Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions

Tiziano Piccardi
EPFL
tiziano.piccardi@epfl.ch

Martin Gerlach
Wikimedia Foundation
mgerlach@wikimedia.org

Robert West
EPFL
robert.west@epfl.ch

ABSTRACT

"Wiki rabbit holes" are informally defined as navigation paths followed by Wikipedia readers that lead them to long explorations, sometimes involving unexpected articles. Although wiki rabbit holes are a popular concept in Internet culture, our current understanding of their dynamics is based on anecdotal reports only. To bridge this gap, this paper provides a large-scale quantitative characterization of the navigation traces of readers who fell into a wiki rabbit hole. First, we represent user sessions as navigation trees and operationalize the concept of wiki rabbit holes based on the depth of these trees. Then, we characterize rabbit hole sessions in terms of structural patterns, time properties, and topical exploration. We find that article layout influences the structure of rabbit hole sessions and that the fraction of rabbit hole sessions is higher during the night. Moreover, readers are more likely to fall into a rabbit hole starting from articles about entertainment, sports, politics, and history. Finally, we observe that, on average, readers tend to stay focused on one topic by remaining in the semantic neighborhood of the first articles even during rabbit hole sessions. These findings contribute to our understanding of Wikipedia readers' information needs and user behavior on the Web.

ACM Reference Format:
Tiziano Piccardi, Martin Gerlach, and Robert West. 2022. Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3487553.3524930

1 INTRODUCTION

If you ever visited Wikipedia for simple fact-checking and then, in a few minutes, found yourself learning that *life does not evolve wheels*[1] and that *unsold video games may end up buried in the desert*[2], there is a high chance that you fell into a "wiki rabbit hole", a term inspired by Lewis Carroll's novel *Alice's Adventures in Wonderland,* where the main character reaches an astonishing world by following a white rabbit deep into its burrow.

Similarly, a wiki rabbit hole, sometimes called a wiki hole or wiki black hole[3], is a popular concept in Internet culture often described as a long navigation session where readers, following multiple links, get lost in Wikipedia and learn about a diverse set of topics. The reason that motivates readers to engage in long navigation sessions, with jumps across different topics, is often associated with boredom [22] and a busybody-style of curiosity [13].

Given the substantial time we spend consuming online content, understanding the dynamics of how we seek knowledge can offer useful insight into our information needs and support the design of better systems centered around users' interests. Previous work provided an overall characterization of reading sessions on Wikipedia [17]. However, reading sessions are typically short (78% consisting of a single pageload) such that the population-wide average does not adequately capture the behavior contained in the long tail. Therefore, this study serves as a complimentary analysis of how readers browse Wikipedia by focusing on long reading sessions associated with rabbit hole navigation.

This paper characterizes the long tail of these navigation traces by focusing on sessions with paths generated by at least ten sequential internal clicks. We use server logs collected for one month from English Wikipedia to describe the exploration dynamics and how readers navigate during the long reading sessions. We characterize the paths of readers who engage in extended navigation by describing common *structural properties* such as the shape of the navigation traces, *temporal patterns* such as the time spent on Wikipedia and the daily rhythm, and *topic-based characteristics* like the semantic diffusion of the navigation from the origin.

This paper is organized as follows. First, we introduce the related work (Sec. 2) and the data (Sec. 3). Then, we operationalize wiki rabbit holes and describe the characteristics of these sessions in terms of structure (Sec. 4.1), time properties (Sec. 4.2), and topics (Sec. 4.3). Finally (Sec. 5), we summarise the findings and discuss the implications.

2 RELATED WORK

In recent years, readers' behavior on Wikipedia has been explored in different context by characterising interaction with citations [14, 18], external links [19], images [20], and navigation behaviors.

Readers' navigation from article to article received significant attention from researchers using different approaches such as wikigames [25], browser history sharing [13], server logs, and public clickstream [1, 21, 26]. These analyses address two different types of navigation on Wikipedia: natural and targeted navigation.

Natural navigation refers to the typical *out-of-lab* usage of Wikipedia; the digital traces left by the readers are associated with self-motivated learning or to satisfy personal information needs

[1] https://en.wikipedia.org/wiki/Rotating_locomotion_in_living_systems
[2] https://en.wikipedia.org/wiki/Atari_video_game_burial
[3] https://en.wikipedia.org/wiki/Wiki_rabbit_hole

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524930

[22]. Previous studies [2, 17] shows that readers in a natural setup tend to have short sessions, with the vast majority (78%) of the sessions composed by a single pageload. When the navigation goes beyond the first page, an analysis based on the server logs revealed the users' propensity to stop the exploration when they reach a low-quality article. In other investigations, researchers found that Wikipedia articles relay different traffic volumes based on their topics [3] and the type of page [5]. Readers also show preferences for links that appear at the top of the page and are semantically closer to the current article [4, 10]. Reading preferences were shown to fall into 4 types of behaviors described as focus, trending, exploration and passing[11].

