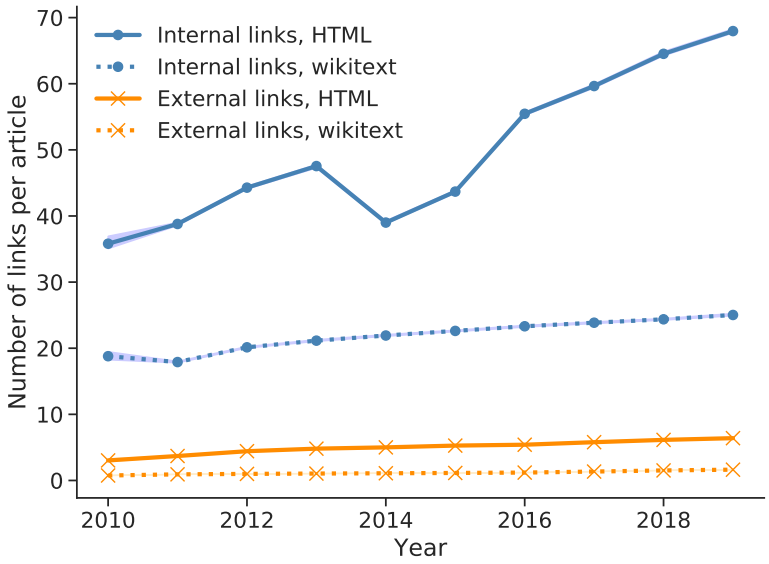


Figure 7.3: Number of links extracted from wikitext and HTML, averaged over 404K articles created in 2009; 95% error bands estimated via bootstrap resampling.

wikitext to HTML tends to pull in information implicit in external templates and modules that are invoked by the wikitext.

In this section, we illustrate the shortcomings of wikitext by showing that a large fraction of the hyperlinks apparent in the parsed HTML versions of Wikipedia articles are not visible in wikitext, thus providing researchers with a strong argument for using WikiHist.html instead of raw wikitext dumps whenever their analyses require them to account for all hyperlinks seen by readers [45, 46, 138, 200].

Prevalence of HTML-only links over time. First we quantify the difference in the number of links that can be extracted from the wikitext vs. HTML versions of the same article revisions. To be able to determine whether the difference has increased or decreased with time, we study the 10 years between 2010 and 2019. In order to eliminate article age as a potential confound, we focus on the 404K articles created in 2009. For each article created in 2009, we study 10 revisions, viz. the revisions available at the start of each year between 2010 and 2019. For each revision, we extract and count internal links (pointing to other English Wikipedia articles) as well as external links (pointing elsewhere) in two ways: (1) based on the raw wikitext, (2) based on the HTML available in WikiHist.html.[XIII]

Fig. 7.3 shows the number of links per year averaged over the 404K articles, revealing a large gap between wikitext and HTML. The gap is significant (with non-overlapping error bands) for

[XIII] As internal links, we consider only links pointing to articles in the main namespace and without prefixes, thus excluding talk pages, categories, etc. We exclude self-loops. In all analyses, if the same source links to the same target multiple times, we count the corresponding link only once. To extract internal links from wikitext, we used a regular expression crafted by Consonni *et al.* [33].

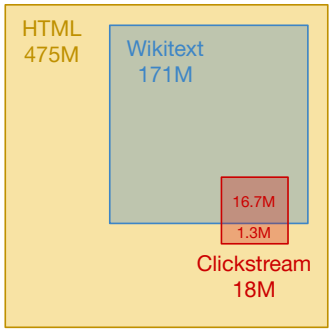


Figure 7.4: Venn diagram of number of links in wikitext and HTML revisions of 1 January 2019, and in Clickstream release of January 2019.

both internal and external links, but is much wider for internal links. Notably, for most years we can extract more than twice as many links from HTML as from raw wikitext, implying that researchers working with raw wikitext (presumably the majority of researchers at present) see less than half of all Wikipedia-internal links.

Via manual inspection we found that most of the links available in HTML only (henceforth "HTML-only" links) are generated by templates and Lua modules to enhance the navigation, e.g., in infoboxes on the top right of pages or as large collections of related links at the bottom of pages.[XIV]

Popularity of HTML-only links. Next we aim to determine how important HTML-only links are from a navigational perspective, operationalizing the importance of a link in terms of the frequency with which it is clicked by users of Wikipedia. If, for argument's sake, HTML-only links were never clicked by users, these links would be of little practical importance, and the necessity of working with WikiHist.html rather than raw wikitext dumps would be less pronounced. If, on the contrary, HTML-only links were clicked as frequently as links also available in wikitext, then researchers would see a particularly skewed picture by not observing over half of the available links.

Click frequency information is publicly available via the Wikipedia Clickstream dataset,[XV] which counts, for all pairs of articles, the number of times users reached one article from the other via a click, excluding pairs with 10 or fewer clicks. We work with the January 2019 Clickstream release.[XVI]

[XIV] The noticeable dip in 2014/2015 of the number of internal links extracted from HTML (top, blue curve in Fig. 7.3) was caused by the introduction of a then-popular Lua module called *HtmlBuilder*, which, among other things, automated the insertion of certain links during wikitext-to-HTML parsing. The module was later deleted and could not be recovered (cf. Sec. 7.2.4), thus leading to those links being unavailable in WikiHist.html and therefore to an underestimation of the true number of links present during the time that *HtmlBuilder* was active.

[XV] https://dumps.wikimedia.org/other/clickstream/

[XVI] Since redirects have been resolved in the Clickstream, we also do so for links extracted from wikitext and HTML in this analysis.

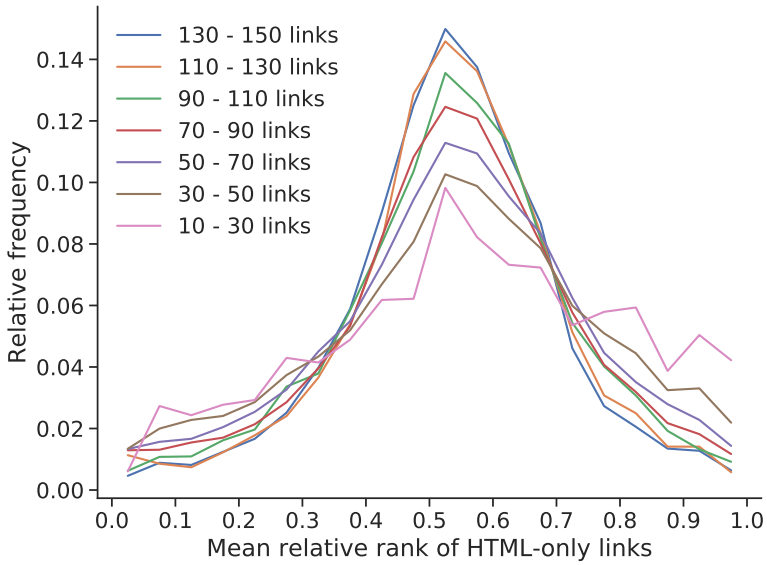


Figure 7.5: Histograms of mean relative rank of HTML-only links among all HTML links in terms of click frequency, averaged over 405K articles. One curve per out-degree bracket.

The situation is summarized as a Venn diagram in Fig. 7.4. On 1 January 2019, there were 475M internal links in WikiHist.html (extracted from 5.8M articles). Out of these, only 171M (36%) are also present in wikitext, and 18M (3.8%) are present in the Clickstream (i.e., were clicked over 10 times in January 2019). Strikingly, out of the 18M links present in the Clickstream, 1.3M (7.2%) cannot be found in wikitext, accounting for 6.1% of all article-to-article clicks recorded in the Clickstream. That is, joining Clickstream statistics with the contents of the respective articles is not fully feasible when working with raw wikitext. With WikiHist.html, it is.

We now move to quantifying the navigational importance of the 1.3M Clickstream links available in HTML only, relative to the set of all 18M Clickstream links available in HTML. (In this analysis, we consider only the 18M links present in the Clickstream.) For each of the 405K articles containing at least one HTML-only link, we sort all links extracted from WikiHist.html by click frequency, determine the relative ranks of all HTML-only links, and average them to obtain the mean relative rank of HTML-only links in the respective article. In the extreme, a mean relative rank of zero (one) implies that the HTML-only links are the most (least) popular out-links of the article.

Fig. 7.5 shows histograms of the mean relative rank of HTML-only links. To exclude the total number of out-links as a confound, we stratify articles by the number of out-links and draw a separate histogram per stratum. If HTML-only links were the least important links, the histograms would show a sharp peak at 1; if HTML-only links were no different from the other links, the histogram would show a sharp peak at 0.5. We clearly see that reality resembles the latter case much more than the former case. From a navigational perspective, HTML-only

links are as important as the links also present in wikitext, and to disregard them is to neglect a significant portion of users' interests.

