Understanding User Surfing on the Webgraph from Passive Measurements

Luca Vassio¹, Zied Ben Houidi², Mohamed Lamine Lamali², Idilio Drago¹, Marco Mellia¹, Renata Teixeira³

¹ Politecnico di Torino, Italy
{lucavassio, idilio.drago, marco.mellia}@polito.it
² Alcatel-Lucent Bell Labs, France
{zied.benhoudi, mohamed.lamine.lamali}@alcatel-lucent.com
³ Inria, France
renata.teixeira@inria.fr

Abstract—The webgraph represents web pages as vertices and links as edges. Webgraphs are typically built by active crawling websites and, as such, miss an aspect that is fundamental to several applications: how users actually surf the web. This paper revisits how users explore the webgraph. We define for each user the surfgraph, which includes only pages visited and links followed. We rely on HTTP traces collected at a large European ISP to extract surfgraphs, and characterize and quantify how web content is consumed. Browsing behavior varies widely among users and depending on the device they use. We study structural properties of the surfgraphs, revealing two common patterns of users’ exploration of the webgraph, namely extensive exploration and intentional visits. We show how users mainly find content, with search engines and social networks playing a key role. When users visit pages promoted by a social network, however, they often do not engage in further browsing.

Index Terms—Webgraph, surfing behavior, social network, search engine, passive measurements.

I. INTRODUCTION

The webgraph is a directed graph, in which vertices represent web content (i.e., web pages or domains) and where there exists a directed edge from vertex A to B if A contains a URL that points to B. Such a graph is typically obtained thanks to robots that actively crawl the web. The webgraph has been extensively studied in the literature [1], [5], [15] for a variety of purposes like improving search engines [3], [14], or comparing rankings in the web [7], [12].

The webgraph built by robots misses a key aspect: the users’ interaction with the web, or how users explore the webgraph. This paper characterizes such interactions by defining the surfgraph for each user: the subset of the webgraph visited by the user in a limited time interval. The vertices in the surfgraph are the webpages that a user visited, and the edges are links that a user followed when surfing the web.

Surfgraphs are related to clickstreams, which capture the path users take when navigating through sites. Although clickstreams have been studied in the past [2], [6], [13], these studies either focused on data from the early 2000s or mostly on user interactions with search engines. The evolution of the Internet, however, changed how users interact with the web. For example, we witness the predominance of a handful of big players [9] and the rise of mobile devices in recent years. It is unclear how such changes influence users’ walks on the webgraph. Given the difficulties of obtaining large-scale traces of users’ interactions with the web, general properties of the portion of the webgraph effectively visited by each user are mostly unknown. This knowledge is fundamental for a variety of applications, from the design of efficient crawling strategies to advertisement placement; from the detection of malicious activity to sociological studies.

Building surfgraphs is challenging as we must be in a position to observe users as they surf the web. Passive measurements collected from network links offer a privileged vantage point in this respect. In this work we use a one-week long HTTP log collected in 2013 at large residential European ISP to build surfgraphs for approximately 20,000 households (Sec. III). We then perform a thorough analysis of these surfgraphs to characterize and quantify how users consume content in the modern web.

Our study of surfgraphs yields a number of interesting findings. First, we show that there is a large diversity on how users interact with the web (Sec. IV). Some users barely visited any page during the week of our study, whereas others visited more than 200 pages per day. The number of visits depends on the device: users visit on average more pages from PCs than smartphones. Second (Sec. V), our analysis of surfgraphs shows that users browse the web in two modes: extensive exploration, where they explore a high number of pages, often with the help of search engines and social networks, and intentional visits, where they directly target websites. In both browsing modes, however, users often stay confined in a very low number of domains. Lastly, we study how users discover web content (Sec. VI). As expected, search engines and, to a lesser extent, Online Social Networks (OSNs) play a key role in helping users exploring content. Interestingly, when users visit a page promoted by a social network, they often do not continue browsing from that page. This result suggests possible limits of social media marketing. Although some of our findings are intuitive, our contribution lies in quantifying these intuitions for a large user population.