In contrast, targeted navigation typically involves instructions for the reader to reach a specific article on Wikipedia. For example, wikigames [25] lets users navigate on a gamified platform to reach a target article in as few clicks as possible starting from a random Wikipedia page. This line of research aims to understand how humans navigate information networks and what strategies they employ to find a predetermined piece of information. Researches found [7, 24] that initially, users tend to jump to high degree nodes that act as navigational hubs, and then they converge to the destination in increasingly small steps in the semantic space. Given the clear definition of success in a wikigame –reaching the target article–, researchers also explored what makes people leave the navigation, discovering that diverging in topic space from the target leads to frustration and giving up.

Finally, a study based on a combination of voluntary sharing of the navigation history and survey-based qualitative methods [13] analyzed curiosity as a driving mechanism for navigation. They found that different curiosity patterns lead to distinct navigation behaviors when looking at the knowledge networks constructed by the readers.

3 DATA

To characterize the sessions of readers that fall into a wiki rabbit hole, we rely on server logs collected over four weeks in March 2021. The data is pre-processed as described in previous work [17], e.g., we removed bots and all the activities of users that were logged in, edited at least one article during the data collection, or accessed Wikipedia from countries with less than 300 pageloads per day in order to preserve readers' privacy. Then, we made the requests from different countries comparable by converting the timestamps in local time and dropped all the sensitive information such as geo-location, user-agent, and IP addresses. To combine requests coming from the same client, we preserve for each request an anonymous user identifier generated from the original user-agent string and the IP address. Since large organizations may have many clients sharing these two properties, some identifiers may actually represent requests coming from different readers. To mitigate this issue, we removed all the activities associated with user identifiers with more than 2800 pageloads in 4 weeks (i.e. more than 100 pageviews per day on average). The resulting dataset comprises 6.25B pageviews generated by 1.47B different user identifiers.

Aggregating the sessions. We aggregate the session into navigation trees as described in previous works [16, 17]. Given the complex navigation patterns of Web users, comprised of multi-tab and backtracking behavior, the structure of the navigation path is typically a tree. To reconstruct sequences of pageviews from individual clicks, we use the HTTP referrer field that allows the browser to specify the origin of each request. First, we use this information to generate *root-only* trees for all the requests coming from URLs that are not Wikipedia articles. Then, for all the loads that originated from internal navigation, we assign each request as a child of the node representing the most recent time the user loaded the source article. The resulting dataset is composed of 3.7B navigation trees.

4 THE WIKI RABBIT HOLE

Wiki rabbit hole operationalization. In this work, we consider as *wiki rabbit hole* sessions with a navigation tree with a minimum depth of ten steps. This constraint means we include all the trees where the longest path root-leaf is at least nine sequential clicks. This definition ensures that large shallow trees where the readers remained around the starting article opening many tabs, in the present work, are not considered rabbit hole sessions. Given the long-tailed nature of the tree size distribution, this filter leaves us with 216M pageviews aggregated in 8.97M trees—0.24% of the original navigation sessions. In comparison, this number is much larger than the roughly 40k active editors per month in English Wikipedia[4], suggesting that the majority of these trees comes from actual readers.

Frequent entry points. In total, 846K articles acted, at least once, as the entry point for a rabbit hole session, i.e. they appear as the root of a tree. The most frequent articles that served as starting points for long sessions are popular pages such as Elizabeth II (37.8K), Deaths in 2021 (23.7K), 2020 United States presidential election (17.6K), Wikipedia (15.8K), and Prince Philip, Duke of Edinburgh (15.0K). These pages are overall very popular and associated with events that received significant attention at the time of the data collection. In total, 11 out of the top 20 articles have some connection with the British Royal family that, in March 2021, received a burst in media attention. Interestingly, the list of the most common pages where rabbit hole sessions start includes the article Wikipedia. A substantial portion of readers use this page as the starting point for navigation, from which they move to the desired destination using the internal search. By inspecting the articles loaded from Wikipedia, it emerges that readers continue the navigation using the blue links available in the body of the page only in 9.3% of the cases.