Beyond hyperlinks. This section illustrated the added value of WikiHist.html over raw wikitext dumps using the example of hyperlinks, but hyperlinks are not the only information to remain hidden to researchers working with wikitext only. Templates and modules invoked during the parsing process may also add tables, images, references, and more.

7.5 Discussion

To date, Wikipedia's revision history was available only in raw wikitext format, not as the HTML that is produced from the wikitext when a page is requested by clients from the Wikipedia servers. Since, due to the expansion of templates and modules, the HTML seen by clients tends to contain more information than the raw wikitext sources, researchers working with the official wikitext dumps are studying a mere approximation of the true appearance of articles.

WikiHist.html solves this problem. We parsed English Wikipedia's entire revision history from wikitext (nearly 1 TB bzip2-compressed) to HTML (7 TB gzip-compressed) and make the resulting dataset available to the public.

In addition to the data, we also release the code of our custom architecture for parallelized wikitext-to-HTML parsing, hoping that other researchers will find it useful, e.g., for producing HTML versions of Wikipedia's revision history in languages other than English.

Conclusion. This chapter introduced WikiHist.html, a large dataset containing the full history of English Wikipedia in HMTL. We described the dataset and the method to allow researchers to reproduce the results. We discussed some use-cases when working with HTML gives some advantages, and we publicly release the data.

Table 7.1: JSON schemas of WikiHist.html dataset. All fields in HTML revision history are copied from wikitext dump, except html, which replaces the original text.

(a) HTML revision history (Sec. 7.2.1)

Field name	*Description*
id	id of this revision
parentid	id of revision modified by this revision
timestamp	time when revision was made
cont_username	username of contributor
cont_id	id of contributor
cont_ip	IP address of contributor
comment	comment made by contributor
model	content model (usually wikitext)
format	content format (usually text/x-wiki)
sha1	SHA-1 hash
title	page title
ns	namespace (always 0)
page_id	page id
redirect_title	if page is redirect, title of target page
html	revision content in HTML format

(b) Page creation times (Sec. 7.2.2)

Field name	*Description*
page_id	page id
title	page title
ns	namespace (0 for articles)
timestamp	time when page was created

(c) Redirect history (Sec. 7.2.3)

Field name	*Description*
page_id	page id of redirect source
title	page title of redirect source
ns	namespace (0 for articles)
revision_id	revision id of redirect source
timestamp	time at which redirect became active
redirect	page title of redirect target (in 1st item of array; 2nd item can be ignored)

Conclusion Part III

8 Discussion

Information seeking is an essential behavioral process that allows people to learn, make decisions, and make sense of the surrounding world. Uncovering the dynamic that guides people in finding information has immediate implications for better understanding our cognitive processes and designing systems that can accommodate our needs better.

In this thesis, we provided the first overview of the online knowledge-seeking patterns by focusing on the case of encyclopedic content consumption. Our work is based on a large-scale analysis of the logs collected from Wikipedia, the largest encyclopedia in English with billions of views every month. Wikipedia was defined as a "living laboratory"[I] [177] to investigate human online behavior, and it represents the ideal candidate to examine how we consume knowledge.

This thesis introduces two major contributions. The first contribution is comprised of three large-scale observational studies of navigation logs that give us a comprehensive picture of the users' behavior. The second contribution is a set of tools to support our work and foster future Wikipedia research. In the following sections, we summarise and discuss the implications and limitations of the overall findings.

8.1 Navigation on Wikipedia

This thesis represents a step in the direction of understanding how we interact with knowledge. We advanced the understanding of how we consume online knowledge by offering the first complete overview of readers' interaction with the content of Wikipedia. We centered our study around three stages of navigation on Wikipedia to describe 1) how readers reach the platform, 2) how readers navigate the platform, and 3) how readers leave the platform.

[I] As long as the studies do not disrupt the functionalities or the content of the platform: https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not

8.1.1 Summary of findings

We discovered that sessions are generally short, and readers stop the navigation more than 68% of the time after the first pageload (Sec. 3.5.3). Although it is hard to establish if the readers were satisfied with the content obtained without qualitative investigations, this observation may suggest that modern search engines efficiently point the users directly to the desired content. Their efficiency enables the tendency to access the desired content immediately and leave. In line with previous studies [56], this pattern suggests that readers tend to exhibit aspects of random surfer behavior. On the other hand, internal navigation may be associated with specific intentions, such as in-depth learning or fighting boredom that, on a large scale, occur less often than shallow fact-checking. Additionally, our analysis highlights that search engines have a navigational role when readers are interested in more information on a topic and willing to engage in longer sessions. Readers often prefer to use external search even when the link to the desired content is available on the current page (Sec. 3.4).

Our quantitive analysis shows that readers have different behavior when accessing Wikipedia from articles associated with entertainment topics and biographies. They tend to have longer navigation sessions, and their explorations tend to generate wider trees than navigations originating from articles about STEM topics. As observed in previous work [108], this finding may suggest a less focused reading intent in articles about entertainment that may be more often connected with exploration driven by boredom. Articles about human aspects such as entertainment, history, and politics are also common Wikipedia entry points where readers are more prone to fall into the so-called wiki rabbit holes [141], long internal explorations that often lead the readers to unexpected articles. Similarly, readers manifest the human social nature by engaging more with the citations in biographies and with content associated with human factors such as relationships (wife, family, daughter) and life events (wedding, born, died).

On the contrary, official links receive more engagement on pages about business and education (Sec. 5.3.2). An in-depth investigation of the type of links used more often and the dynamic of engagement with them shows that readers on some links engage consistently more and faster. The platform acts as a gateway for business websites and content that is typically not easily accessible from search engines, with an estimated economic value of this outgoing traffic of several million dollars per month. This observation, combined with the frequent click of references routing the readers to open access documents, highlights the previously largely undocumented and underestimated navigation role of Wikipedia for scientific and business content.

Finally, we described the important role of the article quality in the readers' navigation behavior. We found that the exploration of a path by following internal links has a higher chance of terminating in low-quality articles. This finding is aligned with the definition of information scent used in information foraging theory. Humans hunting for information follow the scent with higher chances of leading to the desired content; when scent loses intensity, they move

to more promising information sources. At the same time, low-quality articles exhibit higher engagement with the citations suggesting that the unsatisfied readers abandon the platform to satisfy their information need somewhere else.

8.1.2 Implications and next steps

Our findings highlight that Wikipedia is an intricate system; it fulfills a diverse set of needs that vary across numerous features of the readers, including temporal and geographical properties. The results of our large-scale analysis on the readers' behavior have implications from Wikipedia and the Web ecosystem.

Formalising logs analysis. In our analysis, we employed large-scale logs that give us a population-scale overview of the behavior on Wikipedia. One of the critical challenges we face in our work is the lack of standard pipelines to process and aggregate these logs. Unlike fields such as NLP or computer vision, where the preprocessing steps are de facto standardized, modeling behavior from access logs does not have well-established approaches. The definition of a session in Web navigation is often ambiguous and hard to define. Our work described in Chapter 3 offered two approaches based on trees and temporal sequences with their relative advantages and disadvantages. We provided an operationalization of the navigation sessions —and engagement with Wikipedia content— that can serve as an analytical framework for future research based on request logs. Future work should investigate how to generalize these approaches to define standard aggregation methods and common tasks for different use cases.

Similarly, methods to embed users based on their navigation traces need more attention. When working with logs, identifying two users that generate navigation trees with similar structures and across similar topics is a common use case, but we lack a general solution. By borrowing from user engagement research [100] —that has overall different objectives— future work should investigate how to formalize this problem.

Toward a deeper understanding of the user behavior. The navigation logs show that Wikipedia fulfills various information needs and readers exhibit diverse navigation patterns. Using large-scale digital traces offers important advantages over other methods when we are interested in quantitative measurement of a phenomenon or behavior [161]. However, purely log-based analysis has its limitations, and it should be considered complementary and not as a substitutive approach. Previous work indicates that big-data analyses are not immune to biases introduced by algorithmic dynamics [106, 196], data collection problems, preprocessing errors, and measurement errors [105, 196].

Additionally, a pressing limitation in our studies is the inability to offer any definitive answer on why precisely a reader engaged in a specific behavior. The information-seeking behaviors we described in this thesis may be associated with different unobservable intents and goals that will require more investigation.

Future work should extend what we know about readers motivations [108, 171] and focus on isolating different forms of information behavior to understand the reader's intent. By enriching the behavioral patterns with qualitative feedback, we can understand the user's objective and design ways to facilitate more efficient access to the desired information. For example, using only logs when we observe interactions with the content of the article *Pliny the Elder*[II] —the author of *Naturalis Historia*, the first pseudo-encyclopedia, and dead in Pompeii during the eruption—, it is nearly impossible to predict the reason and intent of the visit. The reader may be engaging with the article after watching a movie about the Pompeii catastrophe (entertainment intent), after learning about his frantic run to the beach protecting his head with a pillow (curiosity intent) [112], or even to complete homework of Latin (education intent). Each of these visits may leave in the logs very different digital fingerprints.

