



How Do Home Computer Users Browse the Web?

KYLE CRICHTON, NICOLAS CHRISTIN, and LORRIE FAITH CRANOR, Carnegie Mellon University

With the ubiquity of web tracking, information on how people navigate the internet is abundantly collected yet, due to its proprietary nature, rarely distributed. As a result, our understanding of user browsing primarily derives from small-scale studies conducted more than a decade ago. To provide an broader updated perspective, we analyze data from 257 participants who consented to have their home computer and browsing behavior monitored through the Security Behavior Observatory. Compared to previous work, we find a substantial increase in tabbed browsing and demonstrate the need to include tab information for accurate web measurements. Our results confirm that user browsing is highly centralized, with 50% of internet use spent on 1% of visited websites. However, we also find that users spend a disproportionate amount of time on low-visited websites, areas with a greater likelihood of containing risky content. We then identify the primary gateways to these sites and discuss implications for future research.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Web-based interaction*; • **Information systems** → World Wide Web;

Additional Key Words and Phrases: Web browsing, user behavior, personal computers, measurement

ACM Reference format:

Kyle Crichton, Nicolas Christin, and Lorrie Faith Cranor. 2021. How Do Home Computer Users Browse the Web?. *ACM Trans. Web* 16, 1, Article 3 (September 2021), 27 pages.
<https://doi.org/10.1145/3473343>

1 INTRODUCTION

It is 7:15 am on a Friday, and Jordan has just finished checking emails on their laptop when an innocent mistake causes serious trouble. Jordan wants to install a new application from a reputable company on their computer. The software is so popular that Jordan knows it by name, or so they think. Unfortunately, they open a new browser tab and mistype the application's name into the address bar, leading Jordan to download the software from an obscure website. One hundred and twenty-six seconds after making that fateful error, Jordan becomes the owner of one newly installed popular application and a host of malicious software add-ons. Fortunately, the story does not end in tragedy as Jordan quickly notices the adware redirecting web traffic to unwanted sites and is able to uninstall the malicious software within the next 20 minutes.

This work was partially funded by the National Security Agency (NSA) Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140). The research was also partially funded through a Carnegie Bosch Institute (CBI) Fellowship.

Authors' address: K. Crichton, N. Christin, and L. F. Cranor, Carnegie Mellon University, CIC 2222E, 4720 Forbes Avenue, Pittsburgh, PA 15213; emails: {crichton, nicolasc, lorrie}@cmu.edu.



This work is licensed under a [Creative Commons Attribution-NoDerivs International 4.0 License](https://creativecommons.org/licenses/by-nd/4.0/).

© 2021 Copyright held by the owner/author(s).

1559-1131/2021/09-ART3 \$15.00

<https://doi.org/10.1145/3473343>

Jordan's experience is not a hypothetical scenario: this participant (whose real name is not Jordan) and their experience come directly from our research on the web browsing of home computer users. Unfortunately, many people who face similar circumstances are not so fortunate. Web browsing mishaps can result in the loss of personal or financial information, stolen credentials or accounts, and the installation of malware. In the end, it was not a technical problem that led to the installation of adware on Jordan's computer, it was the actions and decisions of the user. Researchers and industry professionals aiming to improve end user security must first understand the answer to a more general question: how do users browse the internet? The more we understand what user behavior is normal, the better we can understand how users are lured into dangerous websites and identify anomalous behavior. This insight can help inform a broad spectrum of internet-related research ranging from predicting user exposure to malicious websites [46] to improving how users search the web [8].

However, our current knowledge of user browsing behavior is limited. This information comes primarily from one of two sources. The first is from the large-scale collection of web requests made by users. Most often this data is captured by major technology companies and infrequently released except as aggregate measures. Several large datasets have been accumulated by the research community [1, 26, 29, 32]. However, these measurements rely on server-side data collection which have several known issues that limit their accuracy [29, 54]. The second source of information is a series of small-scale studies that collected client-side data directly from participants [6, 14, 36, 43, 48, 63]. Although these studies avoid the measurement issues of their counterparts, they sacrifice size for accuracy. Of these studies conducted over the past three decades, only one has recruited more than 25 participants.

Given this trade-off in the existing literature, we propose examining user browsing behavior using a dataset that harnesses the advantages of both types of studies. This data was provided by the **Security Behavior Observatory (SBO)** [17], a longitudinal study that collected *in situ* measurements of the personal computer use of 623 U.S.-based Windows users between December 2014 and July 2019. While mobile internet use continues to rise, browsing on personal home computers remains a vital component of a user's online experience and has become even more relevant with the increase in individuals working from home. Mobile users, particularly those on smartphones, have been shown to exhibit different browsing behaviors due to the markedly different user interface experience [50]. As such, our findings are limited to the specific context of U.S.-based Windows personal computer users.

In this study, we first orient the new results from the SBO in the context of previous work and identify five changes in browsing behavior over time. Most importantly, we observe a substantial increase in the use of multiple browser tabs while navigating the web. Based on this trend, we quantify the accuracy gained from including tab information in web browsing research, one of the known limitations of server-side measurements. We then present further evidence to support previous claims that the commonly used concept of an "average internet user" does not exist. Moving beyond what is done in previous work, we leverage the rich data provided by the SBO to contribute a detailed snapshot of how users browse the internet. In doing so, we identify a highly centralized pattern to user browsing. Last, we pinpoint the three primary pathways through which users leave their usually visited websites and navigate to less-traveled areas of the internet that they have never visited before.

2 RELATED WORK

Previous work studying web browsing behavior has primarily taken one of two approaches: large-scale measurements of web requests or smaller user-based studies. In contrast, our work employs a hybrid approach to capture the advantages of each type of study, providing more accurate measure-

ments on a larger scale. We leverage these improvements to answer several key research questions. First, how has user web navigation changed over time? Second, how do users spend time browsing the internet? Third, how do users reach low-visited websites on the periphery of the internet?

2.1 Large-Scale Browsing Studies

Most often, studies of user web browsing employ large-scale data collection to capture the web requests made by thousands of internet users. These studies frequently rely on server-side logging, collaboration with technology companies, or third-party data aggregators to compile such large datasets. Once collected, user behavior is most commonly modeled as a *clickstream*: a series of web requests over time. Clickstream analysis has been used to study a variety of topics related to web browsing including user security [4, 46], Sybil detection [56], user demographic prediction [21], and purchase behavior [28]. In addition, there have been a series of studies conducted using clickstreams in specific browsing contexts such as social media websites [2, 57, 58] and search engines [22, 37]. More closely related to our study, several works have leveraged clickstream models to examine user behavior more broadly. These studies assess the types of websites a user visits, identify patterns in user navigation, and measure the time users spend on web pages, also known as dwell time [1, 26, 29, 32].

However, clickstream studies, which typically employ server-side measurements, suffer from three primary limitations. First, Vassio et al. [54] estimated that as little as 2% of web requests captured in clickstream models are user initiated. The overwhelming majority of web requests are generated automatically to load dynamic content and buffer multimedia. With the pervasiveness of streaming video and dynamic web content in modern websites, the noise that automatic requests introduce into clickstream models makes it difficult to distinguish what is actual user behavior. Second, server-side measurements may miss when users return to a recently visited web page since these pages are often cached by the browser on the client side. Given that users frequently return to pages they have recently visited, clickstreams may miss a substantial amount of user navigation [29, 48]. Third, although clickstream models capture user actions that result in the generation of a web request, they do not reflect when users create, switch, or remove tabs within their browser. Since the introduction of browser tabs in the late 1990s, tab use has grown in popularity. By 2010, browsing sessions that used multiple tabs exceeded those that did not [23]. Although a clickstream model might show a user staying on a single web page until the next web request, the user may have actually been viewing several different web pages in other tabs during this period of time. As a result, clickstreams do not fully capture how users navigate the web, and subsequently their measurement of dwell time is inaccurate.

Of the previous studies mentioned, only Lehmann et al. [29] acknowledged this limitation and attempted to account for tab navigation by inferring tab changes from server-side logs. Although this represents an improvement over basic clickstream models, the proposed solution was unable to distinguish tab changes from back button navigations and could not tell if more than one tab change had occurred between requests.

2.2 User-Based Browsing Studies

In contrast, client-side studies of user behavior do not face the same limitations as server-side measurements. This is because these studies can log user activity directly rather than interpreting it from captured web requests. As a result, user-based studies have served as the foundation of our web browsing knowledge, providing insightful snapshots of user behavior between 1994 and 2011 [6, 14, 36, 43, 48, 63]. However, due to both the expense and technical challenges of scaling this type of study, there are relatively few and those that exist draw on a small number of participants.

To fill this gap, the SBO was developed to provide client-side data collection from a large panel of users [17]. The SBO initiative builds on the HomeNet project, an early work building panels of

internet users [25]. The data collected through the SBO project has been used to provide insight into a variety of user behaviors including the use of passwords and private browsing [18, 45]. Another contemporary project similar to the SBO has been used to study malware attacks [31]. However, this recent study has not examined user browsing behavior more broadly. For this study, we obtained a subset of the SBO data and analyzed it to better understand browsing behavior generally and provide a much-needed update to previous user-based browsing studies.

2.3 Changes in Web Navigation

Given that the most recent user-based browsing study was conducted nearly a decade ago, it is important to assess how user navigation has changed over time. Previous user-based studies of browsing behavior have focused on the frequency of various navigational actions, such as clicking a link or submitting a form, and high-level browsing metrics, like the number of web pages or the breadth of websites visited [6, 36, 43, 48, 63]. In particular, several of these studies' primary aim has been to measure the revisitation rate: the frequency that a user returns to web pages they have previously visited compared to those they have not [36, 43, 48, 63]. These studies, conducted between 1994 and 2011, provide a series of valuable snapshots of user behavior. As such, the first aim of this study is to compare our measurements from the SBO to that of previous work, identifying changes and trends over time.

One pattern that previous work has already identified is the increase in the use of multiple browser tabs to switch between web pages [14, 23, 29]. This behavior, known as tabbed or parallel browsing, has been studied in the context revisitation rates [63] and search behavior [22, 23]. However, the effect that tabbed browsing has on user browsing generally and the way in which it is measured is not well understood [29]. This gap in knowledge motivated our analysis of tabbed browsing behavior among SBO users and our quantification of the error inherent in web measurements that fail to account for the use of multiple browser tabs.