These pages are observed frequently in this list because of the attention received overall. We observe a different pattern when we normalize the number of times an article acts as the entry point for a rabbit hole by its global popularity. By limiting to pages that received at least 100 pageviews, we find 1980 Arkansas gubernatorial election (34% as rabbit hole entry), It Is the Law (33%), Alexander, Duke of Schleswig-Holstein-Sonderburg (30%), Solidarity Party of Afghanistan (29%), and 2006 Florida gubernatorial election (29%).

[4] https://stats.wikimedia.org/#/en.wikipedia.org/contributing/active-editors/normal|line|2-year|(page_type)~content*non-content|monthly

Tree	Winston Churchill	UFC 260	WandaVision	RuPaul’s Drag Race (season 1)	Elizabeth II
	Anthony Eden	UFC on ABC: Till vs. Vettori	Filmed Before a Live Studio Audience	RuPaul’s Drag Race (season 2)	George VI
	Harold Macmillan	UFC on ESPN: Whittaker vs. Gastelum	Don’t Touch That Dial	RuPaul’s Drag Race (season 3)	Edward VIII
	Alec Douglas-Home	UFC 261	Now in Color	RuPaul’s Drag Race (season 4)	George V
	Harold Wilson	UFC on ESPN: Reyes vs. Procházka	We Interrupt This Program	RuPaul’s Drag Race (season 5)	Edward VII
	James Callaghan	UFC on ESPN: Sandhagen vs. Dillashaw	On a Very Special Episode...	RuPaul’s Drag Race (season 6)	Queen Victoria
	Margaret Thatcher	UFC 262	All-New Halloween Spooktacular!	RuPaul’s Drag Race (season 7)	William IV
	John Major	UFC Fight Night: Font vs. Garbrandt	Breaking the Fourth Wall (WandaVision)	RuPaul’s Drag Race (season 8)	George IV
	Tony Blair	UFC Fight Night 189	Previously On	RuPaul’s Drag Race (season 9)	George III
	Gordon Brown	UFC 263	The Series Finale	RuPaul’s Drag Race (season 10)	George II of Great Britain
	David Cameron	UFC Fight Night 190		RuPaul’s Drag Race (season 11)	George I of Great Britain
	Theresa May	UFC Fight Night 191		RuPaul’s Drag Race (season 12)	Anne, Queen of Great Britain
	Boris Johnson	UFC 264		RuPaul’s Drag Race (season 13)	
		UFC Fight Night 192			
Count	850	555	503	452	406

Table 1: The five most frequent trees result from using navigational links in the infobox and have a chain-like structure.

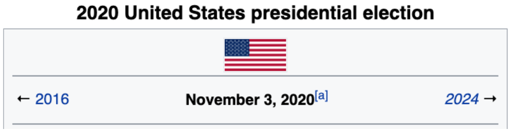


Figure 1: Navigational links in the infobox are often used to engage in long sessions.

Infobox navigational links. A manual inspection of the top 1000 articles that serve most frequently as an entry point for the rabbit hole reveals that many articles refer to recurrent events with seasonal repetitions, such as elections (i.e. 1946 Dutch general election), sports events (i.e. 1979 NBA All-Star Game), and award ceremonies (i.e. 2019 Academy Awards). In total, 68.4% of the articles in this list contain the word “election” in their title, and 87.3% contain a four digits year. When considering the entire dataset, the pages about elections are 0.5% of all the entry-points articles and cover 3.4% of the trees.

By inspecting these navigation traces, we observe that readers often engage in long sessions using the navigational links available in the infobox. Articles about recurrent events typically have links to move to the previous or following occurrence of the same event. For example, a reader that follows this pattern may open the article 2020 United States presidential election and then navigate to 1960 United States presidential election by repeatedly clicking the year of the previous election. Figure 1 shows an example of links used for this type of navigation.

Frequent pathways. English Wikipedia in March 2021 had around 252M links between its articles[5]. Given the large exploration space that the readers can access, it is rare that two sessions generate the same navigation tree. We observe that for the sessions with the rabbit hole pattern, 0.74% of the trees appear more than once, and 0.03% more than ten times. The usage of the navigational links in the infobox typically shapes these navigation traces. Table 1 shows the five most frequent paths with the respective number of occurrences. Given the readers’ behavior that generated these trees, they exhibit a chain-like structure without branching.

A manual inspection of the top 100 common trees shows that the most frequent pathway (90 occurrences) that does not rely on navigational links of the infobox is a path from Egg to Philosophy, following only the first link in the article. This behavior may be caused by readers curious to verify the popular Wikipedia property that following the first link of each page recursively leads to the philosophy article [8, 9, 23].