Additionally, future user modeling should consider the behavioral patterns in their entirety by considering all the possible interactions with the page, such as the relationship between engagement with images, previews, internal and external links.

Implications for Wikipedia content. From our analyses, we have strong evidence that Wikipedia receives significantly more engagement on content regarding human factors. Longer sessions on entertainment and biographies or high click-through rate for references about recent and life-related events paint the picture that many readers get informed about other people's lives from Wikipedia. Probably not surprisingly, some of our analyses suggest that some of the readers use Wikipedia as a gossip magazine. These findings pose the interrogative on where to set the notability threshold, what should be considered encyclopedic material about the person worth including, and how we can ensure content from a neutral standpoint. Future work on biased representation on Wikipedia could take into account readership metrics to detect the content that needs more urgent verification.

Implications for the debate on Wikipedia role. Previous work [77] showed that Wikipedia, besides many other intrinsic values, such as its educational role, has indirect economic impacts in the real world. Our work complements these findings by showing that the platform has a large-scale invisible economic impact on the Internet economy. Many websites benefit from the traffic received through Wikipedia, and in the case of business activities, these users can be potential customers. Although our estimation is a back-of-the-envelope approximation, the hypothetical economic value offered by Wikipedia gives important hints to fuel the debate around the impact of free software and open knowledge. One common misconception is that there is no real value in free licenses and platforms based on them, and often the economic benefits generated by free platforms are hard to track, undermined, and, ultimately, disregarded. Our findings show that the benefit is measurable, and its direct estimation can foster policy conversations to keep the knowledge ecosystem open.

[II] https://en.wikipedia.org/wiki/Pliny_the_Elder

Implication for improving Wikipedia. Our work is not only important for researchers but also for the contributors who work to improve Wikipedia. Understanding the dynamics of knowledge consumption could empower the community to make informed decisions around the organization of Wikipedia content. Previous work [107] found a misalignment between the amount of attention that editors and readers give to the same articles. Readership-driven prioritization, such as improving pages subject to frequent abandonment or adding links to incomplete citations used frequently, can help editors improve Wikipedia. Editors can improve the content to fulfill all the frequent information needs and offer a richer experience.

Additionally, these improvements can extend beyond Wikipedia when external websites that offer information not available in the encyclopedia can make informed decisions on the content to improve. For example, organizations like the Internet Archive that aim to offer free universal access to a large archive of digitalized books[III] can benefit from modeling readers' behavior. These books are manually searched and scanned by a crowd of volunteers that then enrich Wikipedia by adding the links as article references. Large-scale efforts like this one could benefit from prediction models that can prioritize the work of the volunteers to attempt to match the information needs of the readers.

At the same, we can use these findings as a starting point to investigate future developments of the Wikipedia platform. By adding the readers' behavior in the design loop, developers and UX designers can implement new forms of knowledge access. Modeling how readers consume the content to infer their intention can help augment the navigation. In the future, adaptive interfaces optimized for different knowledge consumption patterns can be employed to offer customized Web experiences. For example, the potential of Wikipedia can be pushed further in the education domain, especially in countries with a limited access to the education system. A reading pattern that suggests the user engaged in an in-depth studying session can be used to generate a collection of articles that can progressively guide the reader in mastering the topic. On the contrary, a pattern associated with curiosity can be used to engage the reader in more intriguing, obscure, bizarre, or entertaining pathways.

Implications for theoretical next steps. We envision that these findings will set the basis for future contributions in knowledge modeling research. We can enhance the relationship between concepts by considering the navigational aspects and encoding how humans traverse the knowledge graph. Additionally, the work presented in this thesis provides a first large-scale characterization of access to knowledge that can have implications for developing theoretical frameworks to describe our navigation patterns. Understanding what drives the readers in following specific trails can inform researchers about the specific properties of the information scent that guide our search for information online. These findings can be instrumental in developing novel theories on how humans move in information networks.

Our work also leads to the exciting opportunity to learn and share these organic digital traces developing in the knowledge space. In 1945, Vannevar Bush sketched his vision of an in-

[III] https://archive.org/details/internetarchivebooks

formation management device —the “memex”—that would allow users to not only retrieve documents quickly but to also easily interlink documents [23]. With the advent of the Web, the hyperlink structure envisioned by Bush has since become a reality—but Bush’s vision went further: he saw the trails taken by users as first-class citizens of the hypertext environment, as important as the text content itself: “Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified” [23]. Our technological reality has not entirely caught up with Bush’s vision yet, and the present work should be seen as a small step toward achieving it: we have started by describing the “associative trails running through” Wikipedia, and we hope that its future versions will build on these insights to incorporate tools and features that will allow readers to continually benefit from each other’s encyclopedic trailblazing.

8.1.3 Limitations

The content of this thesis offers the first overview of the patterns associated with knowledge consumption focusing on the special case of browsing Wikipedia. However, despite Wikipedia being one of the most visited websites and a very common choice to satisfy many information needs, our findings represent only one piece of the puzzle.

Since the main focus of this thesis is encyclopedic knowledge consumption, general online knowledge consumption that extends beyond Wikipedia is not investigated. For instance, as described in Chapter 3, most of the external traffic of Wikipedia originates from search engines, which constitute the central hub to find content, sometimes even replacing internal links. Using knowledge panels sourced from Wikipedia, they satisfy the users’ information without directing any traffic to the platform. This limited view and the crucial role of search engines pose some challenges in drawing general conclusions about general knowledge consumption on the Web and how the document was searched. These observations stress the need for further work in understanding how readers reach Wikipedia, possibly by investigating the search queries used.

Additionally, the work described in this thesis does not investigate different websites that may satisfy the user’s information needs, such as MOOCs, Youtube, and Q&A websites. Our findings, supported by previous work, highlight how content consumption patterns are affected by many factors such as topic, device, time, and even the layout of the document. These properties can influence the user’s behavior and lead to different conclusions based on the context of knowledge consumption. Future work should investigate how people satisfy their information needs by considering the broader navigation on the entire Web.

8.2 Tools and datasets

The second major contribution of this thesis is the release of WikiPDA, a method to obtain cross-lingual topic models, and WikiHist.html, a large dataset with the full history of Wikipedia

in HTML. Both our contributions are publicly released to the community to support Wikipedia research.

WikiPDA can enable researchers to address a series of questions that were hard to answer before, such as language-specific topic biases, distances between articles in different languages, and language-independent articles classification. Our framework based on bag-of-links, instead of bag-of-words, is customized for Wikipedia, but our results suggest that with a dedicated entity linker, the same method is effective for every document. Using regression analysis on the topics distribution extracted from all the articles in 28 languages, we observed topic biases, such as a higher presence of comics in Dutch or Ice Hockey in Korean. Our analysis showed that language-specific topic biases are measurable. These findings can be used to understand cultural differences, measure what biases readers are exposed to, or find what topics editors should pay more attention to when contributing to a specific language edition. Another crucial aspect where WikiPDA can help the community is detecting malicious coordinate behaviors that could manipulate the content. This adversarial behavior could be perpetrated by organized groups such as governments or companies interested in adapting the content to fit a specific agenda not aligned with Wikipedia's mission: product promotion, removing undesired facts. Currently, Wikipedia does not have any system to detect and visualize edits that individually look innocuous but collectively skew the content in an undesired direction. WikiPDA could help mitigate this issue by monitoring long-term information spread and global trends across languages.

WikiHist.html offers an opportunity for all researchers interested in modeling the evolution of Wikipedia. Thanks to a full expansion of the templates, the full history in HTML allows researchers to overcome the limitation of wikitext and to have a more accurate representation of the exact content visualized by the readers at different times in the past.

8.2.1 Impact on the Wikipedia ecosystem

These methodological contributions already bring concrete implications for the progress and research on Wikipedia. The approach employed in WikiPDA based on bag-of-links has been explored in a subsequent Wikimedia Foundation work [85]. Their implementation offers encouraging results, and they aim to expand the official ORES topic model to all the Wikipedia languages. At the same time, the release of WikiHist.html enabled discussions on the need for a Wikipedia dump in HTML format. Although WikiHist.html remains the only historical dataset available, starting from November 2021, the Wikimedia Foundation began the regular release of the monthly snapshot in HTML[IV].

8.3 Future research opportunities

Wikipedia is a complex system composed of linked concepts in multiple languages, multimedia items, and human behavioral traces. The richness of its data enables scholars to address a

[IV] https://dumps.wikimedia.org/other/enterprise_html/

tremendous amount of research questions. In the past, a substantial amount of work has gone into using content to train AI models and understand the peer production dynamics. However, as presented in this thesis, given the large user-base that visits its articles every day, Wikipedia is also an ideal candidate to investigate human navigation on the Web, our behavior around online knowledge, and improve our web experiences.