In addition, Obendorf et al. [43] present an argument against the concept of an average internet user. Although this argument is supported by some subsequent findings outside of the context of user browsing, it is in direct tension with the continued development of software and design of research studies that are predicated on the assumption that an average user exists [15]. Given this contradiction, and the changing nature of user browsing behavior, we provide an updated assessment of the average user as a concept and evaluate clustering browsing behavior into user archetypes.

2.4 How Users Browse the Web

Large-scale server-side browsing studies have examined user behavior across a variety of different contexts including the use of search engines [8, 22, 23] and social media sites [2, 32]. However, these studies do not cover how users browse the internet as a whole and the many different types of websites it contains. Although user-based studies paint a broader picture of web behavior, their focus on high-level metrics like revisitation rates do not provide enough actionable information for researchers and practitioners. This methodological gap warrants additional study of user browsing behavior using an exploratory lens. Leveraging the advantages of the SBO, we contribute a detailed picture of how users spend time on the internet and how they navigate between sites across the web.

2.5 The Periphery of the Internet

In particular, the time that users spend on low-visited websites is of great interest to researchers because it has been demonstrated that these sites have a higher likelihood of containing risky or malicious content [24, 33, 35, 46]. These sites, which we refer to as periphery websites, are defined

Table 1. Differences between SBO Participant Demographics and Those of the General Population of the United States

Demographic	SBO	US
Age 18–34	73.0%	23.3% [52]
Female	60.7%	50.8% [52]
Computer-Related Field	27.1%	2.3% [51]
Bachelor's Degree or Higher	58.9%	32.3% [53]

by having a low global traffic rank. A better understanding of how users end up on these types of sites can aid in the development of protections for users. Beyond security, this information has applications for work related to spam [5, 60] and misinformation [42]. In this article, we identify the main pathways that lead users away from their routine web browsing and bring them to low-trafficked periphery websites. These pathways represent areas of opportunity for the design of future interventions to aid and protect users.

3 THE SBO

The data for this research was collected through the SBO platform and was provided to our research team for the purposes of this study. The SBO was a longitudinal study of user web browsing and computer use that filled a gap in the literature between large-scale web measurement studies and small laboratory experiments. Between December 2014 and July 2019, the project recruited and collected data from 623 participants who were primarily based in the Pittsburgh, Pennsylvania, metropolitan area of the United States. On average, participants remained in the study for just under 2 years ($\mu = 1.76$, $\sigma = 1.05$). As a part of the SBO study, participants voluntarily consented to have their home Windows computers instrumented with a variety of sensors that collected data automatically, encrypted it locally, and sent it to a secure centralized repository [17]. These computers were used primarily at home, a couple of computers were shared among users in the same household, and several participants had multiple computers across the lifetime of the study. Despite being located at home, most computers were used for a combination of work and leisure. However, it is not certain the extent to which a user's complete computer use was captured, as participants may have maintained additional home computers or used computers outside of their residence. Sensors automatically collected data whenever the user's computer was powered on. As compensation, individuals received \$30 for enrollment and \$10 per month for the duration of their participation. Individuals were free to discontinue their participation in the study at any point. The protocol for this project received approval from the Institutional Review Board at Carnegie Mellon University.

As shown in Table 1, the SBO sample is generally younger, with the majority of participants falling between the ages of 18 and 34 years (73%), and contains slightly more women (61%) than the general population. In addition the sample overrepresents participants who have either education or professional experience in a computer-related field (27%) and those who have received a bachelor's degree or higher (59%).

The data specifically used in the course of this research was collected through the SBO between March 3, 2016, and July 28, 2019. This subset includes information from 257 participants across 608 different computers.¹ Although the SBO has collected a larger pool of data, our analysis was limited to the subset for which there was complete information from the SBO tabbed browsing sensor and the SBO system-level sensors. These constraints restricted the dataset to 53M raw browsing records

¹There are a greater number of computers than participants because some users registered multiple computers or replaced their computer partway through the study.

Table 2. Summary of the Raw Data Provided by the SBO and How It Was Processed and Used as a Part of This Study

Sensor	Records	Description
Navigation	45.1M	Automatic and user initiated web navigation
Tab	8.1M	Tab creation, activation, and removal events
System	8.0M	Indicator that sensors were active, sent at regular intervals
Operating System	2.4M	Computer startup events among others
Power	2.3M	Computer suspend and resume events
Session	0.6M	Login, logoff, lock, and unlock events
Foreground Window	40.6M	Application that was active in the foreground of the computer
Mouse	35.6M	Indicators of mouse movement collected at regular intervals

and 89M system records.² Due to the sensitive nature of the data collected, the risk of leaking personally identifiable information through data contained in URLs, and the possibility that the highly detailed information collected in this study could be used for de-anonymization [40]; this dataset will not be released publicly.

4 METHODS

To conduct our analysis, we had to integrate and transform the raw data we obtained from the SBO. SBO information was provided in eight data extracts, each corresponding to one of the sensors instrumented on the clients' machine. As we summarize in Table 2, we can divide the eight SBO extracts into two primary categories: browsing and system. The browsing sensors (top two rows in the table) collected navigation and tab data using an extension installed in the participant's Google Chrome browser. The navigation sensor captured user-initiated web requests (13.2%) in addition to those automatically generated by web pages (86.8%). The remaining six system sensors, which collected data through software installed directly on the client machine, captured a variety of information on the computer's state, the application currently being viewed, and the user's mouse inputs.

Unlike in server-side measurements, our sensors collect web request data in which each request is labeled as initiated by the user or generated automatically. In this analysis, we excluded automatically generated web requests. Although automatically loaded content does provide additional information about the websites that load them and can represent an important attack vector if that content originates from a malicious third party [38, 41], this study focuses on how users navigate the web rather than the consequence of their navigation. Following similar logic, we specifically labeled adware sites in our data but did not include other forms of malicious content. When a user attempts to use an adware-afflicted browser, their intended navigation is usually redirected first to the adware site and then to an unintended destination. Unlike phishing or malware sites, which can be harmful to a user once they land there, adware sites are part of the navigation itself and therefore important in explaining how users end up traversing across the web. Thus, we set aside the assessment of the automatically generated content and malicious websites for future research. As a result, we use the term *web request* to refer to user-initiated web requests only, and *web navigation* refers to the actions taken by the user in the course of browsing the web whether or not they generated a web request. A visit to a *website* indicates a web request was issued for a

²Incomplete SBO data is primarily a result of some participants receiving a lightweight version of the SBO software, which limited data collection to a subset of sensors to avoid affecting the participant's daily use. In a few cases, sensors broke down after the user installed other software, applied updates, or changed their computer settings during the course of the study.

page with a different domain name than the page the user is currently on. A visit to a *web page* corresponds to a web request for a different URL, but not necessarily a different domain name.

For each participant, we combined data provided by each of the sensors to reconstruct a single timeline of events. During the process, we took several steps to cleanse and prepare the data. First, we compared the data collected from the browser extension and the system sensor to identify gaps where there was missing information. As a result, we excluded data from our study in cases where there was either browsing data but no system data or there was system data indicating the participant was actively using Google Chrome and no browsing data. Second, we accounted for occasional lag time found in the tab sensor by overlaying the data from the tab sensor with that obtained from other sensors and re-aligning the tab events. Third, we added events to the reconstructed activity trace in cases where there was an obvious and clearly identifiable gap in the chain of events. In these cases, we added the missing event to the timeline. For example, when we observed an application being used before a user had actually logged in to their computer, we added a login event preceding it, or when a user clicked a link in a tab that had not yet been created, we added a tab creation event immediately before it. When missing events were detected but the obvious course of events was not clear, the data surrounding that gap was discarded from the analysis. In total, missing events accounted for 1% of the entire dataset. Finally, we overlaid the mouse movement information to determine when a user was actively interacting with the application or web page they were viewing. We report this *active use* period for all measurements of time spent.

We calculated the *revisitation rate*, a metric used to capture how frequently users return to previously visited web pages, in two ways. We first followed the conventional definition introduced by Catledge and Pitkow [6] in 1995. They define the revisitation rate as the ratio of the number of web requests to previously visited pages over the total number of requests. The second method, proposed by Zhang et al. [63] in 2011, accounts for tabbed browsing by including tab changes into the definition of a web request. We calculated this value, which we refer to as the *effective revisitation rate*, as the number of web requests and tab changes to previously visited pages divided by the total number of requests and changes. Neither definition considered an expiration time for revisitation. Hence, once a page was visited, subsequent navigations back to that page were considered revisits indefinitely.

We also supplemented the information collected by the SBO with data from external sources and manual coding by our research team. We collected the popularity rank of the top 10M websites from Open PageRank and matched the list against domains in the SBO dataset [44]. Other sources of website rank, including Amazon Alexa, were evaluated, but they only provided the top 1M websites. To capture websites that were further in the tail of the distribution, we opted to use Open PageRank. In addition, we applied hierarchical category labels to each website that users visited and each application they used on their computer. Since no corpus existed for categorizing software applications, the research team manually coded this data. The coding structure with examples can be found in Appendix A.1.

The categories of websites we used were based on a consolidated DMOZ coding structure using 65 codes across 15 high-level categories [13]. The 15 website categories were used to define the different types of *web activity* that a user engaged in. Time spent actively visiting a website of a certain category was considered equivalent to engaging in that web activity. Documentation of the coding hierarchy can be found in Appendix A.2. Our research team manually coded the top 2,000 websites to ensure accuracy across 84% of browsing records. To categorize the remaining 121,000 websites that fell in the tail of the browsing distribution, we evaluated services like Amazon Alexa but found website coverage to be less than 30%. Instead, we employed a hybrid technique. If the domain existed in the DMOZ dataset, we used that category label directly. If it did not appear

in DMOZ, we used a convolutional neural network that had been trained to predict the website category based on the domain name. We trained the algorithm using the labeled DMOZ data and configured it with parameters similar to those employed in previous work [46, 64]. Although the cross-validated accuracy of the model, approximately 40%, appeared at first to be insufficient, we found that the majority of the mislabeled sites were between highly similar categories. For example, industrial websites were often labeled as business, and streaming websites were often labeled as movies and television. By conservatively grouping these similar categories, the accuracy of the model improved to 72%. This led to an overall accuracy of 76% across unique domains and 94% across the entire dataset. As such, we concluded that the classifier sufficed for the purposes of our study.