Frequent exit. Complementary to the common entry points, we look at the last page that a reader loaded in the session. The most frequent article where the users leave the navigation is the article Philosophy which occurs as the last document more than 20K times. This article is then followed by popular content such as Elizabeth II, Joe Biden, and 2020 United States presidential election. To remove the popularity bias, we normalize using the total number of pageloads of the articles and limit to pages loaded at least 100 times in the month of the data collection. Philosophy remains a frequent last article acting as an exit point in 12.9% of the cases, but the rank is dominated by Ichiki Kitokurō (31.1%), Are You In?: Nike+ Original Run (27.9%), What’s Your Favorite Color?: Remixes, B-Sides and Rarities (25.8%). A manual inspection of the top 100 reveals a high presence of historical figures and music albums.

Frequent origin. Thanks to the referrer field of the requests that serve as roots of the trees, we can determine how readers reached Wikipedia. The majority of the rabbit hole sessions (63.8%) start from requests coming from search engines, followed by an unspecified origin (14.1%) that may include searches from toolbars and revisiting patterns where the user picked the article from the browser URL autocomplete. The other traffic sources are the main page (13.4%), the Wikipedia internal search results (4.3%), and Wikipedia in other languages than English (3.3%). External websites –including social media– contribute 0.5% as the traffic source for the rabbit hole sessions.

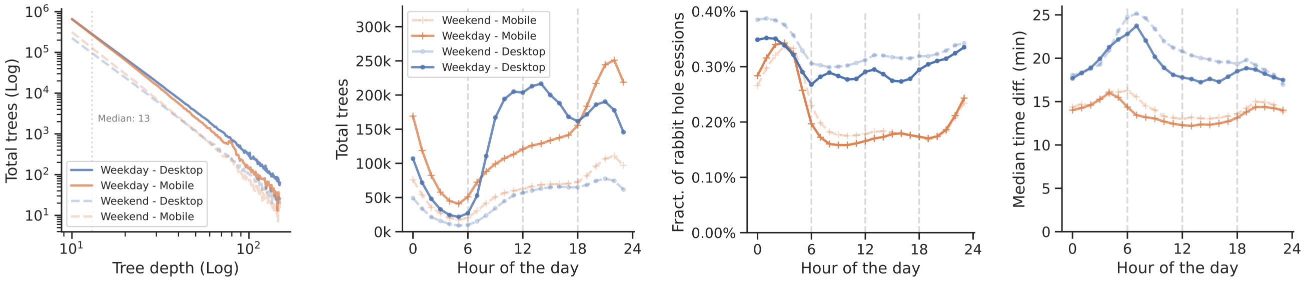
4.1 Structural patterns

As observed in previous work [17], the depth and size distributions of the trees generated by readers exploring Wikipedia show a long tail (approximately a straight line in a log-log plot). The trees that we retained as rabbit hole sessions have a median of 18 pageloads (Q1 = 14, Q3 = 28). We observe a small but significant[6] tendency to have larger trees from mobile (median = 19) compared to desktop (median = 18) and during the weekend (Saturday-Sunday – median = 19) compared to the working days (Monday-Friday – median = 18).

Trees depth. The median depth of the trees, i.e. the longest root-to-leaf path, is 13, meaning that half of the rabbit hole sessions do not extend beyond 12 clicks away from the first page. Figure 2a

[5] When considering the full HTML instead of the wikitext of the article this number rises to more than 400M [15].

[6] Wilcoxon signed-rank test: $p < 0.001$.



(a) Total by depth **(b) Total by time** **(c) Proportion by time** **(d) Time from first to last click**

Figure 2: (2a) Depth distribution of the trees with a rabbit hole pattern. (2b) Total volume of wiki rabbit hole sessions by the time of the day. (2c) Proportion of wiki rabbit hole sessions by the time of the day. (2d) Median time passed from the first to the last pageload of the session. All plots show the distributions broken down by access device and weekday-weekend.

shows the distribution of the tree depths. The depth of the trees does not show any significant difference based on device or day of the week.

Trees sizes. The distribution of the size of navigation trees, i.e. the number of articles loaded in a session, is skewed, with an average of 24 pageloads per session and a median of 18 pageloads (Q1 = 14, Q3 = 28). By limiting to pages that acted at least 100 times as entry points of the rabbit hole, we observe that the articles that generate the largest trees are Eurovision Song Contest 1956 (median = 42), Reference ranges for blood tests (median = 38), List of Major League Soccer transfers 2021 (median = 36), *1st Academy Awards* (median = 35), and Lists of UK top-ten singles (median = 34).