Recommendation systems. Future recommender systems can use our insights, on the one hand, to design better experiences for the readers and, on the other hand, to prioritize the writing of missing content for the editors. For example, readers with patterns that reassemble a session with a learning objective can be guided in the necessary steps to understand the subject, whereas readers navigating entertainment articles may be more interested in having higher visibility on recent content. At the same time, measuring the readers' information needs allows the editors to identify and improve articles that, because of some properties —such as low quality—, may lead people to abandon the navigation. Similarly, deeper investigations of the readers' behavior can uncover other elements of the articles that receive significant engagement without satisfying the information needs, such as incomplete references or poor-quality images. In doctoral work not included in this thesis [140], we explore this aspect by designing and testing with users a system that provides the editors a template with the relevant sections to write.

Measuring socio-cultural differences. Wikipedia offers opportunities to explore the daily patterns of content consumption on a large scale. Our work can be extended to understand how readership behavior differs during the day and across different countries. For example, socioeconomic conditions could be associated with diverse information needs that influence reading patterns. Additionally, ORES [67] and the progress in cross-lingual topic models like WikiPDA [144] can enable researchers to compare articles of different languages. The dataset offered by Wikipedia gives an unprecedented view on a global scale of how people behave online, and understanding these aspects is important to comprehend implicit cultural differences and potentially adapt the platform to the different information needs. Most of the current work on readers' behavior is focused on the English edition of Wikipedia, and expanding to multiple languages can give us insights into the shared behavior. Similarly, combining cross-topic and cross-lingual studies with readership patterns can expose potential biases to which readers are unknowingly exposed. At the same time, qualitative studies on the readers' needs and expectations regarding their information ecosystem can be used to design novel strategies and offer access to knowledge in alternative formats.

Engagement with Wikipedia content. Future investigations could also go beyond the article-to-article navigation and focus on the consumption patterns on the page. The usage patterns of essential elements of the Wikipedia articles, such as multimedia items, links previews, or the article's history page, remain still largely unexplored. For example, analyzing how readers consume images and videos could give us great insights into the role of multimedia elements in the comprehension of the article's subject, and the link preview can expose different types of

intents when readers seek information. Additionally, consumption patterns can be influenced by many internal and external factors, such as layout change, the controversy of the content, or external events. Future work should investigate how these properties impact navigation behavior.

Navigation beyond Wikipedia. As discussed in the limitations of the present work (Sec. 8.1.3), Wikipedia is not an isolated island on the Web, and when readers look for information may navigate multiple websites. In our work, we shed light on the behavior on Wikipedia, but we are limited to the source and destination of the traffic within one hop of distance. Our observation may be only a small piece of larger and more complex information-seeking behavior on the Web. Future work should investigate how people find the desired information by jumping from website to website and how Wikipedia fits in their navigation patterns. Researchers studied some of these aspects in the early stage of the Web, but previous work observed that navigation patterns changed over time [36]. Recent transformations in computer literacy and Web topology may need us to revise what we know about navigation on the broad Web.

Quantifying different values of Wikipedia. Additional work is also needed in understanding the role of Wikipedia in the broader Web and its multiple types of value offered. In this thesis, we presented its navigational value by acting as a stepping stone to external websites and the economic value that it could provide to the owners of the linked websites. Besides these and the importance for the success of search engines [120, 191], Wikipedia offers *for free* a large variety of societal benefits that are worth quantifying. Some examples include its educational role, potential societal development, and the indirect economic impact of its data. The assessment of its value is crucial for legal and policy conversations centered around the future of the open knowledge ecosystem.

Validity of conclusions based on Wikipedia data. Finally, we need novel principles analysis frameworks to assess the validity of the findings beyond Wikipedia. In our and similar works, Wikipedia is used as a reference to study the knowledge-seeking patterns on the Web, but we lack a formal framework to measure how much these conclusions can generalize.

Bibliography

[1] Eytan Adar, Jaime Teevan, and Susan T Dumais. “Large scale analysis of web revisitation patterns”. In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2008.

[2] Eytan Adar, Jaime Teevan, and Susan T Dumais. “Resonance on the web: web dynamics and revisitation patterns”. In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2009.

[3] Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. “Find it if you can: a game for modeling different types of web search success using interaction data”. In: *Proc. Conference on Research & Development in Information Retrieval (SIGIR)*. 2011, pp. 345–354.

[4] Réka Albert, Hawoong Jeong, and Albert-László Barabási. “Diameter of the world-wide web”. In: *nature* 401.6749 (1999), pp. 130–131.

[5] Alphabet Inc. *Alphabet announces fourth quarter and fiscal year 2019 results*. 2019. URL: https://web.archive.org/web/20210104101629/https://abc.xyz/investor/static/pdf/2019Q4_alphabet_earnings_release.pdf.

[6] Ashton Anderson, Ravi Kumar, Andrew Tomkins, and Sergei Vassilvitskii. “The Dynamics of Repeat Consumption”. In: *Proc. International World Wide Web Conference (WWW)*. 2014.

[7] Dan Andreescu, Kinneret Gordon, Isaac Johnson, and Nicholas Perry. *Searching for Wikipedia*. https://techblog.wikimedia.org/2021/06/07/searching-for-wikipedia/. accessed: 13 October 2021. 2021.

[8] Akhil Arora, Martin Gerlach, Tiziano Piccardi, Alberto García-Durán, and Robert West. “Wikipedia Reader Navigation: When Synthetic Data Is Enough”. In: *Proc. International Conference on Web Search and Data Mining (WSDM)*. 2022.

[9] Sumit Asthana and Aaron Halfaker. “With few eyes, all hoaxes are deep”. In: *Proc. ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*. 2018.

[10] Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. “Towards a science of user engagement (position paper)”. In: *WSDM workshop on user modelling for Web applications*. 2011.

[11] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. "Dbpedia: A nucleus for a web of open data". In: *The semantic web*. Springer, 2007, pp. 722–735.

[12] Peter C Austin. "An introduction to propensity score methods for reducing the effects of confounding in observational studies". In: *Multivariate behavioral research* 46.3 (2011), pp. 399–424.

[13] Mamoun A Awad and Latifur R Khan. "Web navigation prediction using multiple evidence combination and domain knowledge". In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37.6 (2007), pp. 1054–1062.

[14] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. "Faces engage us: Photos with faces attract more likes and comments on instagram". In: *Proc. SIGCHI conference on human factors in computing systems*. 2014.

[15] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. "Omnipedia: Bridging the Wikipedia language gap". In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2012.

[16] Nicola Barbieri, Fabrizio Silvestri, and Mounia Lalmas. "Improving post-click user engagement on native ads via survival analysis". In: *Proc. International World Wide Web Conference (WWW)*. 2016.

[17] Eda Baykan, Monika Henzinger, Ludmila Marian, and Ingmar Weber. "Purely URL-based topic classification". In: *Proc. International World Wide Web Conference (WWW)*. 2009.

[18] Austin R. Benson, Ravi Kumar, and Andrew Tomkins. "Modeling User Consumption Sequences". In: *Proc. International World Wide Web Conference (WWW)*. 2016.

[19] Ivan Beschastnikh. "Wikipedian Self-Governance in action: Motivating the policy lens". In: *Proc. International Conference on Web and Social Media (ICWSM)*. 2008.

[20] David M. Blei. "Probabilistic topic models". In: *Communications of the ACM* 55.4 (2012), pp. 77–84.

[21] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[22] Jordan Boyd-Graber and David M. Blei. "Multilingual topic models for unaligned text". In: *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*. 2009.

[23] Vannevar Bush *et al.* "As we may think". In: *The atlantic monthly* 176.1 (1945), pp. 101–108.

[24] Ewa S. Callahan and Susan C. Herring. "Cultural bias in Wikipedia content on famous persons". In: *Journal of the American Society for Information Science and Technology* 62.10 (2011), pp. 1899–1915.

[25] Lara D Catledge and James E Pitkow. "Characterizing browsing strategies in the World-Wide Web". In: *Computer Networks and ISDN systems* 27.6 (1995), pp. 1065–1073.

[26] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. "Reading tea leaves: How humans interpret topic models". In: *Proc. Advances in Neural Information Processing Systems*. 2009.

[27] Xiaoxi Chelsy Xie, Isaac Johnson, and Anne Gomez. "Detecting and gauging impact on Wikipedia page views". In: *Proc. International World Wide Web Conference (WWW)*. 2019.

[28] Ed H Chi, Peter Pirolli, Kim Chen, and James Pitkow. "Using information scent to model user information needs and actions and the Web". In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2001.