5 FINDINGS

We contribute six main findings detailed in the sections that follow. First, in comparison to previous studies, user browsing and navigational habits have changed over time. Second, the failure to account for the use of browser tabs in web research leads to inaccurate measurements. Third, users exhibit a wide range of browsing habits and do not easily fall into categorical types. Fourth, web browsing consumes the majority of users' time spent on their home computer, eclipsing the use of all other applications among 75% of our participants. Fifth, users spend the majority of their browsing time on a few popular websites but also spend a disproportionate amount of time on low-visited websites on the edges of the internet. Sixth, three primary gateways are used to navigate to low-visited sites: search engines, other low-visited sites, and intermediary mid-tier websites.

5.1 Changes in Browsing over Time

This study represents a substantial contribution in a series of web measurement studies conducted over the past three decades. With the most recent study having been run in 2011, our results help illustrate the evolution of web browsing over the subsequent 8 years. As summarized in Table 3, this study was conducted on a substantially larger scale, collecting more than 5.2M web requests across 257 participants over a period of 352 days on average. Of the previous studies, only one had more than 25 participants. In addition, the browsing data for this study was collected from Google Chrome users. In 2019, Google captured 61.8% of the browser market share as opposed to XMosaic (obsolete), Netscape (obsolete), and Firefox which maintained 34.1% of the market share at its peak in 2010 and only 4.8% in 2019 [55].

Our results indicate five primary changes in user browsing over the past three decades. First, the use of multiple tabs to switch between web pages has almost doubled as a proportion of all web navigation since 2011, rising from 28.6% to 54.4%. Since being introduced in the 1990s, tabbed browsing has grown in popularity among computer users by providing the ability to switch rapidly between tasks, open multiple links in the background, maintain frequently used pages, and provide short-term bookmarks [14, 23]. However, our findings represent a leap in adoption with tab navigation exceeding the number of web requests made by users for the first time.

Second, the use of the back button to open previously visited web pages has fallen from over 30% of web requests in the 1990s to 10.5% in the 2000s to 2.1% in our study. Previous work has suggested that back navigation has been replaced by the use of multiple tabs [14, 43]. Although this would fit the trends we observe across studies, ours included, we did not find evidence of an inverse relationship when examining the use of tabs and back navigation between users in our study.

Third, search engine functionality directly embedded in the browser has become a staple of user web browsing comprising 11.5% of all navigation. We find that the use of search terms in the address bar has eclipsed that of supplying URLs directly. The ability to provide search terms to the

Table 3. Comparison of User Browsing Behavior in the SBO to Major Measurement Studies

	Catledge & Pitkow [6]	Tauscher & Greenberg [48]	McKenzie & Cockburn [36]	Obendorf et al. [43]	Zhang & Zhao [63]	SBO
Overview						
Study period	1994	1995–96	1999–2000	2004–05	2011	2016–19
Browser used	XMosaic	XMosaic	Netscape	Firefox*	Firefox	Chrome
Tabs supported	No	No	No	Yes	Yes	Yes
Study-Level Metrics						
Participants	107	23	17	25	20	257
Mean length (days)	21	42	119	105	31	352
Web requests	31,134	84,841	83,411	137,272	80,811	5,273,926
Tab changes	N/A	N/A	N/A**	–	32,432	5,704,476
Distinct URLs	–	–	17,242	65,643	49,450	2,097,782
Pages visited per day	14	21	41	90	130	163
Traffic to Google	N/A	N/A	–***	16.6%	–	16.3%
Revisitation rate	61%	58%	81%	45.6%	39.3%	59.8%
Eff. revisitation rate	–	–	–	–	59.6%	78.1%
Web Requests by Type						
Link	45.7%	43.4%	–	43.5%	–	47.2%
Direct access****	12.6%	13.2%	–	9.4%	–	9.2%
Search	N/A	N/A	N/A	N/A	N/A	11.5%
New tab/window	0.2%	0.8%	–	10.5%	–	7.6%
Form submit	–	4.4%	–	15.3%	–	12.7%
Back	35.7%	31.7%	–	14.3%	–	2.1%
Reload	4.3%	3.3%	–	1.7%	–	3.7%
Forward	4.3%	3.3%	–	0.8%	–	0.5%
Other	–	2.3%	–	4.8%	–	5.5%

“N/A” represents data that was not available at the time of collection (e.g., because the feature did not exist), and “–” represents data that was available but not collected.

*60% used an instrumented version of Firefox that logged detailed user behavior, 40% used a different browser where only web requests were captured.

** Although tabs were invented and introduced in some other browsers in the mid-1990s, Netscape/Mozilla/Firefox did not support tabs until around 2002.

*** Although Google was still a small search engine at the time, it was steadily gaining market share.

**** Includes URLs typed directly to the address bar (93.9%) and bookmarks (6.1%).

address bar was first introduced when Google Chrome was released in 2008 and was not available in Firefox 3.0, the browser used in the most recent 2011 browsing study. Within this category of navigation, we observe two distinct types of search behavior. In 99.9% of cases where search terms were provided to the address bar, the user submitted the query directly and was redirected to a search engine web page displaying the query results. In the remaining cases, the user selected one of the browser’s cached suggestions that appeared as a set of drop-down options as they typed and was navigated directly to the destination website.

Fourth, the average revisitation rate, both conventional ($\mu = 59.7\%$, $\sigma = 8.3\%$) and effective ($\mu = 78.1\%$, $\sigma = 5.8\%$), is high. To interpret this rate, we make direct comparisons to the findings of Zhang and Zhao [63], who report both the conventional and effective rates. In both cases, we find a much higher rate of revisitation among users in our study. To compare to studies prior to 2011, we use the effective revisitation rate which accounts for the increased use of tabbed browsing observed in our study. By comparison, our revisitation rate is on the higher end of the spectrum, nearing that found by McKenzie and Cockburn [36]. Revisitation rates lack a mechanism to account for the length of time that data is collected. A revisit to a web page after 2 minutes is treated no differently than a revisit to that same web page after 2 years. Given the relatively short duration of previous studies, we speculated that the longer time period covered by the SBO might account for the increased revisitation. To test this hypothesis, we recalculated our effective revisitation rate limiting the data of all participants to the first 20 days of their participation, allowing us to roughly match the quantity of data collected by Zhang and Zhao [63]. However, our result using this method ($\mu = 76.6\%$, $\sigma = 6.7\%$) did not shift substantially. As a result, we conclude that users in our sample traveled back to sites they have previously visited at a higher rate than in most other studies.

Fifth, the number of pages visited per day has steadily increased over time. This is driven in part by changes user browsing patterns. In general, users are spending more time online as work,

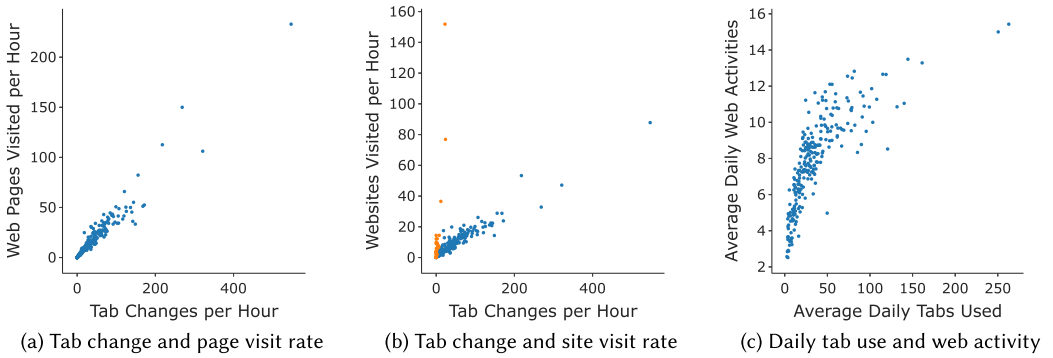


Fig. 1. Illustration of the positive correlations between tabbed browsing and web page visit rate (a), website visit rate (b), and daily web activities (c). All rates are reported per hour of active browsing time. In (b), two kinds of browsing emerge: linear browsers who visit websites sequentially (orange) and multitaskers who jump between websites using multiple tabs (blue).

shopping, and entertainment become increasingly online activities. In addition, we find that users who switch between tabs more frequently visit more web pages. As we have already observed, switching between tabs as a form of web navigation has grown by 79% since 2011. Therefore, it follows that tabbed browsing behavior accounts for some of this growth. However, although behavioral factors do contribute to this trend, we hypothesize that technological changes are as much, or likely more, of a factor driving the increase in average web page visits. This includes rising broadband internet speed and the expansion of content delivery networks that deliver web pages to the browser faster.

5.2 Patterns in Tabbed Browsing Behavior

Investigating the increase in the use of multiple browser tabs further, we find several behavioral trends associated with users switching between tabs. Figure 1 displays three graphs that illustrate how the use of tabs is associated with, from left to right, the rate of visiting web pages, the rate of visiting websites, and the average daily web activities. As mentioned previously, users who switch between tabs more frequently also view more web pages. As shown in Figure 1(a), a linear relationship ($r = 0.97$) exists between the rate that users switch tabs and the rate that they visit web pages. On average, users visit an additional web page for every 2.2 tab changes they make. This contradicts previous findings that showed greater tab switching reflects higher levels of multitasking but results in the same number of page views [23].

If we examine the number of websites a user visits, rather than pages, we find that two types of behavior emerge. As shown in Figure 1(b), we predominately find a linear trend that follows the leftmost figure: users who switch between tabs more frequently tend to visit more sites. Shown in blue, these multitasking users visit an additional website for every five to six tab changes made. In contrast, we find a smaller subset of users who exhibit linear browsing preferences. These users, highlighted in orange, visit websites sequentially while infrequently switching between tabs.

Finally, as illustrated in Figure 1(c), we observe a non-linear relationship between daily tab use and the average number of web activities that a user engages in. Here a distinct web activity corresponds to the user visiting a website of a specific category such as entertainment, social media, or communication. As such, users who visit websites across a broader set of categories per day are also said to engage in a greater number of web activities. Although the rate at which users switch tabs is weakly correlated with the number of activities ($r = 0.35$), there is a much

Table 4. Inclusion of Tabbed Browsing in Browsing Studies Is Important to Accurately Capture User Behavior and Measure How Long Users Spend on a Given Web Page

	User Actions		Dwell Time	
	Frequency	Percentage	Coverage	Accuracy
Web requests only	5,273,926	48.0%	80.7%	63.0%
With tabbed browsing	10,978,402	100.0%	88.9%	100.0%

stronger correlation between the number of tabs used and web activities per day ($r = 0.67$). This relationship exhibits a logarithmic form with large increases in web activity at low levels of tab use and smaller increases at high levels of tab use. This indicates that multitasking behavior does not necessarily mean that users engage in a greater number of web activities. However, the use of multiple tabs appears to facilitate users engaging in a greater number of activities throughout the day.