Branching factor. When readers follow multiple links on a page, they generate a fork in the navigation tree. By the trees' construction, this branching can happen when the user moves forward in the exploration by opening multiple browser tabs or backtracking the navigation to a previous point with the back button of the browser. The macro average degree of the trees is 1.36 (median = 1.2), while trees have, on average, a maximal width considering all branches of 3.9 nodes (median = 3). The breakdown by device shows a higher tendency to explore multiple paths from the same article for mobile sessions (average = 1.42, median = 1.26) compared to desktop sessions (average = 1.30, median = 1.15). This observation is reflected in the average maximum width of the tree (mobile: average = 4.20, median = 3 vs. desktop: average = 3.58, median = 2). Overall, 22% of the trees have a chain-like structure with an average branching factor of zero.

4.2 Temporal patterns

Daily pattern. Previous work [17] showed that the readers follow a circadian rhythm with reduced access at night. During the working days, the requests received by English Wikipedia are balanced across desktop and mobile until the evening, when mobile access increases drastically, generating more than double the traffic of desktop devices.

For rabbit hole sessions, we find a different pattern shown in Figure 2b. The absolute number of rabbit hole sessions is significantly higher from desktop during the working daytime, whereas mobile access dominates during evening and night. This inversion between the desktop and mobile is not present during the weekend when mobile traffic remains the most common rabbit hole access method.

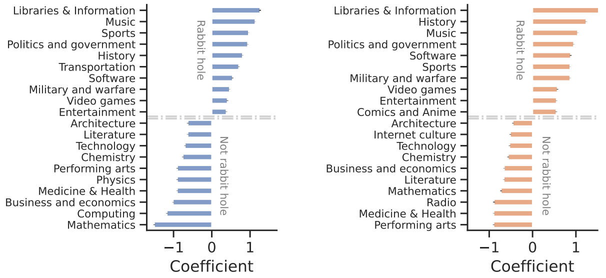
In Figure 2c we show the relative proportion of sessions with a rabbit hole pattern (compared to all sessions) for the two access methods throughout the day. In general, the fraction of rabbit hole sessions is higher at night. On desktop devices, the portion of deep trees is consistently higher, with a further increase during the weekend. For mobile devices, the sessions started during the night (i.e. 0.34% at 3 AM) is double compared to the working hours (i.e. 0.17% at 3 PM).

Time spent in the rabbit hole. We approximate the time readers spent in the rabbit hole, by computing the time difference between the first and last pageload of the navigation trees[7]. The median time is 15 minutes and 49 seconds ($Q1 = 6m42s$, $Q3 = 43m52s$), with differences between desktop and mobile. Figure 2c shows that during the day, the rabbit hole sessions started from desktop devices keep the readers consistently for more time than from mobile. Readers spend in median 5 minutes more in the rabbit hole from desktop (18m42s vs. 13m43s)—this difference is statistically significant[6].

4.3 Topic patterns

Entering the rabbit hole. To comprehend the dynamics that bring readers down the rabbit hole, we use regression analysis to study what topics are associated with the first page of the session. We train a logistic regression to predict if a reader will fall into the rabbit hole by using the properties (i.e. topics) of the first article. We generate the dataset by selecting the first articles of all the rabbit hole sessions as positive samples and an equal number randomly picked from the non-rabbit hole trees as negative ones. This step leaves us with around 18M samples equally distributed between the two classes. Each article is then represented with a vector representing

[7] Typically, the actual time spent reading is higher



(a) Desktop **(b) Mobile**

Figure 3: Coefficients of the logistic regressions that predict if a reader will end up in a wiki rabbit hole given the topics of the first page.

the topics probabilities obtained from ORES[8] [6]. This sampling approach takes into account the popularity of the topics, ensuring that a highly popular topic is proportionally represented also in the negative class. We train two models independently for the samples generated from desktop and mobile devices, obtaining an AUC on a testing set respectively of 0.61 and 0.62.

Figure 3 summarises the coefficients of the topics most associated positively and negatively with the rabbit hole pattern. Similar to previous findings [17], the readers who fall into a wiki rabbit hole typically start from entertainment, sport, politics, and history articles. At the same time, STEM, Medicine, and Business are overall topics where the readers engage less in deep explorations. The coefficients obtained for the regression of the two devices show variation in their ranking, but they offer qualitatively the same conclusions. In order to confirm the robustness of these findings, we use a linear regression model to predict the exact number of articles loaded during a navigation session, which yields qualitatively similar results (not shown).