II. RELATED WORK

Previous work on clickstream analysis has mostly characterized user browsing sessions using particular or limited
vantage points. For example, Huberman et al. [11] study the number of pages that a user visits within a single website, Kammenhuber et al. [13] study the relations between queries on a search engine and followed paths, while Olmedilla et al. [16] focus primarily on the categories of websites that mobile users access. In contrast, we rely on aggregated HTTP traces from a large ISP to build surfgraphs, which summarize all browsing activity of each user, and focus on understanding general properties of the surfgraphs. The work of Xie et al.[17] is the closest to ours. It presents a methodology to rebuild browsing sessions from packet traces and illustrates the methodology in a case study, in which packet traces are used to uncover dominant players in the web, characterize session lengths and compare mobile and wireline traffic. We instead characterize the structure of surfgraphs as a whole, not only to pinpoint dominant players and content promoters, but also to understand how users interact with such players, highlighting users’ engagement with websites connected to promoters few hops away.

III. SURFGRAPH EXTRACTION

We extract surfgraphs from passive measurements collected from a residential European ISP network. A probe running Tstat [8] has been installed in a Point of Presence, where it observes the traffic of \( \approx 20,000 \) households, each using (and uniquely identified by) a static IP address assigned per each gateway. Users connect to the Internet at home, using a single access gateway, via Ethernet or WiFi. We use the HTTP traces that Tstat exports. For each HTTP request, the traces contain a timestamp, the client IP address, the URL, the referer, and the user-agent field. The data collection and analysis take the following steps to protect users' privacy. First, client IP addresses are anonymized. Second, all URLs are truncated to remove any query parameters. Last, we save only the information strictly needed to build the surfgraphs.

A. Dataset

We focus on a one-week long trace collected in 2013, for which Tstat logged more than 190 million HTTP requests. Getting surfgraphs from the raw HTTP traces is challenging because modern web pages are fairly complex. To render a page, browsers have to fetch all the objects that compose it, because modern web pages are fairly complex. To render a page, browsers have to fetch all the objects that compose it, and able to work without access to full packet traces. This methodology can reach a precision of 95% with a recall of 80%. Using the most precise variation, we identify 5 million user visits, corresponding to about 3.7 million distinct URLs. In this paper we study surfgraphs obtained by considering each via a different URL. We employ the method developed in our previous work to identify URLs corresponding to actual user visits [10], because it is as accurate as the state of the art [17] when compared on the same ground truth trace, and able to work without access to full packet traces. This methodology can reach a precision of 95% with a recall of 80%. Using the most precise variation, we identify 5 million user visits, corresponding to about 3.7 million distinct URLs. In this paper we study surfgraphs obtained by considering the entire week of data collection. We plan to study how surfgraphs evolve over time in future work (Sec. VII). Distinct URLs constitute the surfgraph vertices. We extract edges by using the referer field: a directed edge exists from the URL in the referer to the requested URL. If a URL request does not include referer, this vertex is inserted into the graph without any edge. This process identifies more than 3 million edges. The surfgraphs have less edges than nodes because of the presence of pages accessed directly from the browser, e.g., using a bookmark, typing in the URL, or using apps in mobile terminals.

B. Surfgraph Types and Annotations

Our goal is to study users’ interactions with the web. However, our data do not provide any user identifier. The IP addresses of the access gateways present in our data aggregate all devices and browsers of several users in the households. We thus study two types of graphs to approximate the characteristics of users’ surfgraphs:

(i) Household graph: It captures how users living in the household interact with the web. We generate one graph per anonymized client IP address, which always corresponds to the same household, thanks to the static allocation policy of IP addresses of the ISP. We study \( \approx 20,000 \) households.

(ii) Surfer graph: We define a surfer as a user interacting with the web on a given application, identified by the user-agent. We extract surfers by generating one graph for each pair of IP address and user-agent. We study \( \approx 40,000 \) surfers. Obviously, a surfer is not equivalent to a single physical person since the latter could use different browsers and have multiple associated surfers, or different people may use the same device or browser to surf the web.

For both types of graphs, we also create domain graphs, where all web pages belonging to the same domain are collapsed in the same vertex.

Our analysis studies the influence of device types on how users interact with the web. We annotate surfgraphs with the device type – PCs, tablets or smartphones – based on the Python User-Agent module.\(^1\) In our dataset, 82% of visits come from PCs, 6.1% from smartphones, and 6.5% from tablets. We ignore the remaining 5.4% of visits that are not classified, possibly coming from smart-TVs, bots, and browsers with unrecognized user-agent.

C. Data Limitation

Our construction of surfgraphs from HTTP traces comes with two caveats.