[29] Flavio Chierichetti, Ravi Kumar, Prabhakar Raghavan, and Tamas Sarlos. "Are Web Users Really Markovian?" In: *Proc. International World Wide Web Conference (WWW)*. 2012.

[30] Jörg Claussen, Tobias Kretschmer, and Philip Mayrhofer. "The effects of rewarding user engagement: The case of Facebook apps". In: *Information Systems Research* 24.1 (2013), pp. 186–200.

[31] Andy Cockburn and Bruce McKenzie. "What do web users do? An empirical analysis of web use". In: *International Journal of human-computer studies* 54.6 (2001), pp. 903–922.

[32] Alexis Conneau and Guillaume Lample. "Cross-lingual language model pretraining". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 7059–7069.

[33] Cristian Consonni, David Laniado, and Alberto Montresor. "WikiLinkGraphs: a complete, longitudinal and multi-language dataset of the Wikipedia link networks". In: *Proc. International Conference on Web and Social Media (ICWSM)*. Vol. 13. 2019, pp. 598–607.

[34] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, *et al.* "Predicting protein structures with a multiplayer online game". In: *Nature* 466.7307 (2010), pp. 756–760.

[35] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. "An experimental comparison of click position-bias models". In: *Proc. International Conference on Web Search and Data Mining (WSDM)*. 2008.

[36] Kyle Crichton, Nicolas Christin, and Lorrie Faith Cranor. "How Do Home Computer Users Browse the Web?" In: *ACM Transactions on the Web (TWEB)* 16.1 (2021), pp. 1–27.

[37] William Cronon. "Scholarly authority in a Wikified world". In: *Perspectives in History* (2012).

[38] Alexander Dallmann, Thomas Niebler, Florian Lemmerich, and Andreas Hotho. "Extracting semantics from random walks on wikipedia: Comparing learning and counting methods". In: *Proc. International Conference on Web and Social Media (ICWSM)*. 2016.

[39] Mukund Deshpande and George Karypis. "Selective markov models for predicting web page accesses". In: *ACM transactions on internet technology (TOIT)* 4.2 (2004), pp. 163–184.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2019.

[41] Megan Rose Dickey. "Google.org donates $2 million to Wikipedia's parent org". In: *TechCrunch* (Jan. 2019). URL: https://web.archive.org/web/20210213130411/https://techcrunch.com/2019/01/22/google-org-donates-2-million-to-wikipedias-parent-org/.

[42] Denis Diderot. *Encyclopédie*. Vol. 14. Garnier frères, 1876.

[43] Dimitar Dimitrov, Florian Lemmerich, Fabian Flöck, and Markus Strohmaier. "Different topic, different trafic: How search and navigation interplay on wikipedia". In: *The Journal of Web Science* 1 (2019).

[44] Dimitar Dimitrov, Florian Lemmerich, Fabian Flöck, and Markus Strohmaier. "Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia". In: *Proc. Conference on Web Science (WebSci)*. 2018.

[45] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. "Visual Positions of Links and Clicks on Wikipedia". In: *Proc. International World Wide Web Conference (WWW)*. 2016.

[46] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. "What Makes a Link Successful on Wikipedia?" In: *Proc. International World Wide Web Conference (WWW)*. 2017.

[47] Kristofer Erickson, Felix Rodriguez Perez, and Jesus Rodriguez Perez. "What is the Commons worth? Estimating the value of Wikimedia imagery by observing downstream use". In: *Proc. International Symposium on Open Collaboration (OpenSym)*. 2018.

[48] Ethan Fast, Binbin Chen, and Michael S. Bernstein. "Empath: Understanding topic signals in large-scale text". In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[49] Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. "Finding news citations for Wikipedia". In: *Proc. Conference on Information and Knowledge Management*. 2016.

[50] Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. "Multiwibi: The multilingual wikipedia bitaxonomy project". In: *Artificial Intelligence* 241 (2016), pp. 66–102.

[51] Andrea Forte, Vanesa Larco, and Amy Bruckman. "Decentralization in Wikipedia governance". In: *Journal of Management Information Systems* 26.1 (2009), pp. 49–72.

[52] Wikimedia Foundation. *Medium-term plan 2019: The model for engagement*. https://meta.wikimedia.org/wiki/Wikimedia_Foundation_Medium-term_plan_2019#The_model_for_engagement. 2019.

[53] Wai-Tat Fu and Peter Pirolli. "SNIF-ACT: A cognitive model of user navigation on the World Wide Web". In: *Human–Computer Interaction* 22.4 (2007), pp. 355–412.

[54] Yuji Fujita, Yuichi Kichikawa, Yoshi Fujiwara, Wataru Souma, and Hiroshi Iyetomi. "Local bow-tie structure of the web". In: *Applied Network Science* 4.1 (2019), pp. 1–15.

[55] Evgeniy Gabrilovich and Shaul Markovitch. "Wikipedia-based semantic interpretation for natural language processing". In: *Journal of Artificial Intelligence Research* 34 (2009), pp. 443–498.

[56] Florian Geigl, Daniel Lamprecht, Rainer Hofmann-Wellenhof, Simon Walk, Markus Strohmaier, and Denis Helic. "Random Surfers on a Web Encyclopedia". In: *Proc. International Conference on Knowledge Technologies and Data-Driven Business*. i-KNOW '15. 2015.

[57] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. "Global disease monitoring and forecasting with Wikipedia". In: *PLoS Computational Biology* 10.11 (2014), e1003892.

[58] Ruili Geng and Jeff Tian. "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage". In: *IEEE Transactions on Human-Machine Systems* 45.1 (2015), pp. 84–94.

[59] Patrick Gildersleve and Taha Yasseri. "Inspiration, Captivation, and Misdirection: Emergent Properties in Networks of Online Navigation". In: *Complex Networks IX* (2018), pp. 271–282.

[60] Patrick Gildersleve and Taha Yasseri. "Inspiration, captivation, and misdirection: Emergent properties in networks of online navigation". In: *Complex Networks IX*. 2018, pp. 271–282.

[61] Carlos Gómez, Brendan Cleary, and Leif Singer. "A study of innovation diffusion through link sharing on stack overflow". In: *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE. 2013, pp. 81–84.

[62] Casper Grathwohl. "Wikipedia comes of age". In: *Chronicle of Higher Education* 57 (2011).

[63] Aditya Grover and Jure Leskovec. "node2vec: Scalable feature learning for networks". In: *Proc. International Conference on Knowledge discovery and Data Mining (SIGKDD)*. 2016.

[64] Amit Gupta, Rémi Lebret, Hamza Harkous, and Karl Aberer. "280 birds with one stone: Inducing multilingual taxonomies from Wikipedia using character-level classification". In: *Proc. AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[65] Amit Gupta, Francesco Piccinno, Mikhail Kozhevnikov, Marius Pasca, and Daniele Pighin. "Revisiting taxonomy induction over wikipedia". In: *Proc. International Conference on Computational Linguistics (COLING)*. 2016.

[66] Aaron Halfaker. “Interpolating quality dynamics in Wikipedia and demonstrating the Keilana effect”. In: *Proc. International Symposium on Open Collaboration (OpenSym)*. 2017.

[67] Aaron Halfaker and R.Stuart Geiger. “ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia”. In: *Proc. Human-Computer Interaction (HCI)*. 2019.

[68] Aaron Halfaker, Os Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. “User Session Identification Based on Strong Regularities in Inter-Activity Time”. In: *Proc. International World Wide Web Conference (WWW)*. 2015.

[69] Shudong Hao and Michael Paul. “Learning multilingual topics from incomparable corpora”. In: *Proc. International Conference on Computational Linguistics*. 2018.

[70] Shudong Hao and Michael J. Paul. “An empirical study on crosslingual transfer in probabilistic topic models”. In: *Computational Linguistics* 46.1 (2020), pp. 95–134.

[71] Brent Hecht and Darren Gergle. “Measuring self-focus bias in community-maintained knowledge repositories”. In: *Proc. International Conference on Communities and Technologies*. 2009.

[72] Brent Hecht and Darren Gergle. “The tower of Babel meets web 2.0: User-generated content and its applications in a multilingual context”. In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2010.

[73] Denis Helic. “Analyzing user click paths in a Wikipedia navigation game”. In: *Proc. International Convention MIPRO*. 2012.

[74] Denis Helic, Markus Strohmaier, Michael Granitzer, and Reinhold Scherer. “Models of human navigation in information networks based on decentralized search”. In: *Proc. ACM Conference on Hypertext and Social Media (HT)*. 2013.

[75] Eelco Herder. “Characterizations of User Web Revisit Behavior.” In: *LWA*. 2005.

[76] Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, and Sara Y. Del Valle. “Forecasting the 2013–2014 influenza season using Wikipedia”. In: *PLOS Computational Biology* 11.5 (2015), e1004239.