Diffusion in topic space. Often the navigation of readers falling in the rabbit hole is imagined as a long session that brings the users on a random page of Wikipedia. We verify this assumption by comparing the readers' trajectories in the topics space with a null model obtained from an unbiased random walker. We are interested in observing how the trajectories diffuse in the space with respect to the origin and if the long navigation paths converge, in multiple steps, to random clicking behavior. Since we need trajectories, we extract from the trees the longest path from the root to one of the leaves. If the tree has multiple longest paths of the same length, we select one random from these candidates. In 84% of the cases, the trajectory selected is also the path leading to the last article loaded by the reader.

To compare the two sets of trajectories, we proceed in two steps: first, we create a matched dataset by running for each of the 8.9M readers-generated paths a random walk that, starting from the same article, generates a sequence of the same length. For each step, the next article is selected randomly from the list of links available on the page. Then, we assign the respective ORES topics vectors to each article visited in all trajectories.

Overall, the comparison shows that readers tend to stay semantically close to the first page, even for long sessions. Figure 4b shows the PCA representation of a random sample of trajectories where the first page is centered at the origin. The marginal density distribution shows that the user-generated paths have a higher concentration close to the first page loaded when compared to the larger spread of random exploration. This intuition is reinforced by computing the Mean Squared Displacement (MSD). MSD is typically used in physics to measure the dispersion of a particle from the starting position.

MSD is formally defined[9] as

$$\mathrm{MSD} \equiv \langle |\mathbf{x}(t) - \mathbf{x_0}|^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{x}^{(\mathbf{i})}(t) - \mathbf{x}^{(\mathbf{i})}(0)|^2 \qquad (1)$$

where N is the number of trajectories, vector $\mathbf{x}^{(\mathbf{i})}(t)$ is the position of the navigation i at step t, and vector $\mathbf{x}^{(\mathbf{i})}(0) = \mathbf{x}_{\mathbf{0}}^{(\mathbf{i})}$ is the position of the first article for the trajectory i.

Figure 4c shows that the dispersion coefficients of the readers' navigation, stratified for different tree sizes and positions in the path, are almost half compared to the diffusion of a random walk.

To ensure these findings are not skewed by the sessions generated from using the navigational links of the inboxes, we repeated this analysis by removing all the trees with a chain-like structure. We observe no variations in the patterns of Figure 4, concluding that readers tend to stay semantically close to the origin –compared to a random walk– even for very long sessions.

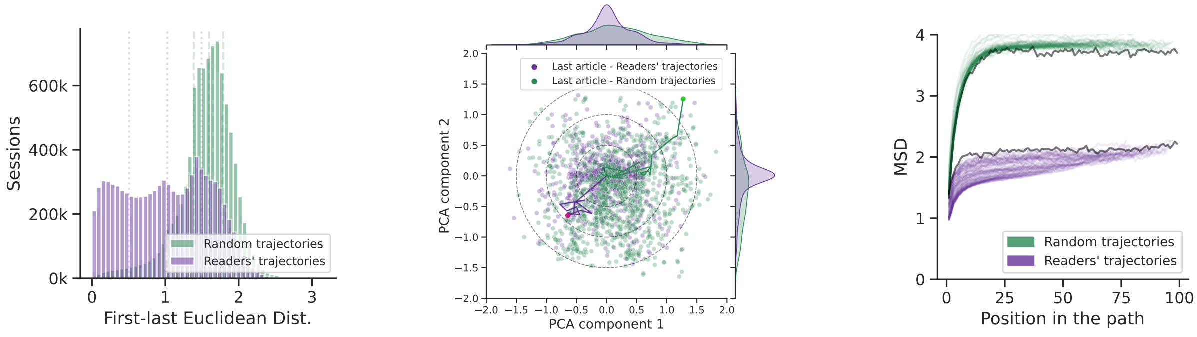
When navigation approaches random. Figure 4a shows that a portion of the sessions that reach semantic locations comparable to a random walk on the links network. We focus on this portion of sessions by selecting the last quartile (last 25% – 2.2M paths) of the distribution of the Euclidean distances from the starting page. These trajectories represent the session of readers that, in the longest path, loaded an article semantically very far from the first page.

Figure 5a shows the mean square displacement of these sessions that reveals that they are close to the trajectories generated by a random walk. Interestingly, the readers' trajectories show a final steep increase in the distance from the origin, suggesting that the readers abandon the path exploration after a fast drift from the first page. This finding represents an extreme case of the behavior already observed in previous work [17].