1) Our analysis is limited to the non-encrypted part of the web. We cannot study pages accessed with HTTPS. To gauge the effect of HTTPS, we collect additional statistics using the DN-Hunter [4] plugin of Tstat. DN-Hunter logs each TCP flow with the server Fully Qualified Domain Name (FQDN) the client resolved via previous DNS queries. This allows to know which service (FQDN) the client contacted when accessing a given server IP address. HTTPS was only 7% of the total number of FQDNs in our 2013 dataset (serving 12% of bytes). We plan to repeat the analysis with newer traces to see how this evolved. The major HTTPS services are Google and Facebook, accounting for 70% of the HTTPS traffic. Google and Facebook appear to us as “walled” domains and we miss what users do when inside. Yet, as soon as a user

\(^1\)http://pypi.python.org/pypi/user-agents
clicks on a link pointing to any page outside those domains, the browser is directed first to an internal HTTP page, then to the actual external link. As such, the referer field of the second request allows us to study when users leave those domains. Only the remaining HTTPS traffic (3.6% of the overall traffic) is thus completely invisible to us.

2) The surfgraphs we study capture how the users of the ISP we monitor interact with the web. Users in different parts of the world will certainly access other parts of the web. However, our monitoring covers tens of thousands of households and our data appears fairly representative of the country we monitor when comparing penetration rates of OSs and browsers with available market studies.

IV. AMOUNT OF CONSUMED CONTENT

A. Visited Pages and Domains

We study the number of pages and domains visited per surfers and households to quantify users’ content consumption. Fig. 1(a) and Fig. 1(b) summarize how surfers and household respectively interact with the web. We plot the CDFs of (i) the total number of visits; (ii) the number of unique visited pages; and (iii) the number of unique visited domains.

The average number of visits during the week is 87 for surfers and 385 for households, which is surprisingly low. However, numbers are very disperse. On the one hand, we see in Fig. 1(b) that 20% of the households visit less than 10 pages in the entire week. On the other hand, a significant number of users heavily explores the web, with few surfers visiting up to 2,000 pages (Fig. 1(a)).

When we compare the number of visits to the number of unique pages and domains visited by households and surfers, we see that users frequently revisit the same page and domain. Finally, we observe that the number of surfers per household is generally low, with 33% of the households having a single surfer and 90% of them having less than 10 surfers. Households with multiple surfers tend to present more activity, which influences the shape of the distribution in Fig. 1(b).

B. Impact of User Device

We investigate if the device type has an impact on surfing behavior. We focus on the page surfgraphs – the analysis for domains lead to similar conclusions. Fig. 2(a) and Fig. 2(b) show the CDFs of the total number of pages visited respectively per surfer and per household. Curves distinguish PCs, tablets, and smartphones.

The number of pages visited on smartphones is significantly smaller than the number of pages consumed from PCs (with tablets in between). While this is not surprising, we can precisely quantify it. For instance, surfers typically visit on average 4.7 times more pages on a PC than on a smartphone, whereas on a tablet they visit 2.3 times more pages than on a smartphone. Similar results are seen for households.

Notice the high percentage of surfers (and households) that visit a negligible number of pages. For example, the percentage
of households that visits 10 or less pages in a week is ≈ 20% for PCs, increasing to ≈ 40% for smartphones (Fig. 2(b)). This underlines the fact that mobile devices are used for occasional browsing at home. Indeed, PCs seem to be the favorite means to browse the web, with ≈ 20% of the PC surfers visiting more than 100 pages (Fig. 2(a)). Overall, results confirm that PCs and, to a lesser extent, tablets are the preferred means to surf the web.

V. STRUCTURE OF THE SURFGRAPHS

A. How Users Browse the Web

We now characterize properties of the surfers’ (page) surfgraphs. A surfgraph represents the portion of the webgraph that an individual user explores in a limited time interval. The structure of the surfgraph reveals whether the user browses many disjoint/independent areas of the webgraph or, instead, simply follows pages of a single region of the webgraph.

We consider for each surfgraph the number and sizes of Weakly Connected Components (WCC), i.e., any maximum sub-graph such that there exists a path between any pair of vertices, considering undirected edges. The number of WCCs reflects how many disjoint/independent sub-graphs a user explores; the WCC sizes tell about how extensive the user’s explorations are.

We observe that most surfgraphs exhibit a particular shape: a big WCC that includes a large fraction of the surfgraph, and several small WCCs. Fig. 3 represents the CDFs of two metrics across surfers who visited at least 10 pages. The Biggest WCC size/Graph size curve shows the ratio between the size of the biggest WCC and the size of the entire surfgraph. We can see that for ≈ 40% of the surfers, the biggest WCC represents more than one third of the entire graph. By manually inspecting the biggest WCCs, we notice that they usually include search engines and/or Online Social Networks, reinforcing their role as hubs that connect other visited pages that we will show in Sec. VI.