[77] Marit Hinnosaar, Toomas Hinnosaar, Michael E Kummer, and Olga Slivko. “Wikipedia matters”. In: *SSRN* (2019).

[78] Matthew Hoffman, Francis R. Bach, and David M. Blei. “Online learning for latent Dirichlet allocation”. In: *Proc. Advances in Neural Information Processing Systems*. 2010.

[79] Thomas Hofmann. “Probabilistic latent semantic analysis”. In: *Proc. Conference on Uncertainty in Artificial Intelligence*. 1999.

[80] Yuheng Hu, Shelly Farnham, and Kartik Talamadupula. “Predicting user engagement on twitter with real-world events”. In: *Proc. International Conference on Web and Social Media (ICWSM)*. 2015.

[81] Bernardo A Huberman and Lada A Adamic. “Growth dynamics of the world-wide web”. In: *Nature* 401.6749 (1999), pp. 131–131.

[82] Jagadeesh Jagarlamudi and Hal Daumé. “Extracting multilingual topics from unaligned comparable corpora”. In: *Proc. European Conference on Information Retrieval*. 2010.

[83] Honey Jindal, Neetu Sardana, and Raghav Mehta. “Efficient web navigation prediction using hybrid models based on multiple evidence combinations”. In: *International Journal of Computers and Applications* 42.7 (2020), pp. 715–728.

[84] Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Jeff Donahue, and Sarah Tavel. “Visual search at Pinterest”. In: *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015.

[85] Isaac Johnson, Martin Gerlach, and Diego Sáez-Trumper. “Language-agnostic Topic Classification for Wikipedia”. In: *Companion Proceedings of the Web Conference 2021*. 2021.

[86] Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. “Global gender differences in Wikipedia readership”. In: *Proc. International Conference on Web and Social Media (ICWSM)*. 2020.

[87] Michael Johnston. “Wikipedia revenue analysis: How a wiki could make $2.3B a year”. In: *MonetizePros* (June 2013). URL: https://web.archive.org/web/20201112024633/https://monetizepros.com/features/analysis-how-wikipedia-could-make-2-8-billion-in-annual-revenue/.

[88] Martin Josifoski, Ivan S. Paskov, Hristo S. Paskov, Martin Jaggi, and Robert West. “Crosslingual document embedding as reduced-rank ridge regression”. In: *Proc. International Conference on Web Search and Data Mining (WSDM)*. 2019.

[89] Brian Keegan, Darren Gergle, and Noshir Contractor. “Hot off the wiki: dynamics, practices, and structures in Wikipedia’s coverage of the Tōhoku catastrophes”. In: *Proc. WikiSym - International Symposium on Wikis and Open Collaboration*. 2011.

[90] Faten Khalil, Jiuyong Li, and Hua Wang. “An integrated model for next page access prediction”. In: *International Journal of Knowledge and Web Intelligence* 1.1-2 (2009), pp. 48–80.

[91] Muneo Kitajima, Marilyn H Blackmon, and Peter G Polson. “A comprehension-based model of web navigation and its application to web usability analysis”. In: *People and computers XIV—Usability or else!* Springer, 2000, pp. 357–373.

[92] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. “End-to-end neural entity linking”. In: *Proc. Conference on Computational Natural Language Learning* (2018).

[93] Tobias Koopmann, Alexander Dallmann, Lena Hettinger, Thomas Niebler, and Andreas Hotho. “On the Right Track! Analysing and Predicting Navigation Success in Wikipedia”. In: *Proc. Conference on Hypertext and Social Media (HT)*. 2019.

[94] Yehuda Koren, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems". In: *Computer* 42.8 (2009), pp. 30–37.

[95] Kayvan Kousha and Mike Thelwall. "Are Wikipedia citations important evidence of the impact of scholarly articles and books?" en. In: *Journal of the Association for Information Science and Technology* 68.3 (2017), pp. 762–779.

[96] Sean Kross, Eszter Hargittai, and Elissa M Redmiles. "Characterizing the Online Learning Landscape: What and How People Learn Online". In: *Proc. Human-Computer Interaction (HCI)* 5.CSCW1 (2021), pp. 1–19.

[97] Juhi Kulshrestha, Marcos Oliveira, Orkut Karacalik, Denis Bonnay, and Claudia Wagner. "Web Routineness and Limits of Predictability: Investigating Demographic and Behavioral Differences Using Web Tracking Data". In: *arXiv preprint arXiv:2012.15112* (2020).

[98] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Dandapani Sivakumar, Andrew Tompkins, and Eli Upfal. "The Web as a graph". In: *Proc. ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2000.

[99] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. "Trawling the web for emerging cyber-communities". In: *Computer networks* 31.11-16 (1999), pp. 1481–1493.

[100] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. "Measuring user engagement". In: *Synthesis lectures on information concepts, retrieval, and services* 6.4 (2014), pp. 1–132.

[101] Daniel Lamprecht, Dimitar Dimitrov, Denis Helic, and Markus Strohmaier. "Evaluating and Improving Navigability of Wikipedia: A Comparative Study of Eight Language Editions". In: *Proc. International Symposium on Open Collaboration (OpenSym)*. 2016.

[102] Daniel Lamprecht, Kristina Lerman, Denis Helic, and Markus Strohmaier. "How the structure of Wikipedia articles influences user navigation". In: *New Review of Hypermedia and Multimedia* 23.1 (2017), pp. 29–50.

[103] Paul Laufer, Claudia Wagner, Fabian Flöck, and Markus Strohmaier. "Mining cross-cultural relations from Wikipedia: A study of 31 European food cultures". In: *Proc. ACM Web Science Conference*. 2015.

[104] Michaël R Laurent and Tim J Vickers. "Seeking health information online: does Wikipedia matter?" In: *Journal of the American Medical Informatics Association* 16.4 (2009), pp. 471–479.

[105] David Lazer, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. "Meaningful measures of human society in the twenty-first century". In: *Nature* 595.7866 (2021), pp. 189–196.

[106] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. "The parable of Google Flu: traps in big data analysis". In: *science* 343.6176 (2014), pp. 1203–1205.

[107] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. “Reader Preferences and Behavior on Wikipedia”. In: *Proc. Conference on Hypertext and Social Media (HT)*. 2014.

[108] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. “Why the world reads Wikipedia: Beyond English speakers”. In: *Proc. International Conference on Web Search and Data Mining (WSDM)*. 2019.

[109] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. “Analysis of references across Wikipedia languages”. In: *Information and Software Technologies* 756 (2017), pp. 561–573.

[110] Xiaoliang Ling, Weiwei Deng, Chen Gu, Hucheng Zhou, Cui Li, and Feng Sun. “Model ensemble for click prediction in Bing search ads”. In: *Proc. International World Wide Web Conference (WWW)*. 2017.

[111] Chao Liu, Ryen W White, and Susan Dumais. “Understanding web browsing behaviors through Weibull analysis of dwell time”. In: *Proc. Conference on Research & Development in Information Retrieval (SIGIR)*. 2010.

[112] David M Lydon-Staley, Dale Zhou, Ann Sizemore Blevins, Perry Zurn, and Danielle S Bassett. “Hunters, busybodies and the knowledge network building associated with deprivation curiosity”. In: *Nature Human Behaviour* 5.3 (2021), pp. 327–336.

[113] Laurens van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[114] Nizar R Mabroukeh and Christie I Ezeife. “Semantic-rich markov models for web prefetching”. In: *Proc. International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2009, pp. 465–470.

[115] Fritz Machlup. “The study of information: Interdisciplinary messages”. In: (1983).

[116] Lauren A Maggio, John M Willinsky, Ryan M Steinberg, Daniel Mietchen, Joseph L Wass, and Ting Dong. “Wikipedia as a gateway to biomedical research: The relative distribution and use of citations in the English Wikipedia”. In: *PLOS ONE* 12.12 (2019).

[117] M Mangel, WH Satterthwaite, P Pirolli, B Suh, and Y Zhang. “Invasion biology and the success of social collaboration networks, with application to Wikipedia”. In: *Israel Journal of Ecology and Evolution* 59.1 (2013), pp. 17–26.

[118] Paolo Massa and Federico Scrinzi. “Manypedia: Comparing language points of view of Wikipedia communities”. In: *Proc. International Symposium on Open Collaboration (OpenSym)*. 2012.

[119] Gene McKenna. “Experiment shows up to 60% of “direct” traffic is actually organic search”. In: *Search Engine Land* (Sept. 2014). URL: https://web.archive.org/web/20201020025709/https://searchengineland.com/60-direct-traffic-actually-seo-195415.

[120] Connor McMahon, Isaac Johnson, and Brent Hecht. "The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies". In: 2017.

[121] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Nielsen, and Arto Lanamäki. ""The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia". en. In: *Journal of the Association for Information Science and Technology* 66.2 (2015), pp. 219–245.