5.3 Implications of Tabbed Browsing for Web Research

In addition to observing several behavioral trends related to tabbed browsing, we also find that the increased use of tabbed browsing behavior among users poses significant challenges for web and browsing research. As discussed previously, we observed that the use of multiple browser tabs to switch between pages has exceeded the number of traditional web requests for the first time. Since switching between tabs does not generate a web request on its own, less than half of a user's interactions with their browser are captured in server-side web logging or clickstream studies that focus exclusively on web requests. Although the sequence of links that bring a user from one web page to another remains unaffected, attempts to gain a deeper understanding of user behavior (e.g., analysis of how users gather information online) based on these methods are likely to be biased.

Furthermore, we find that the measurement of dwell time, an important metric in industry and research, is quite inaccurate if tabbed browsing is not accounted for. Dwell time refers to the period of time that a user spends on a given web page and is used in search engine optimization, advertising, and a variety of other applications. This is usually measured by computing the time between user web requests during the course of a browsing *session*. Typically, 10 to 30 minutes of inactivity is used to denote session completion [2, 8, 9, 19, 20, 26, 29, 32, 34, 46, 49, 56, 62]. We evaluated this method by applying it to a clickstream version of our own data and then contrasting the results, which are summarized in Table 4. Using a 20-minute threshold, we find that traditional clickstream models ignore nearly 20% of user dwell time. Furthermore, of the 80% that is captured, 37% of the time the user is on a different page than the one indicated by the clickstream model.

Accounting for tabbed browsing can remedy the dwell time accuracy issue and improve data coverage to 88.9%. The latter is done by extending the user browsing session beyond the last web request to the last tab navigation before 20 minutes of inactivity. The remaining 11.1% of coverage was obtained in our models by capturing mouse movements and the use of non-browser applications, eliminating the need to define a browsing session in the first place.

5.4 There Is Indeed No Average User

In line with the findings of Obendorf et al. [43], we do not find evidence that would support the concept of an average internet user. Within our sample of participants, we do not observe a normal distribution of user behavior or a single cluster of shared habits. Instead, we notice a broad spectrum of user behavior. For example, in Table 5, we find that users on average spend 2.6 hours per day on their computer. However, across users, we find a wide range of activity with one

Table 5. Differences in Average Daily Browsing Behavior across Users When Browsing at Least One Website in a Given Day

Activity per Day	Mean	Std. Dev.	Min	Max
Hours spent using computer	2.6	1.9	0.3	11.1
Applications used	7.3	2.9	1.7	17.0
Hours spent browsing	1.7	1.35	0.1	10.2
Percentage of time browsing	67.5%	14.7%	16.0%	96.2%
Websites visited	20.1	13.7	2.9	142.0
	(20.5)	(15.7)	(2.9)	(143.0)
Pages visited	154.4	122.7	6.1	702.0
	(162.7)	(164.1)	(7.8)	(1,885.1)
Tabs used	35.9	31.7	2.5	250.5
	(36.8)	(34.8)	(2.5)	(263.0)
Tab changes	129.2	119.3	2.4	673.0
	(133.0)	(133.3)	(2.4)	(1,072.7)
Web activities	7.6	2.3	2.5	14.5
	(7.6)	(2.4)	(2.5)	(14.5)
Web activity changes	38.2	30.5	2.0	225.0
	(41.8)	(56.0)	(2.0)	(780)

Values are reported both excluding and including, in parentheses, an outlier who had a browser hijacker installed on their machine that multiplied the actual user activity captured by the sensor.

participant having spent less than 20 minutes per day on their computer and another who spent over 11 hours on average. A similarly broad spectrum of habits is observed across application use, browsing time, websites visited, page views, and web activities. We did observe one outlier whose high browsing use was a result of being in the study for a short duration (7 days) and having a browser hijacker installed on their machine. This adware multiplied the user's actual activity by redirecting web traffic and opening unwanted advertisements. We report average behavior across users both excluding and including, in parentheses, this user.

Although we did not expect to find a single average user, we had hypothesized that we would find several different web browsing behavior profiles based on the duration of user browsing, the frequency of multitasking, and the types of websites typically visited. However, the spectrum of browsing behaviors we observe does not lend itself to clustering users into discrete groups. The research team applied unsupervised clustering (DBSCAN) using several of the key metrics but did not find any distinct clusters. Furthermore, we did not find any strong relationship between these behaviors and different classifications of users based on their primary web activities. For example, user behavior was independent of whether the user primarily spent their time on social media websites as compared to those who mostly used their browser for work. As such, we echo the findings of previous work in concluding that there is no such thing as an “average user” [15, 43].

5.5 How Users Browse the Web

To provide initial context to user browsing behavior, we examine the role of the web browser in relation to other uses of the computer. Based on the collection of application data, a unique feature of the SBO, we observe that participants primarily use their computers to browse the internet. Figure 2 shows how users spend their time on the computer, grouped into categories. Dots represent individual measurements; the adjacent whisker plots provide information about the distribution of these measurements. As the figure shows, browsing serves as the primary function of

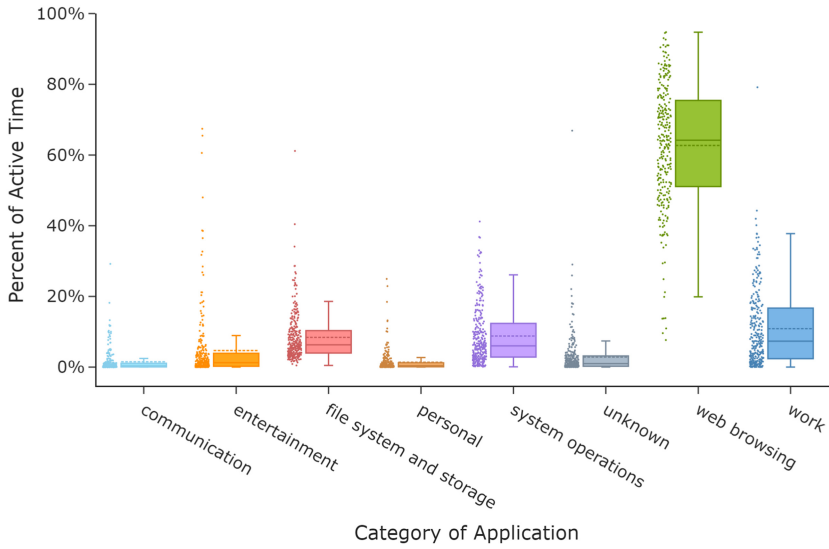


Fig. 2. How users spend their time on their computer. Each dot corresponds to a specific user. Whisker plots aggregate this data over all participants to provide the lower and upper fences (first quartile – 1.5 the inter-quartile range, and third quartile + 1.5 the inter-quartile range). Within the boxes, the dashed line represents the mean and the solid line the median. Users mostly spend their computer time browsing the internet.

nearly all of the participants’ computers. On average, browsing consumes 63% of the time users spend on their machine. However, we observe a wide range of use spanning from 8% to 95% of users’ time. Outside of browsing, users spend most of their time split between working using professional or educational software, configuring aspects of the computer or operating system, and directly interacting with the file system through actions like copying or moving files manually.

For most users, using entertainment applications outside the web browser does not consume a large portion of a participant’s computer time. This is likely due to the traditional forms of entertainment, like movies and television, becoming increasingly available for streaming directly in the browser rather than requiring a separate media player. In general, this follows a growing trend in which the browser is becoming the centerpiece of the personal computer experience. Video games make up the majority of non-browser entertainment consuming 56.7% of users’ time in that category. Although some games have become browser based, many still require their own software applications to play.

Within the web browser, we find that users spread their browsing time across many different types of websites. Figure 3 illustrates how users spend their time on the web, grouped by website categories or “web activities.” On a daily basis, users engage in seven to eight web activities on average. Most commonly, users spend their time on work (16%), social media (15%), and entertainment websites (14%), with some users dedicating well over 50% of their browsing time to these activities. These popular activities align closely with the findings of An et al. [1], a study of the internet use of households in Belgium that was conducted at the network level.

However, web browsing is more than just how users spend their time on the internet. The way in which users navigate between different types of websites is equally important. Figure 4 combines these two dimensions to illustrate how users browse the internet. In this graph, categories of websites are represented by circular nodes and are color coded by type in the same manner as Figure 3. The 32 most visited websites, which account for 50% of user dwell time, are separated

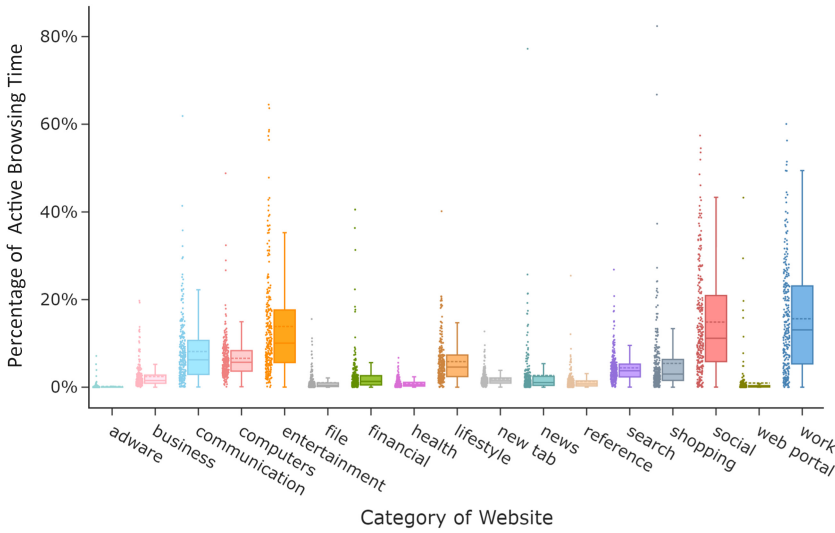


Fig. 3. How users spend their browsing time. As before, each dot corresponds to a specific user. Whisker plots aggregate this data over all users to provide the lower and upper fences (first quartile – 1.5 the inter-quartile range, and third quartile + 1.5 the inter-quartile range). Within the boxes, the dashed line represents the mean and the solid line the median. Users spread their time across many different activities when browsing the web. The most popular activities involve social media, work, and entertainment.

out and clustered together with the other websites of the same category. The size of each node represents the cumulative time users spent browsing those types of websites or website. The lines connecting various nodes indicate users navigating between different websites. The thicker the lines, the more users are navigating between the two types of sites. As such, this represents how frequently users are switching between different web activities. The color of the line indicates that users are primarily navigating away from websites of the matching color and landing on websites of another category. For example, the thick gray line connecting new tabs and Google conveys that more users open a new browser tab and then navigate to Google rather than vice versa.