The Mean WCC size/Graph size curve in Fig. 3 depicts the ratio between the mean size of the remaining WCCs, if any, and the size of the graph. We can see that mean sizes of the remaining WCCs are rather small, confirming that there are many small WCCs. A new WCC component is created when there is no referer in a user’s visit to a page. Thus, we conjecture that the large number of small WCCs happens because of direct visits to websites (i.e., sites users are familiar with and reach without the assistance of any content discovery mechanisms). For instance, browser bookmarks, or applications that directly access web pages could result on several (small) WCCs. The click extraction methodology has 80% recall and thus could miss links, artificially creating isolated WCCs. However, we obtain similar results when tuning parameters of the heuristic to reach 95% recall, while lowering its precision.

We conclude that most surfers browse the web in two modes: (i) extensive exploration, where they explore a high number of pages in the same portion of the webgraph, often using search engines and social networks (big WCC); and (ii) intentional visits, where they target different domains, independently from popular content discovery means (small WCCs).

B. Characterizing Users’ Paths

To better gauge how extensive the users’ explorations are, we study how many consecutive pages, forming a path, surfers visit. We focus on the surfers’ biggest WCCs and compute their diameters: the longest among the finite directed shortest paths. The size of the diameter gives a hint on how far the user goes in his/her surfgraph.

The diameter is usually short as expected, since most surfers visited few pages during our data capture (recall Fig. 2(a)). On average, users explore longer paths on PCs (5.1 pages) than on tablets (4.1 pages) and smartphones (2.9 pages). However, 9% of the surfers have diameters with more than 10 pages – we notice some surfers with diameters of up to 1,000 pages.

Surprisingly, the number of domains in the diameters is very low and generally stable (mean of 1.8 domains per diameter). This is visible in Fig. 4, which shows boxplots for the number of domains in diameters. We plot independent boxes according to the number of pages in the diameters. We limit the y-axis to 5 to improve visualization, although we observed some few diameters with up to 24 domains, recognized as outliers. Even more surprising, the number of domains does not increase with the number of pages in the diameter. Note how the median number of domains is always 1 or 2, regardless of the number of pages in the diameters. Moreover, when focusing only on diameters with more than 30 pages, 57% of those diameters are
completely within a single domain. We conclude that both long and short explorations are usually restricted to few domains.

VI. CONTENT DISCOVERY

A. Promoting Content

The aggregated surfgraph: To get an insight into the structure of the surfgraphs, we focus on the household graphs for domains. Fig. 5 depicts all household graphs merged into a single graph, which we call the aggregated surfgraph. Only domains visited by at least 20 households are considered to improve visualization. The figure includes the ≈ 10,000 popular domains, out of the ≈ 190,000 original ones. The statistics reported hereinafter refer to the entire aggregated surfgraph.

We can see a giant connected component, in which a central vertex emerges. Not a surprise, Google is at the center of the graph. Its vertex and edges are enclosed by a circle. Next to Google, we see a second large circle enclosing Facebook and its neighborhood: 49% of the domains are directly connected to Google, while 5% are connected to Facebook. It is clear that these players have a central role in promoting (popular) content on the web. More interesting is to note that many domains are connected either to Facebook or to Google, but not to both: indeed, the intersection accounts for 2% of the number of domains. In sum, Facebook and Google are used to reach different types of content.

Finally, notice the large number of vertices in the border of the figure: those are disconnected from the giant component. The isolated vertices correspond to around 27% of the aggregated surfgraph. Manual inspection shows that those vertices mostly are (i) domains users reach without following links (e.g., directly, via bookmarks, or by using specific applications); and (ii) redirection that takes place without sending a referer (e.g., using HTML meta refresh), thus removing possible links in the graph.

Main content promoters: To pinpoint the main content promoters, we rank the vertices in the aggregated surfgraph by decreasing out-degree – i.e., the number of edges leaving the vertices. In our context, the out-degree of a domain measures the number of other domains to which it has generated visits to, and thus promoted content.

Fig. 6 depicts the top-15 domains ranked from the highest to the lowest out-degree. The y-axis (in log scale) is normalized by the total number of domains in the graph. We manually classify domains as (i) search engines; (ii) Online Social Networks (OSNs); or (iii) others. The figure shows that most entries are search engines or OSNs. As we already showed, Google has a dominant position, with about 10 times more out-neighbors than Facebook, which comes second. The rest of the domains are spread in a long tail, connecting a smaller number of domains. Curiously, Babylon, Delta Search, Ask, Conduit, and Avg (which are search engines related to toolbars) are among the top domain promoters. Although they are not very popular, once a user installs one of them, possibly all the user’s web searches come from them.

Main promoted domains: We also report in Fig. 7 the vertices in the aggregated surfgraph with highest in-degree, measuring thus the number of domains with followed links to them. Again, the y-axis (in log scale) is normalized by the total number of domains in the graph. Here the differences between the top domains are not as evident as for the out-degree chart, showing more uniform values. Websites hosting millions of
Fig. 6. Top-15 domains by out-degree. Note the logarithmic y-axis.

Fig. 7. Top-20 domains by in-degree. Note the logarithmic y-axis.

blogs (Wordpress) or videos (Youtube, Vimeo) are among the top ones. Since content hosted by these domains is often present in other websites (e.g., by cross referencing similar blogs, or by embedding videos in other pages), the number of pages that point to them is large. Interestingly, we spot many Adult websites (manually highlighted in Fig. 7). They follow a cross-referencing mechanisms according to which content from one website is cross-referenced in others, thus creating a dense and highly linked lattice that capture a huge number of visited domains.

B. User Interaction with Content Promoters

We now focus again on per-surfer/per-household surfgraphs. We aim to understand (i) how the usage of search engines and OSNs varies across users and devices; and (ii) how users behave after they leave such sites. We manually build a list of the top-80 most “out-connected” search engines and OSNs. To verify to what extent users explore the web after leaving any of these sites, we define reachability: we say that a destination page is reachable from a source page if there exists a path in the graph from the source to the destination page.

Fig. 8(a) shows the CDFs across households of the percentage of visited pages that are directly connected to (Direct curve) or reachable from (Reachable curve) a search engine. Fig. 8(b) shows these metrics for OSNs. We only consider households that visited at least 20 pages.

Focus first on the fraction of pages that are directly connected to a search engine. The fraction of pages per household that are directly connected to a search engine is mostly distributed between 20% and 60%. There is however a big disparity between households, and some clearly “search and click” many more than others. On average, 38% of all pages visited by the households are directly connected to a search engine.

Compare now the fraction of pages reachable and directly connected to search engines. Fig. 8(a) shows that there are many more pages that are reached from search engines than pages directly connected to them (57% on average). This result indicates that users keep browsing the web after leaving a search engine.

OSNs, on the other hand, are very different (check Fig. 8(b)). First, the percentage of pages to which they generate visits is much lower, with about 38% of households that never visit a pages starting from any OSNs. Second, almost all of the households visit a low fraction of pages starting from an OSN.

The Direct and Reachable curves in Fig. 8(b) are not too far from each other, suggesting the “addictiveness” of OSNs. Indeed, 59% of the pages connected to an OSN have no other connections – i.e., users often do not browse further after visiting a page promoted on an OSN. Moreover, in 81% of the cases, all pages visited starting from an OSN are from a single domain. We plan to further explore these observations in future work, since they could give some insights of the efficiency of social media marketing. Publishers who drive user traffic to their websites using social media marketing may not get the full users’ attention.

At last, we analyzed the impact of different devices (omitted for brevity). We find no major differences considering domains visited from search engines. However, smartphone and tablet users give more importance to OSNs than PC users: on average, only around 6% of the domains visited on PCs are connected to an OSN. This grows to 11% when smartphones are considered.

VII. CONCLUSIONS AND OPEN ISSUES

This paper presented a characterization of users’ behavior when surfing the webgraph. We highlighted typical amount of content users consume, characteristics of the structure of users’ navigation, how they interact with content promoted in social networks and search engines and the impact of user devices. Our results are a first step in understanding the surfgraphs, which will improve the knowledge of the web itself and help in designing novel applications that leverage such knowledge.

In future work we plan to extend our analysis to several vantage points and longer captures to understand how the exploration of the webgraph varies considering both space and time. The latter is important, especially given how much the
structure and complexity of web pages has changed over time. We also plan to crawl websites while surfgraphs are extracted from network traffic, and compare structural properties of the surfgraphs with the static Webgraph – e.g., to reveal properties of paths that are not followed by people. Our work relies on User Agent strings to define surfers, thus missing the relation between surfgraphs of a single user on different devices and client browsers. We plan to study strategies to correlate such disconnected surfgraphs of a single user. Finally, we will include HTTPS browsing in our framework. While identifying every page visited by users is not possible in such case, the information at domain level is still available and can provide a coarse view of the paths followed by users while surfing the HTTPS web.

REFERENCES