[122] Rada Mihalcea and Andras Csomai. "Wikify! Linking Documents to Encyclopedic Knowledge". In: *Proc. Conference on Information and Knowledge Management (CIKM)*. 2007.

[123] David Milne and Ian H. Witten. "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links". In: *Proc. AAAI Workshop on Wikipedia and Artificial Intelligence*. 2008.

[124] David Milne and Ian H. Witten. "Learning to link with Wikipedia". In: *Proc. Conference on Information and Knowledge Management (CIKM)*. 2008.

[125] Blagoj Mitrevski, Tiziano Piccardi, and Robert West. "WikiHist. html: English Wikipedia's Full Revision History in HTML Format". In: *Proc. International Conference on Web and Social Media (ICWSM)*. 2020.

[126] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. "Quantifying Wikipedia usage patterns before stock market moves". en. In: *Scientific Reports* 3.1 (2013).

[127] Helen Susannah Moat, Chester Curme, H. Eugene Stanley, and Tobias Preis. "Anticipating stock market movements with Google and Wikipedia". In: *NATO Science for Peace and Security Series C: Environmental Security* (2014). Ed. by Davron Matrasulov and H. Eugene Stanley, pp. 47–59.

[128] Jonathan Morgan and Isaac Johnson. *Social media traffic report pilot*. https://meta.wikimedia.org/wiki/Research:Social_media_traffic_report_pilot. 2020.

[129] Daniel Cheng Moyer, Samuel L Carson, Thayne Keegan Dye, Richard T Carson, and David Goldbaum. "Determining the influence of Reddit posts on Wikipedia pageviews". In: *Proc. International Conference on Web and Social Media (ICWSM)*. 2015.

[130] Jack Muramatsu and Wanda Pratt. "Transparent Queries: investigation users' mental models of search engines". In: *Proc. Conference on Research & Development in Information Retrieval (SIGIR)*. 2001.

[131] Meera Narvekar and Shaikh Sakina Banu. "Predicting user's web navigation behavior using hybrid approach". In: *Procedia Computer Science* 45 (2015), pp. 3–12.

[132] Finn Årup Nielsen. "Scientific Citations in Wikipedia". In: *First Monday* 12 (2007).

[133] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. "Scholia, scientometrics and Wikidata". In: *European Semantic Web Conference*. Springer. 2017, pp. 237–259.

[134] Richard E Nisbett and Timothy D Wilson. "Telling more than we can know: Verbal reports on mental processes." In: *Psychological review* 84.3 (1977), p. 231.

[135] Thanapon Noraset, Chandra Bhagavatula, and Doug Downey. "Adding high-precision links to Wikipedia". In: *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[136] Heather L O'Brien and Elaine G Toms. "What is user engagement? A conceptual framework for defining user engagement with technology". In: *Journal of the American society for Information Science and Technology* 59.6 (2008), pp. 938–955.

[137] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. "Social data: Biases, methodological pitfalls, and ethical boundaries". In: *Frontiers in Big Data* 2 (2019), p. 13.

[138] Ashwin Paranjape, Robert West, Leila Zia, and Jure Leskovec. "Improving Website Hyperlink Structure Using Server Logs". In: *Proc. International Conference on Web Search and Data Mining (WSDM)*. 2016.

[139] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[140] Tiziano Piccardi, Michele Catasta, Leila Zia, and Robert West. "Structuring Wikipedia articles with section recommendations". In: *Proc. Conference on Research & Development in Information Retrieval (SIGIR)*. 2018.

[141] Tiziano Piccardi, Martin Gerlach, Akhil Arora, and Robert West. "Going down the Wikipedia Rabbit Hole: Characterizing the Long Tail of Reading Sessions". In: *Proc. International World Wide Web Conference (WWW) - Companion*. 2022.

[142] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. "On the Value of Wikipedia as a Gateway to the Web". In: *Proc. International World Wide Web Conference (WWW)*. 2021.

[143] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. "Quantifying engagement with citations on Wikipedia". In: *Proc. International World Wide Web Conference (WWW)*. 2020.

[144] Tiziano Piccardi and Robert West. "Crosslingual Topic Modeling with WikiPDA". In: *Proc. International World Wide Web Conference (WWW)*. 2021.

[145] Peter Pirolli and Stuart Card. "Information foraging." In: *Psychological review* 106.4 (1999), p. 643.

[146] Peter L T Pirolli and James E Pitkow. "Distributions of surfers' paths through the World Wide Web: Empirical characterizations". In: *World Wide Web* 2.1 (1999), pp. 29–45.

[147] Alessandro Piscopo and Elena Simperl. "What we talk about when we talk about Wikidata quality: a literature survey". In: *Proc. International Symposium on Open Collaboration (OpenSym)*. 2019.

[148] Simone Paolo Ponzetto, Michael Strube, *et al.* "Deriving a large scale taxonomy from Wikipedia". In: *AAAI*. Vol. 7. 2007, pp. 1440–1445.

[149] Simone Paolo Ponzetto and Michael Strube. "Taxonomy induction based on a collaboratively built knowledge repository". In: *Artificial Intelligence* 175.9-10 (2011), pp. 1737–1756.

[150] Jamie Powell. "Is GMO's Montier right on 'absurd' US stocks?" In: *Financial Times* (Aug. 2020). URL: https://www.ft.com/content/6ad966ed-17c8-484f-85b3-f685c8d28b13.

[151] Daniele Rama, Tiziano Piccardi, Miriam Redi, and Rossano Schifanella. "A Large Scale Study of Reader Interactions with Images on Wikipedia". In: *EPJ Data Science* (2021).

[152] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. "Characterizing and Modeling the Dynamics of Online Popularity". In: *Physical Review Letters* 105.15 (2010).

[153] Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. "Citation Needed: A Taxonomy and algorithmic assessment of Wikipedia's verifiability". In: *Proc. International World Wide Web Conference (WWW)*. 2019.

[154] Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. "A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft)". In: (Aug. 2020). arXiv: 2008.12314 [cs.CY].

[155] Antonio J Reinoso, Jesus M Gonzalez-Barahona, Gregorio Robles, and Felipe Ortega. "A quantitative approach to the use of the Wikipedia". In: *2009 IEEE Symposium on Computers and Communications*. IEEE. 2009, pp. 56–61.

[156] Antonio J Reinoso, Rocıo Munoz-Mansilla, Israel Herraiz, and Felipe Ortega. "Characterization of the Wikipedia traffic". In: *ICIW 2012: Seventh International Conference on Internet and Web Applications and Services*. 2012, pp. 156–162.

[157] Manoel Horta Ribeiro, Kristina Gligorić, Maxime Peyrard, Florian Lemmerich, Markus Strohmaier, and Robert West. "Sudden attention shifts on wikipedia during the COVID-19 crisis". In: *Proc. International Conference on Web and Social Media (ICWSM)*, 2021.

[158] Giovanna Chiara Rodi, Vittorio Loreto, and Francesca Tria. "Search strategies of Wikipedia readers". In: *PLOS ONE* 12.2 (Feb. 2017), pp. 1–15.

[159] Dana Rotman, Sarah Vieweg, Sarita Yardi, Ed Chi, Jenny Preece, Ben Shneiderman, Peter Pirolli, and Tom Glaisyer. "From slacktivism to activism: participatory culture in the age of social media". In: *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 2011.

[160] Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. "A topic-aligned multilingual corpus of Wikipedia articles for studying information asymmetry in low Resource languages". In: *Proc. Language Resources and Evaluation Conference*. 2020.

[161] Matthew J Salganik. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

[162] Aju Thalappillil Scaria, Rose Marie Philip, Robert West, and Jure Leskovec. "The Last Click: Why Users Give up Information Network Navigation". In: *Proc. International Conference on Web Search and Data Mining (WSDM)*. 2014.

[163] Shigehiko Schamoni, Julian Hitschler, and Stefan Riezler. "A Dataset and Reranking Method for Multimodal MT of User-Generated Image Captions". In: *Proc. Conference of the Association for Machine Translation in the Americas*. 2018.

[164] Michael Scholz, Christoph Brenner, and Oliver Hinz. "AKEGIS: Automatic keyword generation for sponsored search advertising in online retailing". In: *Decision Support Systems* 119 (2019), pp. 96–106.

[165] Thomas Shafee, Gwinyai Masukume, Lisa Kipersztok, Diptanshu Das, Mikael Häggström, and James Heilman. "Evolution of Wikipedia's medical content: past, present and future". In: *Journal of Epidemiology and Community Health* (Aug. 2017), jech–2016–208601.

[166] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning". In: *Proc. Annual Meeting of the Association for Computational Linguistics*. 2018.

[167] Aaron Shaw and Eszter Hargittai. "The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing". In: *The Journal of communication* 68.1 (2018).

[168] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. "Predicting economic development using geolocated Wikipedia articles". In: *Proc. KDD*. 2019.

[169] Wei Shen, Jianyong Wang, and Jiawei Han. "Entity linking with a knowledge base: Issues, techniques, and solutions". In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2014), pp. 443–460.

[170] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. "Detecting memory and structure in human navigation patterns using Markov chain models of varying order". en. In: *PLoS ONE* 9.7 (2014), e102070.

[171] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. "Why we read Wikipedia". In: *Proc. International World Wide Web Conference (WWW)*. 2017.

[172] Philipp Singer, Thomas Niebler, Markus Strohmaier, and Andreas Hotho. "Computing Semantic Relatedness from Human Navigational Paths: A Case Study on Wikipedia". In: *Int. J. Semant. Web Inf. Syst.* 9.4 (Oct. 2013), pp. 41–70.

[173] Harshdeep Singh, Robert West, and Giovanni Colavizza. "Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia". In: *Quantitative Science Studies* 2.1 (2021), pp. 1–19.

[174] Yang Song, Xiaolin Shi, and Xin Fu. "Evaluating and predicting user engagement change with degraded search relevance". In: *Proc. International World Wide Web Conference (WWW)*. 2013.

[175] A. Spoerri. “What is popular on Wikipedia and why?” In: *First Monday* 12.4 (2007). URL: https://firstmonday.org/ojs/index.php/fm/article/view/1765/1645.

[176] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. “Yago: a core of semantic knowledge”. In: *Proc. International World Wide Web Conference (WWW)*. 2007.

[177] Bongwon Suh, Gregorio Convertino, Ed H Chi, and Peter Pirolli. “The singularity is not near: slowing growth of Wikipedia”. In: *Proc. International Symposium on Open Collaboration (OpenSym)*. 2009, pp. 1–10.

[178] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network”. In: *2010 IEEE second international conference on social computing*. IEEE. 2010, pp. 177–184.

[179] Jian Tang, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei, and Ming Zhang. “Understanding the limiting factors of topic modeling via posterior contraction analysis”. In: *Proc. International Conference on Machine Learning*. 2014.

[180] Linda Tauscher and Saul Greenberg. “How people revisit web pages: Empirical findings and implications for the design of history systems”. In: *International Journal of Human-Computer Studies* 47.1 (1997), pp. 97–137.

[181] Linda Tauscher and Saul Greenberg. “Revisitation Patterns in World Wide Web Navigation”. In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 1997.

[182] Nathan TeBlunthuis, Tilman Bayer, and Olga Vasileva. “Dwelling on Wikipedia: Investigating Time Spent by Global Encyclopedia Readers”. In: *Proc. International Symposium on Open Collaboration (OpenSym)*. 2019.

[183] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael AS Potts. “Information re-retrieval: Repeat queries in Yahoo’s logs”. In: *Proc. Conference on Research & Development in Information Retrieval (SIGIR)*. 2007, pp. 151–158.

[184] Misha Teplitskiy, Grace Lu, and Eamon Duede. “Amplifying the impact of open access: Wikipedia and the diffusion of science”. In: *Journal of the Association for Information Science and Technology* 68.9 (2017), pp. 2116–2127.

[185] *The Wiki Game - Explore Wikipedia!* https://www.thewikigame.com/.

[186] Neil Thompson and Douglas Hanley. “Science is shaped by Wikipedia: evidence from a randomized control trial”. In: *MIT Sloan Research Paper* 5238.17 (2018).

[187] Michele Tizzoni, André Panisson, Daniela Paolotti, and Ciro Cattuto. “The impact of news exposure on collective attention in the United States during the 2016 Zika epidemic”. en. In: *PLoS computational biology* 16.3 (Mar. 2020), e1007633.

[188] Daniel Torres-Salinas, Esteban Romero-Frías, and Wenceslao Arroyo-Machado. “Mapping the backbone of the humanities through the eyes of Wikipedia”. In: *Journal of Informetrics* 13.3 (2019), pp. 793–803.

[189] Christoph Trattner, Denis Helic, Philipp Singer, and Markus Strohmaier. “Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks”. In: *Proc. International Conference on Knowledge Management and Knowledge Technologies*. 2012.

[190] Sarah K Tyler and Jaime Teevan. “Large scale query log analysis of re-finding”. In: *Proc. International Conference on Web Search and Data Mining (WSDM)*. 2010.

[191] Nicholas Vincent and Brent Hecht. “A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021), pp. 1–15.

[192] Nicholas Vincent, Isaac Johnson, and Brent Hecht. “Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia’s relationships with other large-scale online communities”. In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2018.

[193] Luis Von Ahn. “Games with a purpose”. In: *Computer* 39.6 (2006), pp. 92–94.

[194] Denny Vrandečić and Markus Krötzsch. “Wikidata: a free collaborative knowledgebase”. In: *Communications of the ACM* 57.10 (2014), pp. 78–85.

[195] Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. “Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications”. In: *Information Processing & Management* 51.1 (2015), pp. 111–147.

[196] Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. “Measuring algorithmically infused societies”. In: *Nature* 595.7866 (2021), pp. 197–204.

[197] Vivienne Waller. “The search queries that took Australian Internet users to Wikipedia”. en. In: *Information Research* 16.3 (2011).

[198] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. “A Survey on Session-based Recommender Systems”. In: *ACM Comput. Surv.* 54.7 (July 2021), pp. 1–38.

[199] Robert West and Jure Leskovec. “Automatic Versus Human Navigation in Information Networks”. In: *Proc. International Conference on Web and Social Media (ICWSM)* (2012).

[200] Robert West and Jure Leskovec. “Human Wayfinding in Information Networks”. In: *Proc. International World Wide Web Conference (WWW)*. 2012.

[201] Robert West, Ashwin Paranjape, and Jure Leskovec. “Mining Missing Hyperlinks from Human Navigation Traces: A Case Study of Wikipedia”. In: 2015.

[202] Robert West, Joelle Pineau, and Doina Precup. “Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts”. In: *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*. 2009.

[203] Robert West, Doina Precup, and Joelle Pineau. “Automatically suggesting topics for augmenting text documents”. In: *Proc. Conference on Information and Knowledge Management (CIKM)*. 2010.

[204] Robert West, Doina Precup, and Joelle Pineau. "Completing Wikipedia's hyperlink structure through dimensionality reduction". In: *Proc. Conference on Information and Knowledge Management (CIKM)*. 2009.

[205] Ryen W White and Steven M Drucker. "Investigating behavioral variability in web search". In: *Proc. International World Wide Web Conference (WWW)*. 2007, pp. 21–30.

[206] Tom D Wilson. "Information behaviour: an interdisciplinary perspective". In: *Information processing & management* 33.4 (1997), pp. 551–572.

[207] Tom D Wilson. "Models in information behaviour research". In: *Journal of documentation* (1999).

[208] Tom D Wilson. "On user studies and information needs". In: *Journal of documentation* (1981).

[209] Rachel Withers. "Amazon owes Wikipedia big-time". In: *Slate* (Oct. 2018). URL: https://slate.com/technology/2018/10/amazon-echo-wikipedia-wikimedia-donation.html.

[210] Ellery Wulczyn and Dario Taraborelli. *Wikipedia clickstream*. https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream. 2015.

[211] Sean Xin Xu and Xiaoquan (Michael) Zhang. "Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction". In: *MIS Quarterly* 37.4 (2013), pp. 1043–1068.

[212] Xing Yi. *Dwell time based advertising in a scrollable content stream*. US Patent App. 13/975,157. Feb. 2015.

[213] Mitsuo Yoshida, Yuki Arase, Takaaki Tsunoda, and Mikio Yamamoto. "Wikipedia page view reflects web search trend". In: *Proc. ACM Web Science Conference*. 2015, pp. 1–2.

[214] Paula Younger. "Internet-based information-seeking behaviour amongst doctors and nurses: a short review of the literature". In: *Health Information & Libraries Journal* 27.1 (2010), pp. 2–10.

[215] Olga Zagovora, Roberto Ulloa, Katrin Weller, and Fabian Flöck. "'I Updated the <ref>': The evolution of references in the English Wikipedia and the implications for altmetrics". In: *Quantitative Science Studies* (2020), pp. 1–38.

[216] Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. "Cross-lingual latent topic extraction". In: *Proc. Annual Meeting of the Association for Computational Linguistics*. 2010.

[217] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. "How well do offline and online evaluation metrics measure user satisfaction in Web image search?" In: *Proc. Conference on Research & Development in Information Retrieval (SIGIR)*. 2018.

[218] Kai Zhu, Dylan Walker, and Lev Muchnik. "Content Growth and Attention Contagion in Information Networks: Addressing Information Poverty on Wikipedia". In: *Information Systems Research* 31.2 (June 2020), pp. 491–509.