Using this graph, we observe six common patterns in user browsing behavior. First, search engines (purple nodes) serve as the backbone of user navigation, allowing users to traverse to other areas of the internet. This is exhibited by the thick purple lines, indicating outbound navigation, extending from Google to all other nodes in the graph. This indicates that search engines are heavily used to navigate to across the internet and switch between different web activities. In total, participants used search engines to navigate to more than 62,000 different websites, approximately half of all websites visited.

Second, a common navigational pattern among users is opening a new tab, switching to their search engine, and then navigating outward to other types of web activities. This is illustrated by the thick gray line connecting the new tab node and Google. We hypothesize that this represents a jumping-off point for user browsing and indicates the start of a new browsing task or sub-task.

Third, we observe several categories of websites that have primarily outbound relationships with other areas of the graph. This includes categories like entertainment, shopping, lifestyle, computers, news, reference, and finance. In these cases, users are primarily navigating to these sites from a search engine, spending time on that site, and then leaving for other areas of the internet. We believe that this represents when users are engaging in a specific one-off task like reading an article, watching a video, retrieving a certain piece of information, or handling a transaction.

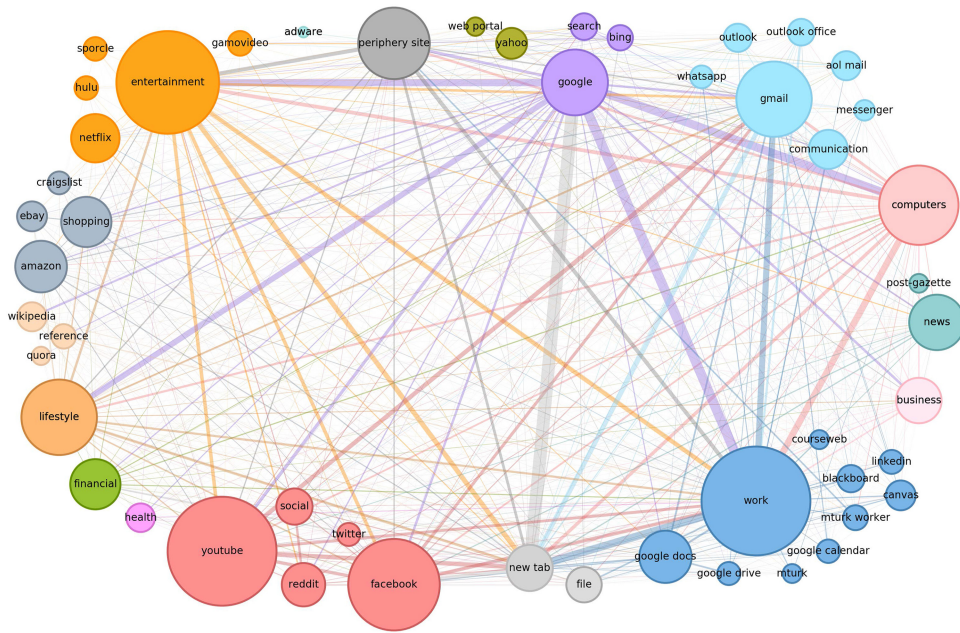


Fig. 4. Graph depicting the time user spend on websites (node size) and how users switch between different websites, or types of website (color coded by web category). The greater the weight of the line between nodes, the more users navigated between the two websites. The color of these lines corresponds to the node that users switch from more frequently than they switch to. One exception is the relationship between reference and work sites, which shared an equal proportion of inbound and outbound traffic.

Fourth, in contrast to the previous point, there are several categories that have primarily inbound relationships from other areas of the graph. Users are navigating to these sites from all different areas of the internet, not just from search engines. This includes social media, communication, and work-related websites. We hypothesize that these websites represent the more routine aspects of a user's web browsing. These sites are the ones most commonly used, those that are kept open in another tab that the user routinely switches back to, or ones that are bookmarked for regular access.

Fifth, we observe a core set of work-related activities and set of leisure activities along with a strong indicator that users multitask between the two frequently. On the right side of the graph, there are strong associations between work, business, email, and computer-related sites. Many websites in the computers category relate to programming and cloud computing. As such, we believe that these sites capture user productivity. On the left side of the graph, there are also strong associations between social media and entertainment activities. However, there are equally strong connections between the set of work and leisure activities as there are within these sets. This indicates that many users multitask, switching between work and leisure activities throughout the day rather than distinct blocks of time dedicated work and leisure individually.

Sixth, we find that large technology companies capture a substantial proportion of our users' attention within specific categories. YouTube and Facebook share the majority of users' time spent on social media, garnering 49% and 35% of total time, respectively. Google maintains a substantial proportion of user attention with Google Search accounting for 76% of search engine use and Gmail handling 54% of browsing time dedicated to communication (light blue nodes). Last, Amazon

Table 6. Percentage of User Dwell Time in the SBO Compared to Industry Measurements of Market Share

Website	SBO Dwell Time	Market Share
YouTube	49%	2% [47]
Facebook	35%	63% [47]
Google	76%	62% [55]
Gmail	54%	27% [27]
Amazon	40%	47% [10]

maintains a 40% share of users' online shopping (dark gray nodes). We compare our findings, based on user dwell time, to publicly available estimates of market share, likely based on website visits, in Table 6.

Although we do not observe a large difference between metrics for Amazon, we do find a substantial difference for YouTube. This is a result of comparing metrics based on dwell time to that using visits. Users who go to YouTube, which is dedicated to video content, likely spend a much longer period of time on each page watching videos as compared to other social media sites that contain a greater amount of static content. Following this logic, we see skew in the opposite direction for Facebook. The overrepresentation observed for Google Search and Gmail are likely a result of collecting our data through the Chrome browser, another Google product. Given the interoperability of Google's tools, it is not surprising to find a bias toward Google-owned websites in our sample.

The dominance of these large technology companies is indicative of a larger trend of centralization to web browsing. Overall, most of our user's browsing time is dedicated to a very small number of popular sites. Of the websites visited by our participants, we observe that users collectively spend 30% of their browsing time on the top 4 websites, 50% of their time on the top 32 websites, and 80% of their time on the top 1,000 websites. Of the 123,646 different websites visited, this concentration of activity covers less than 1% of the total sites. This indicates that users habitually return to the same small set of websites over and over again where they spend the majority of their time. This finding aligns with the high revisitation rates we observe in our sample, further reinforcing that web browsing is both highly centralized and becoming more centralized over time.

The allocation of browsing time within our sample also closely aligns with global web traffic ranks. Figure 5 provides the cumulative distribution of browsing time spent across websites by global popularity, measured by the OpenPageRank initiative's traffic ranking. Overall, we observe a power law distribution in which user browsing time follows an 80/15/5 split. Users spend 80% of their time on *core websites*, the most popular sites that have a traffic rank between 1 and 1M. An additional 15% is spent on *mid-tier websites*, those with a traffic rank between 1M and 10M. Finally, *periphery websites*, which fall outside of the top 10M, capture the remaining 5% of user attention. This distribution of user activity across websites resembles a core-periphery structure that is often used in network theory. In this type of model, there exists a dense, interconnected core and a sparse, loosely connected periphery [3].

Within the top 10M ranked sites, there were several anomalous websites that garnered a disproportionate amount of browsing relative to their rank. Upon closer inspection, these sites were specific to the user's occupation, the university they attended, or their regional bank. As such, these spikes in the curve are not unexpected. Of particular interest is the accumulation of 5% of user browsing time on websites outside of the top 10M. The time users spend on these periphery sites cannot be explained in the same manner.



Fig. 5. The cumulative browsing time spent across websites by their traffic rank. Users spend 80% of their time on core websites (rank 1–1M), 15% of their time on mid-tier websites (rank 1M–10M), and spend 5% of their time on periphery websites (rank 10M+).

5.6 The Periphery of the Internet

Previous work has demonstrated that websites with a lower traffic or page rank have a higher likelihood of containing risky content [24, 33, 35, 46]. Websites that experience high volumes of traffic are able to monetize through advertising, sales, or collecting data and have the resources to maintain the security of their site. Although most websites with low user traffic are completely legitimate, these sites have greater incentives to monetize through less reputable ad networks, contain unwanted trackers, or even contain malicious content. In addition, low-user traffic websites sometimes lack the resources to maintain proper security updates and can be hijacked by malicious actors.

In our data, we find that 89% of websites that were flagged by Google Safe Browsing as containing malware or phishing content were periphery websites that ranked outside of the top 10M sites. In one example, which is eerily reminiscent of earlier security work on web compromises [30], we found a periphery website run by a university research group that had been co-opted by external actors and transformed into a storefront for an illegal online pharmacy.

The left section of Table 7 shows the top 10 categories of websites visited on the periphery of the internet across all of our users. Over half of the visits to unranked periphery websites are related to video streaming, computers, and crowdworking. Many of the websites related to computers appear to be legitimate technical guides or small businesses. However, upon manually examining the websites in the streaming and crowdworking categories, we find more questionable content. Many of the streaming sites on the periphery include potentially pirated video sources. Furthermore, the majority of the crowdworking websites offer rewards and freebies in exchange for opinion surveys or personal information.

Also displayed in Table 7 are the most overrepresented website categories that users visit on the periphery of the internet. This is reporting in terms of relative risk, otherwise referred to as a risk ratio. The risk ratio is calculated by taking the probability that a user visits a website of a given category when browsing periphery sites divided by the probability they land on a website of the same category when visiting core or mid-tier websites. This does not mean that more of these types of websites exist on the periphery of the internet. Rather, users are much more likely to visit these categories of websites when on the periphery. For example, users are 28 times

Table 7. Top 10 Website Categories That Users Visit Most Frequently on the Periphery of the Internet and the Top 10 Overrepresented Categories in the Periphery as Compared to the Distribution of Visits to Core Websites

Top 10 Periphery Website Categories		Top 10 Overrepresented Periphery Categories		
Website Category	Prevalence	Website Category	Prevalence	Risk Ratio
Streaming	22.8%	Adware	2.0%	28.2
Computers	16.0%	Senior Health	0.3%	26.8
Crowdworking	12.4%	Alternative Science	0.2%	18.9
Music and Radio	5.3%	Alternative Health	0.5%	16.3
Education	5.2%	Streaming	22.8%	10.2
Sports	4.0%	Gambling	0.7%	7.4
Business	2.5%	Online Entertainment	1.2%	5.0
Finance	2.2%	Organizations	1.6%	4.3
Industrial	2.1%	Personal Pages	0.1%	3.4
Adware	2.0%	Adult	1.7%	2.8

more likely to land on an adware site when browsing the periphery of the internet than they are when visiting core or mid-tier sites. This relationship is descriptive rather than causal, and is indicative of the type of content that users are more likely to engage with when browsing periphery websites.