MSD measures the diffusion from the original, but it does not capture the relative distance between two sequential pageloads. Figure 5b shows how the sessions evolve. The average Euclidean distances for the consecutive pairs of articles show an initial drop indicating that readers in these special sessions, on average, jump far from the origin in the first step and then tend to move with smaller semantic jumps. Differently from the general behavior where readers favor, as next step, articles semantically close [4, 10, 17], the average distance between two consecutive articles in these special sessions tends to approximate the semantic jumps of a random walk (horizontal line). Additionally, the trajectories show a difference in behavior based on the length of the path. Figure 5b shows that

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

[9] https://en.wikipedia.org/wiki/Mean_squared_displacement

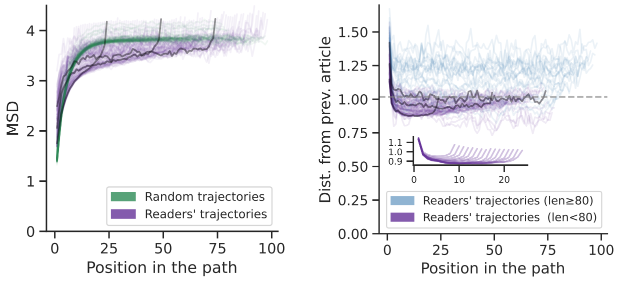


(a) First-to-last distance

(b) PCA projection of navigation trajectories

(c) Mean squared displacement

Figure 4: Sessions diffusion in topics space. (4a): Distribution of Euclidean distances between the first and the last article of the path. Quartiles as vertical lines. (4b): First 2 principal components of the last article of 2000 paths generated by human readers and a random walker in the topic space defined by ORES. The first article of the session centered on the origin. Marginal plots show KDE distributions. The green and purple lines represent two examples of full trajectories. (4c): Mean squared displacement (MSD) of the sessions in the topic space defined by ORES. Each line represents the MSD of all the sessions of one specific length [10-100]. The dark trajectories are added for readability and represent the sessions with 100 pageviews.



(a) Mean squared displacement

(b) Distance from previous

Figure 5: (Figure 5a) Mean squared displacement of the sessions with the largest distance from the first to the last page. (Figure 5b) Euclidean distance from the previous article in the function of the position in the path—divided into two length groups. The horizontal line represents the average distance from the previous article in the case of a random walk. Inset: zoom on the paths with a max of 25 pageloads. Each line represents the sessions of one specific length [10-100]. Dark lines are added for readability, and they represent the trajectories of paths of length 25, 50, and 75.

very long sessions —more than 80 clicks— tend to have more extreme exploration patterns with jumps in semantic space that go further than a random walk. A zoom-in on the paths with less than 25 pageloads shows that the sessions tend to divergence before leaving the exploration, as previously observed in a more general study [17].

Finally, an investigation on time (hour and weekend) and the device used show no significant differences from the other rabbit hole sessions.

5 DISCUSSION

In this paper, we have provided a first data-driven investigation on the patterns associated with long reading sessions in Wikipedia, also known as rabbit hole navigation.

Summary of findings. By investigating the frequent paths taken by readers falling into a wiki rabbit hole, we observed that characteristics of the article layout are associated with deep navigation trees. The presence of navigational links in the infobox to transition between different instances of the same recurrent event supports the type browsing similar to reading a slideshow.

The dynamics of falling into a wiki rabbit hole show differences across time of the day, the device used, and the topic of the first article. The fraction of sessions with deep trees is overall higher on desktop than mobile devices and increases in both cases at night. Confirming popular belief and previous findings on more general behavior [17], articles about entertainment, sport, politics, and history are more common as starting points for rabbit hole sessions.

An investigation of the diffusion in topic space shows that rabbit hole sessions consist of topically coherent articles. On average, even after long sequences of clicks, the visited article is topically much closer to the starting article than compared to a randomly chosen page.

While most reading sessions are quite short [17], here, we find that rabbit hole sessions still constitute a substantial number of sessions of readers in absolute terms – far exceeding the number of active editors in Wikipedia. Overall, we show that rabbit hole sessions can exhibit distinctly different patterns in comparison to the average of all sessions which are dominated by the large number of short sessions.

Limitations and future work. This study can spark future work to comprehend the knowledge consumption patterns and to inform the organization of Wikipedia's content. One important limitation of the present work is that defining the rule to identify a rabbit hole session is not a trivial task. The rabbit hole concept is mainly based on anecdotal examples, and what constitutes a rabbit hole session may be subjective. We focus on the trees with at least a depth of 10 nodes, but other approaches, for example, could employ methods based on the divergence of the reader from the origin or on time spent reading articles, regardless of the number of steps. Another future aspect worth exploring is the diversity of behavior based on geographical and cultural factors. Our analysis is limited to the English edition of Wikipedia, but we are interested in extending it to multiple languages. Additionally, further interactions with the page, such as previews and dwelling time, can be taken into account to paint a more complete picture of the readers' navigation in the rabbit hole. Finally, these findings can be used to better serve readers' needs. For example, the chain-like navigation using the links in the infobox could suggest either i) a desire of the readers to have a more complete overview on a set of articles, or ii) difficulty in finding the article containing relevant content through traditional text-based search.