Of these categories, adware and streaming sites are particularly interesting as they both make up a substantial proportion of visits to periphery sites and are highly overrepresented. In addition, we find that users are much more likely to visit alternative health and science websites on the periphery. These types of websites, which often contain misinformation, have been shown to distort valid scientific information, perpetuate conspiracy theories, and negatively impact individual behavior and decision making [12, 16, 59]. This has important implications for researchers studying the dissemination of science-based information. Although much of the attention in this space has been directed toward social media [7, 39, 61], our results indicate that the scope of this problem extends well beyond these sites. In only 7.4% of these cases did the user navigate directly to an alternative health or science website from a social media site. This rises to 24% when accounting for indirect navigation where users visited a social media website before, but not immediately preceding, the site containing alternative information. Search (34.3% direct, 53.1% indirect) and crowdworking (15.5% direct, 31.0% indirect) websites were the other primary sources of visits to these sites along with a multitude of smaller contributing categories. We also find that users are more likely to engage with pornographic, gambling, and free streaming websites on the periphery of the internet. Last, in a manual inspection by the research team of websites pertaining to senior health organizations and personal pages, we find no unusual activity, just sites representing individuals, small groups, or businesses.

The higher relative risk of users interacting with adware, alternative information, and potentially illegal entertainment websites on the periphery of the internet makes this subset of the web of particular interest to researchers and practitioners. The sources of traffic that transport users to the periphery of the internet, which will be referred to as gateways, represent an important decision point in users' browsing. At these points, users deviate from their browsing routines and traverse outward to a website that is not often visited. To identify these gateways, we assessed the series of websites leading up to the user landing on a low-visited periphery website. Specifically, these events were limited to a user's first visit to a periphery site and exclude any subsequent visits to the same website during the duration of the user's participation within the study. In total, we observed 61,135 instances in the data that fit this criteria. As illustrated in Figure 6, we find

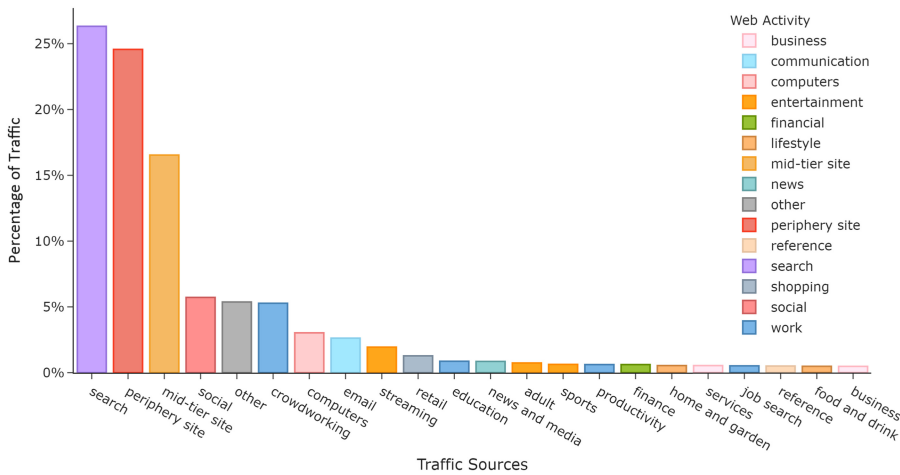


Fig. 6. The types of websites that users most frequently navigate from to reach websites on the periphery of the internet. Periphery (red) and mid-tier (orange) websites are grouped into respective categories to illustrate their use as major gateway to the periphery. All other categories represent traffic from different categories of core websites.

that users navigate through three primary gateways to reach the periphery: search engines, other periphery websites, and mid-tier intermediary websites.

The largest contributors of traffic to periphery websites are search engines, the traffic source for 26% of all new visits. As we have demonstrated previously, search engines serve as a springboard for users to navigate across the internet in general. As such, it is unsurprising that they are also used to land on new periphery sites. The second largest gateway to the periphery, contributing 25% of user traffic, is other low-visited periphery sites. This indicates that users tend to use search engines to jump from core to periphery sites, but once there they tend to traverse to other sites on the periphery. The third gateway to the periphery is mid-tier websites, with ranks between 1M and 10M. These sites contribute an additional 17% of traffic to the periphery. More importantly, we observe that these sites serve an intermediary role in user navigation. Users tend to link from a core website to a mid-tier website, spend some time on that site before using it as a stepping stone to travel further out to a periphery site. In total, these three patterns of navigation account for more than two-thirds of all user visits to new periphery websites.

In addition to the primary gateways, we find that users link from streaming, adware, and adult websites less frequently, but at a disproportionate rate relative to the frequency of users visiting those websites. This bears close resemblance to the types of websites that are overrepresented in the distribution of periphery sites. Interestingly, 29.7% of visits to new free streaming sites originate from core entertainment websites, an indicator that users might first search through normal channels for the content they want before resorting to other free streaming options if not available. Of the other main sources of traffic, we observe a substantial proportion of the visits that come from crowdworking websites are directed to small opinion survey, user testing, and rewards websites which are often a part of crowdworking tasks. Traffic from social media sites lead to one of the most diverse sets of websites by type, second only to search. Users who navigate from retail websites often land on other smaller shopping websites or business homepages. However, it is unclear if these users are intentionally navigating between these specific sites, such as to compare products, or if they clicked on an advertisement that led them there.

6 DISCUSSION

In interpreting our findings, we first consider the limitations of this work before turning to implications for future research and the trends in web browsing observed during this study.

6.1 Limitations

The results of our study are limited in several ways. First, our data is based on the computer and browsing behavior of users on desktop and laptop computers. Given the predominate use of individual apps rather than a browser on tablets and smartphones, browsing behavior on mobile devices is likely very different. Second, our sample is based in the United States and primarily located in one major metropolitan area. Although there were several users who regularly visited international websites or those in different languages, the overall geographic and cultural composition of our sample limits the scope of browsing behavior that we observe. Users in other countries are likely to visit an overlapping, but largely different, set of websites on a regular basis. These differences likely hold across cultural differences within the United States, albeit to a lesser extent. As such, we are likely to find different gateways being used to access periphery websites across disparate populations. Third, our data was collected from users who owned computers running Microsoft Windows operating systems and whose browsing was done using Google Chrome. Fourth, despite the SBO data covering a substantially longer period of time than previous studies, our findings still represent a temporal snapshot of user behavior. As we have demonstrated in comparing to previous work, user browsing behavior has changed over time. With the continued shift in the population of web users, the trend of moving applications directly into the browser, and the emergence of new browsing technology, it is inevitable that user behavior will continue to evolve.

6.2 Implications for User-Based Browsing Research

Although we observe several trends in user behavior and substantial changes over time, like previous work we find no clear definition of an average user. In the development of websites and web applications, creators often design and test products for specific types of users. Related research also falls into the same pattern when testing for security and usability across discrete groups of subjects. Based on our findings, this characterization can be problematic, leading to software built and tested well for some users and not for others. Instead, we recommend that developers and researchers treat internet users as a spectrum rather than a type.

This recommendation has an important implication for the use of a browsing session in future web research. These sessions, which are extensively used to artificially segment a user's web browsing for analysis, are often separated based on a fixed period of user inactivity. However, given that we observe a broad range of browsing habits across users, a single definition of a browsing session may be not be appropriate. Future research is needed to assess browsing sessions as a concept, evaluate whether a different paradigm is required, and determine the optimal unit of measure for analyzing user web browsing.

Until techniques are fully developed to address the noise in server-side measurements stemming from tabbed browsing and automatically generated web requests [54], future research on user browsing behavior should focus on client-side data collection. Conclusions regarding browsing behavior should be drawn from server-side measurements cautiously. However, we acknowledge that client-side measurement studies do pose their own significant challenges. Although server-side measurements can be scaled to thousands of users, such efforts using client-side collection methods are more expensive and technically difficult to manage. As such, future research that helps to mitigate server-side measurement errors would be immensely valuable to the field.

Based on our experience running this study and the limitations of current server-side collection methods, we recommend that studies of user web browsing satisfy the following six conditions. First, collection should be conducted on the user's own machine to obtain accurate measurements and maintain ecological validity. Browser extensions were particularly useful in deploying sensors on the user's computer while minimizing the technical challenges often associated with client-side measurements. Second, the platform for data collection should be lightweight enough so as to prevent interference with the user's normal browsing. Third, web requests initiated by the user must be distinguishable from those generated automatically by web pages. Although Vassio et al. [54] demonstrate a viable method of discerning user web requests from HTTP traffic, this option is not preferable to differentiating requests on the client side if available. The growth of HTTPS traffic, the constant evolution of the internet, and the changes in user behavior over time make this classification task very difficult and will inevitably introduce noise. Fourth, the use of tabs should be captured, especially the events where a user switches the tab that is currently in view. Fifth, events when the web browser application starts, stops, and moves in and out of the foreground should be recorded to bound user browsing time. Sixth, ideally some form of user interaction such as mouse movements, keystrokes, or touch should be collected to determine when the user is actively browsing or has just left their browser open on the screen.

6.3 Implications for Internet-Related Research

The periphery of the internet poses important questions and challenges across several areas of future research. From a security perspective, the greater likelihood of users encountering adware sites on the periphery supports previous findings associating malicious content and low-traffic websites [24, 33, 35, 46]. To better inform and protect users, future research should investigate possible interventions that could be deployed along the common pathways leading to periphery websites. In addition, the substantial overrepresentation of websites related to alternative science and alternative health on the periphery of the internet has important implications for the dissemination of science. Future research in this area will help illuminate the situations in which users encounter this kind of alternative information for the first time. Such insights can provide researchers with a better understanding of how users evaluate information from different sources and how their consumption of information may change after being exposed to different ideas. Our findings indicate that users get to websites containing alternative information from many sources, and social media, although a substantial contributor, is not the primary source. The relationship between other contributing sites, particularly search engines, and alternative information should be studied in greater detail. Last, we observe a large proportion of visits to free streaming websites on the periphery of the internet. This is a negative sign for the entertainment industry and piracy researchers. However, with approximately 30% of visits originating from core entertainment websites first, at least some users appear to be resorting to free streaming sites rather than actively seeking them out to begin with. In these cases, it may be that the show the user is interested in is unavailable or too expensive through legitimate channels. Further investigation into user decision making is warranted.

6.4 Trends in Web Browsing

Finally, two growing trends in web browsing demand further investigation. First, the increasing use of smartphone and tablet devices to browse the internet represents an important set of research that runs parallel to our own. Early mobile browsing studies have begun to draw comparisons and identify differences between mobile and personal computer browsing [11, 50]. However, changes in technology, especially in the area of mobile devices, have rapidly advanced in the past 8 years.

In addition to the internet evolving over this period, mobile devices have become much more powerful and mobile internet speeds have greatly increased. Similar to our findings for desktop users, these developments likely correspond to changes in user browsing. As such, an update on the state of mobile browsing behavior is critically needed.

Second, applications are moving directly into the browser in increasingly greater numbers. This trend has been led in part by collaboration platforms like Google Docs and Microsoft Office Online. In fact, this article was written almost exclusively within a web browser. The recent development of technology like WebAssembly, which enables an even broader range of applications, such as arcade video game emulators, to run in the browser as they would on “native” hardware, points to the continuation of this trend. In the near future, we might find that the browser consumes what remains of users’ non-browser activity. As the browser continues to become the focal point of personal computer use, issues of security, privacy, and usability grow. Addressing these challenges relies on a foundation of knowledge about how users browse the web. Although our work contributes to this effort, it only represents a step in the larger longitudinal study of web browsing. Future contributions to our understanding of user browsing behavior will be needed.

7 CONCLUSION

This study contributes an important update to our knowledge of user web browsing that was established in a series of studies conducted between 1994 and 2011. We identified the key changes in user behavior over time including the large-scale adoption of tabbed browsing, the near obsolescence of back navigation, the expanded use of the address bar as a navigational tool, and the increase in the rate of revisitation to pages already seen before. In addition to identifying these trends, we provided evidence in support of previous findings that there is no “average internet user” and demonstrated the need to account for tabbed browsing in web measurements. In line with broader trends, we observed that 75% of users spend more time on their web browser than all other applications on their computer combined. Within the browser, we proposed that user behavior can be represented using a core-periphery model where users spending the majority of their time browsing a small set of popular websites. In the periphery of the internet, we detected an overrepresentation of visits to questionable websites related to free online streaming, adware, gambling, adult content, and alternative science and health. Motivated by this, we identified three primary gateways to the periphery: search engines, other periphery websites, and mid-tier intermediary sites. These gateways provide an opportunity for interventions to better protect users as they browse the web.

As an increasing number of applications are subsumed by browsers and as mobile web browsing continues to grow, the context in which users browse the internet will continue to evolve. The online activities and challenges that users engage in tomorrow will be different from today, and their behaviors in navigating them will change accordingly. Continued work within the research community to study these changes and compare browsing behavior in different contexts is crucial for the future development of informed internet-related technology, policy, and research.

A APPENDIX

A.1 Application Categories

The hierarchical category structure was developed and manually coded by the research team (Table 8).

Table 8. Hierarchical Category Structure for Labeling Applications That Participants Used on Their Computer

Category Code 1	Category Code 2	Types of Applications
Communication	Communication	Email, chat clients, video conferencing, e-cards
Entertainment	Adult	Pornographic videos
	Animation and Comics	Manga readers, animation tools
	Gambling	Casino applications, betting tools
	Gaming	Video games, streaming software, mod editors
	Literature	E-Readers
	Media Player	Adobe Flash, video players
File System and Storage	Streaming	Netflix, Hulu, Spotify desktop applications
	File System and Storage	File explorer, cloud storage, file transfer
Personal	Audio Editing	Audio and music editing
	Finances	Investment tools, cryptocurrency wallets
	Food and Drink	Recipes, cooking instruction
	Hobbies	Drone software, amateur radio, music tools
	Home and Garden	Home design, 3D designers
	Image Editing	Photo viewers, photo editors, graphic design
	Lifestyle	Fitness, calorie trackers, wearables
	News and Media	News articles, weather apps
	Reference	Maps, dictionary and thesaurus
	Shopping	Retail portals, discount tools, app stores
	Social	Social media desktop applications
System Operations	Video Editing	Video editing, media converting or burning
	Connected Devices	Printers, scanners, USB devices
	File Sharing	Torrents, download managers
	Install/Uninstall	Installing and uninstalling programs
	Navigation	Start menu, search
	Notifications/Popups	Notification windows
	Security	Antivirus, malware removal tools
	System Configuration	Display, audio, power, and network settings
Unknown	System Repair	Data backup, data recovery, system recovery
	Unknown	Typically unreadable sensor information
Web Browsing	Web Browsing	Chrome, Firefox, Internet Explorer, Opera, Tor
Work	Business	Billing, payroll, accounting tools
	Computers and Programming	Program editors, IDEs, debuggers
	Office Tools	Text, word, spreadsheet, and presentation editors
	Productivity	Productivity utilities, data analysis software
	Remote Connectivity	Remote desktop, SSH tools
	Science	3D Modeling, scientific imaging and sensors
	Social Science	Decision and simulation tools

A.2 Website Categories

The hierarchical coding used in this study was based on the DMOZ website categorization structure (Table 9). These categories were simplified down into 65 codes across 15 categories. DMOZ data labeled in this manner was used to train the website category classifier. The 15 high-level categories correspond the types of web activities that a user engaged in.

Table 9. Hierarchical Category Structure for Labeling Websites the User Visited While Browsing the Internet

Category/Activity	Website Codes	Description
Adware	Adware	Unwanted adware websites
Business	Advertising, Business, Government and Politics, Industrial, Legal, Organizations, Non-Profit	Business, government organizations, and non-profits
Communication	Communication, Email	Chat clients, video conferencing, VoIP, and email
Computers	Computers and Electronics	Programming, technical docs, how-to forums, cloud computing
Entertainment	Adult, Animation and Comics, Entertainment, Gambling, Gaming, Literature and Writing, Movies and Television, Music and Radio, Performing Arts, Sports, Streaming, Visual Arts	A variety of entertainment sites, streaming music and video, online gaming, celebrity news, gambling, and pornography
Financial	Financial	Banking, investing, financial advice and news
Health	Alternative Health, Healthcare, Medicine and Conditions, Mental Health and Support, Reproductive Health, Senior Health	Sites related medical conditions, and the healthcare industry; includes mental health, addictions, and support groups
News	News and Media	News, newspapers, weather
Lifestyle	Alternative Science, Ethnic and Cultural, Family, Food and Drink, Genealogy, Hobbies, Home and Garden, LGBTQ, Lifestyle, Outdoors, Personal Pages, Pets, Real Estate, Relationships and Dating, Religion, Services, Social Issues, Subcultures, Travel, Vehicles	Sites related to the users personal life, hobbies, lifestyle choices, and beliefs
Reference	Reference	Wikipedia, Q&A, maps, dictionary
Search	Search Engines	Major search engines
Shopping	Retail, Rewards and Freebies	Shopping and rewards programs
Social	Social Media	Major social media sites
Web Portal	Web Portal	Sites that bring information from many domains together
Work	Crowdworking, Education, Humanities, Job Search, Productivity, Remote Connectivity, Science, Social Sciences	Sites related to the user's occupation or education

ACKNOWLEDGMENTS

Thank you to Sarah Pearman and Jeremy Thomas for their work on the Security Behavior Observatory. Their help answering questions and providing technical assistance was incredibly valuable throughout the course of this research.

REFERENCES

[1] Xueli An, Fahim Kawsar, and Utku Günay Acer. 2017. Profiling and predicting user activity on a home network. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous'17)*. ACM, New York, NY, 494–503. <https://doi.org/10.1145/3144457.3145502>

[2] Fabricio Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgilio Almeida. 2009. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement (IMC'09)*. ACM, New York, NY, 49–62. <https://doi.org/10.1145/1644893.1644900>

[3] Stephen P. Borgatti and Martin G. Everett. 2000. Models of core/periphery structures. *Social Networks* 21, 4 (2000), 375–395. [https://doi.org/10.1016/S0378-8733\(99\)00019-2](https://doi.org/10.1016/S0378-8733(99)00019-2)

[4] Davide Canali, Leyla Bilge, and Davide Balzarotti. 2014. On the effectiveness of risk prediction based on users browsing behavior. In *Proceedings of the 9th ACM Symposium on Information, Computer, and Communications Security (ASIACCS'14)*. ACM, New York, NY, 171–182. <https://doi.org/10.1145/2590296.2590347>

[5] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. 2007. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. ACM, New York, NY, 423–430. <https://doi.org/10.1145/1277741.1277814>

- [6] Lara D. Catledge and James E. Pitkow. 1995. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems* 27, 6 (1995), 1065–1073. [https://doi.org/10.1016/0169-7552\(95\)00043-7](https://doi.org/10.1016/0169-7552(95)00043-7)
- [7] Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. 2015. Why do social media users share misinformation? In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'15)*. ACM, New York, NY, 111–114. <https://doi.org/10.1145/2756406.2756941>
- [8] Zhicong Cheng, Bin Gao, and Tie-Yan Liu. 2010. Actively predicting diverse search intent from user browsing behaviors. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, NY, 221–230. <https://doi.org/10.1145/1772690.1772714>
- [9] Flavio Chierichetti, Ravi Kumar, Prabhakar Raghavan, and Tamas Sarlos. 2012. Are web users really Markovian? In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. ACM, New York, NY, 609–618. <https://doi.org/10.1145/2187836.2187919>
- [10] J. Clement. 2019. Projected retail e-commerce GMV share of Amazon in the United States from 2016 to 2021. *Statista*. Retrieved August 17, 2021 from <https://www.statista.com/statistics/788109/amazon-retail-market-share-usa/>.
- [11] Yanqing Cui and Virpi Roto. 2008. How people use the web on mobile devices. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, New York, NY, 905–914. <https://doi.org/10.1145/1367497.1367619>
- [12] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559. <https://doi.org/10.1073/pnas.1517441113>
- [13] DMOZ. 2020. The Directory of the Web. Retrieved August 17, 2021 from <https://dmoz-odp.org/>.
- [14] Patrick Dubroy and Ravin Balakrishnan. 2010. A study of tabbed browsing among Mozilla Firefox users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY, 673–682. <https://doi.org/10.1145/1753326.1753426>
- [15] Serge Egelman and Eyal Peer. 2015. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop (NSPW'15)*. ACM, New York, NY, 16–28. <https://doi.org/10.1145/2841113.2841115>
- [16] Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion Proceedings of the the Web Conference 2018 (WWW'18)*. 595–602. <https://doi.org/10.1145/3184558.3188730>
- [17] Alain Forget, Saranga Komanduri, Alessandro Acquisti, Nicolas Christin, Lorrie Cranor, and Rahul Telang. 2014. *Security Behavior Observatory: Infrastructure for Long-Term Monitoring of Client Machines*. Technical Report CMU-CyLab-14-009. Carnegie Mellon University.
- [18] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. 2018. Away from prying eyes: Analyzing usage and understanding of private browsing. In *Proceedings of the 14th USENIX Conference on Usable Privacy and Security (SOUPS'18)*. 159–175.
- [19] Kirstie Hawkey and Kori Inkpen. 2005. Web browsing today: The impact of changing contexts on user activity. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems (CHI EA'05)*. ACM, New York, NY, 1443–1446. <https://doi.org/10.1145/1056808.1056937>
- [20] Julia Hoxha. 2012. Semantic formalization of cross-site user browsing behavior. *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence (WI'12)*. <https://doi.org/10.1109/WI-IAT.2012.232>
- [21] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. 2007. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, New York, NY, 151–160. <https://doi.org/10.1145/1242572.1242594>
- [22] Jeff Huang, Thomas Lin, and Ryen W. White. 2012. No search result left behind: Branching behavior with browser tabs. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*. ACM, New York, NY, 203–212. <https://doi.org/10.1145/2124295.2124322>
- [23] Jeff Huang and Ryen W. White. 2010. Parallel browsing behavior on the web. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT'10)*. ACM, New York, NY, 13–18. <https://doi.org/10.1145/1810617.1810622>
- [24] J. James, L. Sandhya, and C. Thomas. 2013. Detection of phishing URLs using machine learning techniques. In *Proceedings of the 2013 International Conference on Control Communication and Computing (ICCC'13)*. 304–309. <https://doi.org/10.1109/ICCC.2013.6731669>
- [25] Robert Kraut, William Scherlis, Tridas Mukhopadhyay, Jane Manning, and Sara Kiesler. 1996. The HomeNet field trial of residential internet services. *Communications of the ACM* 39, 12 (Dec. 1996), 55–63. <https://doi.org/10.1145/240483.240493>
- [26] Ravi Kumar and Andrew Tomkins. 2010. A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, NY, 561–570. <https://doi.org/10.1145/1772690.1772748>
- [27] Litmus Labs. 2020. Email Client Market Share. Retrieved August 17, 2021 from <https://emailclientmarketshare.com/>.

- [28] Choudur Lakshminarayan, Ram Kosuru, and Meichun Hsu. 2016. Modeling complex clickstream data by stochastic models: Theory and methods. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW'16 Companion)*. 879–884. <https://doi.org/10.1145/2872518.2891070>
- [29] Janette Lehmann, Mounia Lalmas, Georges Dupret, and Ricardo Baeza-Yates. 2013. Online multitasking and user engagement. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM'13)*. ACM, New York, NY, 519–528. <https://doi.org/10.1145/2505515.2505543>
- [30] N. Leontiadis, T. Moore, and N. Christin. 2011. Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In *Proceedings of the 20th USENIX Security Symposium (USENIX Security'11)*. 281–298.
- [31] Fanny Lalonde Lévesque, Sonia Chiasson, Anil Somayaji, and José M. Fernandez. 2018. Technological and human factors of malware attacks: A computer security clinical trial approach. *ACM Transactions on Privacy and Security* 21, 4 (July 2018), Article 18, 30 pages. <https://doi.org/10.1145/3210311>
- [32] Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 379–386. <https://doi.org/10.1145/1835449.1835513>
- [33] Long Lu, Roberto Perdisci, and Wenke Lee. 2011. SURF: Detecting and measuring search poisoning. In *Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS'11)*. ACM, New York, NY, 467–476. <https://doi.org/10.1145/2046707.2046762>
- [34] Bonnie Ma Kay and Carolyn Watters. 2008. Exploring multi-session web tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. ACM, New York, NY, 1187–1196. <https://doi.org/10.1145/1357054.1357243>
- [35] J. McGahagan, D. Bhansali, M. Gratian, and M. Cukier. 2019. A comprehensive evaluation of HTTP header features for detecting malicious websites. In *Proceedings of the 2019 15th European Dependable Computing Conference (EDCC'19)*. 75–82. <https://doi.org/10.1109/EDCC.2019.00025>
- [36] B. McKenzie and A. Cockburn. 2001. An empirical analysis of web page revisitation. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*. 1–9.
- [37] Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. 2016. Uncovering task based behavioral heterogeneities in online search behavior. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, New York, NY, 1049–1052. <https://doi.org/10.1145/2911451.2914755>
- [38] H. Mekky, R. Torres, Z. Zhang, S. Saha, and A. Nucci. 2014. Detecting malicious HTTP redirections using trees of user browsing activity. In *Proceedings of the 2014 IEEE Conference on Computer Communications (INFOCOM'14)*. 1159–1167. <https://doi.org/10.1109/INFOCOM.2014.6848047>
- [39] Filippo Menczer. 2016. The spread of misinformation in social media. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW'16 Companion)*. 717. <https://doi.org/10.1145/2872518.2890092>
- [40] A. Narayanan and Vitaly Shmatikov. 2006. How to break anonymity of the Netflix Prize dataset. arXiv:cs/0610105.
- [41] Terry Nelms, Roberto Perdisci, Manos Antonakakis, and Mustaque Ahamad. 2015. WebWitness: Investigating, categorizing, and mitigating malware download paths. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security'15)*. 1025–1040. <https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/nelms>.
- [42] Nam P. Nguyen, Guanhua Yan, My T. Thai, and Stephan Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference (WebSci'12)*. ACM, New York, NY, 213–222. <https://doi.org/10.1145/2380718.2380746>
- [43] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. 2007. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*. ACM, New York, NY, 597–606. <https://doi.org/10.1145/1240624.1240719>
- [44] Open PageRank. 2020. Home Page. Retrieved August 17, 2021 from <https://www.domcop.com/openpagerank/>.
- [45] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. 2017. Let's go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS'17)*. ACM, New York, NY, 295–310. <https://doi.org/10.1145/3133956.3133973>
- [46] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. 2018. Predicting impending exposure to malicious content from user behavior. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS'18)*. ACM, New York, NY, 1487–1501. <https://doi.org/10.1145/3243734.3243779>
- [47] StatCounter. 2020. Social Media Stats United States of America. Retrieved August 17, 2021 from <https://gs.statcounter.com/social-media-stats/all/united-states-of-america>.
- [48] Linda Tauscher and Saul Greenberg. 1997. Revisitation patterns in World Wide Web navigation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'97)*. ACM, New York, NY, 399–406. <https://doi.org/10.1145/258549.258816>

- [49] Paul Thomas. 2014. Using interaction data to explain difficulty navigating online. *ACM Transactions on the Web* 8 (Nov. 2014), 1–41. <https://doi.org/10.1145/2656343>
- [50] Chad Tossell, Philip Kortum, Ahmad Rahmati, Clayton Shepard, and Lin Zhong. 2012. Characterizing web use on smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. ACM, New York, NY, 2769–2778. <https://doi.org/10.1145/2207676.2208676>
- [51] United States Bureau of Labor Statistics. 2019. Employment by Detailed Occupation. Retrieved August 17, 2021 from <https://www.bls.gov/emp/tables/emp-by-detailed-occupation.htm>.
- [52] United States Census Bureau. 2017. ACS Demographics and Housing Estimates. Retrieved August 17, 2021 from <https://data.census.gov/cedsci/table?q=ACS%20Demographics%20housing&tid=ACSDP1Y2017.DP05>.
- [53] United States Census Bureau. 2019. Educational Attainment in the United States: 2018. Retrieved August 17, 2021 from <https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html>.
- [54] Luca Vassio, Idilio Drago, Marco Mellia, Zied Ben Houidi, and Mohamed Lamine Lamali. 2018. You, the web, and your device: Longitudinal characterization of browsing habits. *ACM Transactions on the Web* 12, 4 (Sept. 2018), Article 24, 30 pages. <https://doi.org/10.1145/3231466>
- [55] W3Counter. 2020. Web Browser Market Share. Retrieved August 17, 2021 from <https://www.w3counter.com/globalstats.php>.
- [56] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y. Zhao. 2013. You are how you click: Clickstream analysis for Sybil detection. In *Proceedings of the 22nd USENIX Security Symposium (USENIX Security'13)*. 241–256. <https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/wang>.
- [57] Gang Wang, Xinyi Zhang, Shiliang Tang, Christo Wilson, Haitao Zheng, and Ben Y. Zhao. 2017. Clickstream user behavior models. *ACM Transactions on the Web* 11, 4 (July 2017), Article 21, 37 pages. <https://doi.org/10.1145/3068332>
- [58] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y. Zhao. 2016. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, New York, NY, 225–236. <https://doi.org/10.1145/2858036.2858107>
- [59] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine* 240 (2019), 112552. <https://doi.org/10.1016/j.socscimed.2019.112552>
- [60] Yi-Min Wang, Ming Ma, Yuan Niu, and Hao Chen. 2007. Spam double-funnel: Connecting web spammers with advertisers. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, New York, NY, 291–300. <https://doi.org/10.1145/1242572.1242612>
- [61] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter* 21, 2 (Nov. 2019), 80–90. <https://doi.org/10.1145/3373464.3373475>
- [62] Yunjuan Xie and Vir V. Phoha. 2001. Web user clustering from access log using belief function. In *Proceedings of the 1st International Conference on Knowledge Capture (K-CAP'01)*. ACM, New York, NY, 202–208. <https://doi.org/10.1145/500737.500768>
- [63] Haimo Zhang and Shengdong Zhao. 2011. Measuring web page revisitation in tabbed browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, 1831–1834. <https://doi.org/10.1145/1978942.1979207>
- [64] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 1 (NIPS'15)*. 649–657.

Received February 2021; revised June 2021; accepted June 2021