Conclusion. Readers visiting Wikipedia cover a wide spectrum in terms of their motivations, needs, and prior knowledge [12]. In this study, we focused on a specific setting in which readers embark on in-depth navigation in Wikipedia, i.e. rabbit hole sessions. The characteristics of these sessions differ from the majority of very short sessions suggesting that rabbit hole sessions satisfy succinctly distinct needs of readers. This work thus provides new quantitative insights into how Wikipedia is used by readers which could empower the community to make informed decisions around the organization of Wikipedia's content. More generally, we hope to inspire future research on online knowledge consumption and add a small piece to our understanding of Wikipedia readership.

ACKNOWLEDGMENTS

This project was partly funded by the Swiss National Science Foundation (grant 200021_185043), the European Union (TAILOR, grant 952215), and the Microsoft Swiss Joint Research Center. We also gratefully acknowledge generous gifts from Facebook and Google supporting West's lab.

REFERENCES

[1] Akhil Arora, Martin Gerlach, Tiziano Piccardi, Alberto García-Durán, and Robert West. 2022. Wikipedia Reader Navigation: When Synthetic Data Is Enough. *arXiv preprint arXiv:2201.00812* (2022).

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

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

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

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

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

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

[8] Mark Ibrahim, Christopher M Danforth, and Peter Sheridan Dodds. 2017. Connecting every bit of knowledge: The structure of Wikipedia's First Link Network. *Journal of Computational Science* 19 (2017), 21–30.

[9] Daniel Lamprecht, Dimitar Dimitrov, Denis Helic, and Markus Strohmaier. 2016. Evaluating and improving navigability of Wikipedia: a comparative study of eight language editions. In *Proceedings of the 12th international symposium on open collaboration.* 1–10.

[10] Daniel Lamprecht, Kristina Lerman, Denis Helic, and Markus Strohmaier. 2017. How the structure of Wikipedia articles influences user navigation. *New Review of Hypermedia and Multimedia* 23, 1 (2017), 29–50.

[11] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. 2014. Reader preferences and behavior on Wikipedia. In *Proceedings of the 25th ACM conference on Hypertext and social media.* 88–97.

[12] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19).* ACM, New York, NY, USA, 618–626.

[13] David M Lydon-Staley, Dale Zhou, Ann Sizemore Blevins, Perry Zurn, and Danielle S Bassett. 2021. Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature Human Behaviour* 5, 3 (2021), 327–336.

[14] Lauren A Maggio, Ryan M Steinberg, Tiziano Piccardi, and John M Willinsky. 2020. Meta-Research: Reader engagement with medical content on Wikipedia. *Elife* 9 (2020), e52426.

[15] Blagoj Mitrevski, Tiziano Piccardi, and Robert West. 2020. WikiHist. html: English Wikipedia's Full Revision History in HTML Format. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 878–884.

[16] Ashwin Paranjape, Robert West, Leila Zia, and Jure Leskovec. 2016. Improving Website Hyperlink Structure Using Server Logs. In *Proc. International Conference on Web Search and Data Mining (WSDM).*

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

[18] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2020. Quantifying engagement with citations on Wikipedia. In *Proc. International World Wide Web Conference (WWW).*

[19] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2021. On the Value of Wikipedia as a Gateway to the Web. In *Proc. International World Wide Web Conference (WWW).*

[20] Daniele Rama, Tiziano Piccardi, Miriam Redi, and Rossano Schifanella. 2022. A large scale study of reader interactions with images on Wikipedia. *EPJ Data Science* 11, 1 (2022), 1.

[21] Giovanna Chiara Rodi, Vittorio Loreto, and Francesca Tria. 2017. Search strategies of Wikipedia readers. *PloS one* 12, 2 (2017), e0170746.

[22] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read Wikipedia. In *Proc. International World Wide Web Conference (WWW).*

[23] Robert West. 2011. Wikipedia's fixed point. http://infolab.stanford.edu/~west1/attractor.html.

[24] Robert West and Jure Leskovec. 2012. Human Wayfinding in Information Networks. In *Proc. International World Wide Web Conference (WWW).*

[25] Robert West, Joelle Pineau, and Doina Precup. 2009. Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI).*

[26] Ellery Wulczyn and Dario Taraborelli. 2015. Wikipedia clickstream. https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream.