A URL-based Analysis of WWW Structure and Dynamics

Jeffery Kline*, Edward Oakes†, Paul Barford‡
*Hitwise. Email: jkline@hitwise.com
†University of California, Berkeley. Email: eoakes@berkeley.edu
‡University of Wisconsin, Madison. Email: pb@cs.wisc.edu

***Abstract*—Understanding the evolving characteristics of the World Wide Web is challenging due to its immense size and diversity. In this paper, we investigate Web structure and dynamics by analyzing over 1 trillion URLs requested during Web browsing by a 2 million person user panel over a period of 12 months. We begin by examining the lifetime of URLs and find that in contrast to early studies, the set of URLs visited is highly dynamic and well-modeled by a gamma distribution. Next, we analyze URL-traversal patterns and find that browsing behaviors differ substantially from hyperlink connectivity. One consequence of this is that the structure of the Web that is derived from hyperlink connectivity does not extend directly to actual user behavior. Finally, we consider the commonly used path and query portions of URLs and highlight their characteristics when used by different website genres. These semantic differences suggest that URL structure can broadly classify the kind of resource that a URL references. Our analyses lead to a set of proposed enhancements to the URL standard that would improve Web manageability and transparency and make a step toward the semantic web.**

I. Introduction

Over the past two and a half decades, the World Wide Web has been the dominant application for transmission of rich media on the Internet. Web-based applications such as search, e-commerce, social networking, and entertainment are used by billions of people around the world on a daily basis. However, the elegance and simplicity of essential aspects of the Web, namely clients running browsers connecting to servers delivering content via HTTP, belie its vast infrastructure and complex mechanisms.

Early studies of the Web focused on issues such as hyperlink structure [1], protocol performance [2], and user behavior [3]. These studies provided a basic framework for understanding the Web and for improving features, performance and reliability. Significant changes have taken place since these studies were conducted, including increased complexity of websites and browsers, the proliferation of mobile devices, content delivery networks, cloud-hosted applications, high quality search engines, online advertising, and social media. The rise and continuing evolution of these technologies coupled with the demands imposed by their widespread use strongly suggest the need for on-going empirical study of Web structure and use.

Unfortunately, there are significant challenges to assembling a data set that is sufficient for broad and deep analysis and modeling of the Web. First, there is a vast number of entities such as users, publishers, and infrastructure providers that comprise Web participants. Collecting data from any of these participants at scale involves a significant effort related to instrumentation deployment and management. Second, there is high sensitivity to privacy and new legal requirements such as GDPR and CCPA ([4], [5]) that preclude the use of certain kinds of data in research studies. These laws impose strict penalties for violations. Third is the analytic and modeling challenge of making sense out of diverse, highly complex data and extracting meaningful conclusions and parsimonious models that provide new insights into the evolving characteristics of the Web.

We posit that Uniform Resource Locators (URLs) [6] have unique utility for studying the Web. This was certainly true during the initial years of the Web's evolution, which was a time when hyperlinks, and by extension the URL, served as the primary tool by which users navigated the Web. As a result of this primary use case, user behavior and the Web's link structure were intimately connected. One guiding question of the present study is whether or not this still holds true for the modern Web, which has undergone a significant transformation towards increased automation and personalization. URLs are one of the key features in HTTP requests and are used in a broad range of applications, including personalized content, advertisement delivery, content delivery networks, and more. In addition to base pages, URLs are used to transmit myriad types of information including a user's browsing history, geographical location, and demographic information. Thus the URL serves two purposes: resource location *and* message delivery.

In this paper, we present findings of an empirical study of over 1 trillion URLs requested during Web browsing sessions from Comscore's 2M person user panel collected over a period of 12 months. The Comscore user panel is an opt-in data collection infrastructure[1] that tracks a variety of features during Web browsing sessions. The goal of our work is to present results that both relate to prior studies and convey new characteristics of Web structure and dynamics. While our data set is compelling in terms of scale and rich information content, we grappled with how best to achieve these goals in a way that is novel and provides useful insights. To that end, we focus on three different analyses.

Our first analysis considers the issue of *URL lifetime* from

[1]Comscore is highly sensitive to user privacy issues and follows all industry best practices for disclosure and data handling. See [7] for details.

978-3-903176-17-1 / © 2019 IFIP

a novel point of view, which to the best of our knowledge provides a new characterization. We introduce the notion of the *invoked lifetime* of a URL, which measures the time between a URL's first and last occurrence in a data set. Practically speaking, this relates to "content interest" dynamics in the Web. This concept is relevant to a publisher's editorial decisions and it can inform Web caching operations. We find that the distribution of invoked lifetimes is well-approximated by a parsimonious parametric function closely related to the probability density function of the gamma distribution. Using this model, we make the surprising observation that the expected lifetime of a URL is asymptotically equal to the time window of observation. In light of prior work [8],[9] that describe self-similar features of network and Web traffic and the ubiquity of fat-tailed distributions within network data, it is also surprising that a classical and "well-behaved" family of statistical distributions appears at all. This is also interesting since, within the realm of advertising (*e.g.*, bid requests and advertisements), a significant volume of URLs in Web traffic are nonces.

Our second analysis is focused on the issue of Web structure. In their work studying the structure of the Web, Boldi and Vigna demonstrate that a directed Web Graph constructed using intra-page links has a sparse representation [10]. We revisit this work, but from the perspective of user browsing behavior rather than page links. Our ansatz (and simplified) assumption is that a user's browsing behavior is a connected path within the Web Graph. If this is true, then a user's appropriately encoded browsing behavior should possess key characteristics that Boldi and Vigna first observed within their Web Graph. Empirically, we find that this is not the case. Our analysis brings rigor to the common understanding that users no longer traverse the Web as they did when these earlier studies were performed. It supports the notion that much of today's "Web surfing" is facilitated by search, recommendations, and so on.

Our third analysis considers the structure inherent in URLs by comparing the use of path hierarchy against query string (*i.e.,* before and after the "?" in URLs) on a variety of sites. We find that static content is delivered using path-depth structure while query parameters contain customized content. This path-depth-versus-query-parameter dichotomy exists even though the two morphologies are functionally equivalent. This is consistent with the following observation: in applications where search and resource discovery are either unnecessary (*e.g.* advertising) or managed through some other method (*e.g.* personalized recommendation systems on video sites) the URL serves as a unique identifier and it does not need to reflect the resource's content. The heavy use of query strings in many applications is in stark contrast with historical discussions that envisioned the URL much as one views a node living within a hierarchical file system.

We conclude by highlighting two issues that relate to the URL but are not addressed by the current URL standard. They are privacy and lexical scope.

URLs were designed to be public documents, and the open nature of the URL leads to the informed recommendation [11] that the URL should not contain sensitive information. In spite of this, we find that many URLs hosted on prominent web sites include private, sensitive personal information in clear text.

The URL standard specifies how a user's credentials can be incorporated into the URL. However, Chrome, which is currently the dominant Web browser [12] no longer supports the credential standard. This change is unlikely to cause widespread disruption because its use within Comscore panel web traffic is rare: credentials are managed using alternative methods. Regardless, sensitive information extends far beyond usernames and passwords. We report that URLs embed personal and sensitive information such as financial information, home addresses, student names, medical search terms and the number of children in one's family. Additionally, large corpuses of URL data that contains this information are routinely bought and sold through business-to-business transactions that lack transparency.

Human-readable URLs are still viewed as critical to the modern Web [13]. Widespread adoption of link shorteners [14] and efforts that alter the end-user's relationship with the URL (*e.g.*, the AMP project [15]) highlight the need to continually revisit basic assumptions about the URL, the role that the URL plays in the Web and to understand unanticipated emerging use-cases that rely on the URL.

II. Data and Methodology

We obtain our data from the Comscore user panel[2], whose participation is voluntary and requires informed consent. The panel is comprised of over 2 million desktop users who install software on their computer that monitors the HTTP(S) traffic that occurs between the panelist's machine and the outside world. The purpose of the panel is ultimately to inform publishers and advertisers about the composition of their digital audiences. Participants sign up in exchange for benefits including cash awards, antivirus software, and online credits. Data collection, storage, and analysis are performed in accordance with Comscore's privacy policy [7]. All data collected is analyzed *ex post-facto* to remove entries that are corrupted or associated with invalid or malicious activities. All analyses presented herein are based on coarse, high-level aggregations of data.

Panel records include, among other things, HTTP(S) headers (which contain URLs), response codes, and the names of processes making the requests. Many of these features are also available in standard network packet traces, and indeed, many of the analyses we present below could be accomplished using network packet traces as the data source.

To review, the URL is a hierarchical sequence of components defined in [6] as scheme, authority, path, query, and fragment. The authority component consists of user credentials, hostname and port number. In Web traffic, credentials are typically omitted and the port number is determined by the scheme,

[2]Similar panels are maintained by other commercial entities such as Nielsen, Hitwise, Compete, *etc*.

where port 80 is commonly used for HTTP traffic and port 443 for HTTPS traffic. The host, pathname and query fields are separated by the first occurrences of "//","/" and "?", respectively. In our analyses, we tokenize the URL using standard libraries such as Python's `urllib`. The query parameters of a URL often include inscrutable, nondeterministic values (*e.g.*, cache breakers) whose purpose is to increase the uniqueness of the URL. In contrast, the portion of a URL that contains the domain and path are indicators of more persistent resources.

Our data span May 2017 and April 2018, inclusive. The panel typically reports several billion records per day, and the total number of records used in our analysis is just over 1 trillion. The primary computing platform used for our analysis is a several-hundred node Spark cluster.

III. Results

A large, rich, longitudinal URL data set offers the opportunity to develop many insights on the Web. Our selection of analyses for this paper was motivated by the goal of providing insights on Web structure and user behavior that revisit prior work and offer new perspectives.

A. A Robust Parsimonious Lifetime Model

In this section, we show empirically that a simple parametric function that is closely related to the gamma distribution serves as a good model of URL lifetime. We consider URL lifetime from the perspective of client side requests: a URL is considered "alive" if it is in use by panel participants. This approach stands in contrast to prior work that defines lifetime in terms of server response codes [16], [17]. To this end, we define the *invoked lifetime* of a URL as the time between a URL's first appearance and its last. Our definition of lifetime differs from prior ones in that our focus is on observed behavioral interest in a URL, as indicated by an observed HTTP request. This is distinct from prior work focused on whether a URL successfully resolves.

For the present analysis, we restrict our attention to the domain and path portions of a URL and disregard query parameters. Additionally, while our data source includes all HTTP web requests that are associated with a browser on a single machine, we will only consider here those requests generated by a user who manually types a URL in the browser location bar and the subsequent requests to resources that load as a result of the top-level request.

Let $\mathcal{U}_N$ denote URLs observed between May 1, 2017 and May 1, 2017 + N days. If $u \in \mathcal{U}_N$ is a URL that is initially observed at time t_0 and is last observed at t_1, then the invoked lifetime of u is $t_1 - t_0$.

Figure 1 shows (blue) the number of URLs in $\mathcal{U}_N$ for $N = 61, 123, 184, 245$, and 304 whose lifetime was l, where $0 < l < N$. The function (orange) is described by

$$t \mapsto 2.5 \times 10^6 \, (N - t) \, e^{-0.025(N-t)}$$

is also shown. The parameters of this function were found using an ad hoc binary search strategy. It is noteworthy that the parameters of the model represented in Figure 1 are *held fixed* for the date ranges shown. More generally, for $t > 0$, define

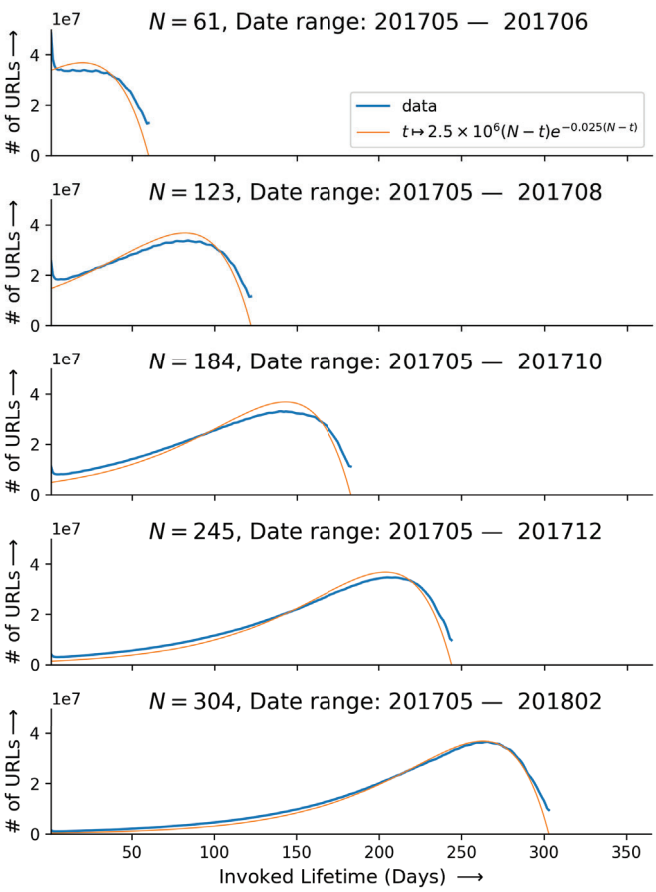

Fig. 1. The distribution of invoked lifetimes of URLs observed over N days (blue). The model function g is also shown (orange). The error of the fit in the bottom axis is $\|d - g\| \,/\, \|d\| \approx 0.08$.

$$g(t) \coloneqq g(t; N, c, \alpha, \tau) \coloneqq c \, (N - t)^{\alpha} \, e^{-\tau(N-t)} I_{[0,N]}(t)$$

The function g is closely related to the pdf of the gamma distribution with decay rate τ and shape parameter α. Unlike many visualizations of web measurements, the lines that represent the data in Figure 1 are remarkably smooth. We believe smoothness is not an artifact of either measurement or our processing, but a genuine feature of the data.

Figure 2 shows URL lifetimes for fixed N, but the URLs in the analysis had volume larger than $s = 0, 10, 10^4$ and 10^5. The parameters of the fitted $g(\cdot; N, c, \alpha, \tau)$ are listed in Table I.

A parameterized model allows additional results about the observed data to be derived. Consider $h(t) \coloneqq h(t; c, \alpha, \tau) \coloneqq ct^{\alpha} e^{-\tau t} I_{(0,\infty)}$, where c normalizes h so that $\int h = 1$. Then h is the pdf of a gamma distribution. If X is a random variable with pdf h having the same decay rate and shape as g of Figure 1, $E(X) = 80$. Applying this to that g, the mean invoked lifetime of a URL in $\mathcal{U}_N$ is, for large N, about $N - 80$. Asymptotically, this equals the length of observation!

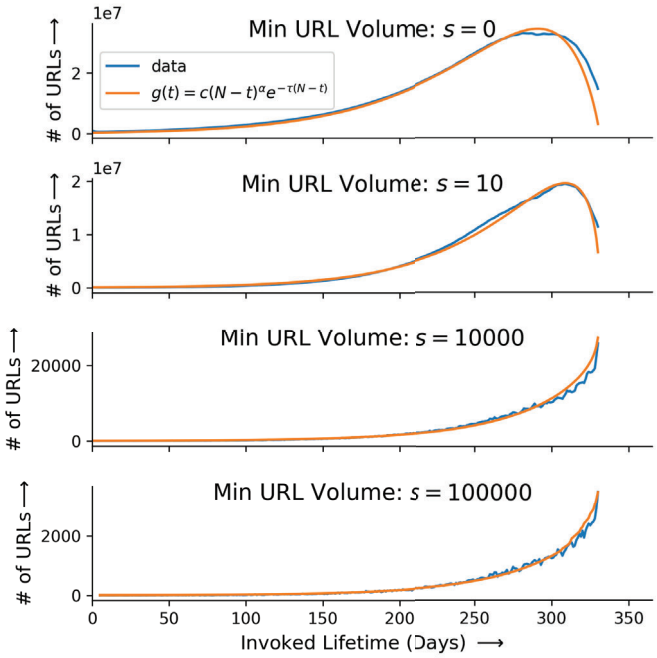

Fig. 2. Same as Figure 1, except that $N = 335$ is fixed and URLs were observed more than $s = 0, 10, 10000, 100000$ times. Model parameters appear in Table I.

TABLE I
DETAILS ABOUT g SHOWN IN FIGURE 2. THE TRANSITION FROM $s = 10$ TO $s = 10,000$ IS SMOOTH SO INTERMEDIATE VALUES ARE OMITTED.

s	N	c	α	τ	$\|d-g\| / \|d\|$
-1	*	2.50×10^6	1.00	0.025	0.080
0	335	3.24×10^6	0.88	0.02	0.095
10	335	6.70×10^6	0.51	0.02	0.059
10,000	335	2.80×10^4	-0.10	0.02	0.123
100,000	335	4.50×10^3	-0.10	0.02	0.111

The gamma distribution is the maximum entropy probability distribution subject to the constraint that $E(X)$ and $E(\log X)$ are held fixed. Thus, g may be completely specified using only these two quantities. We conjecture that the fitted gamma distribution described above may be derived using a first-principles analysis that leverages either this fact or else draws from facts known about gamma processes. Such a derivation would have explanatory power. The general approach envisioned is analogous to analyses that relate the Poisson arrival process to queuing theory and applies the analysis to explain empirically-observed distributions of queue wait times. We leave this sort of first-principles analysis for future work.

The URLs that have both large volume and long invoked lifetime are the most significant. For the popular-and-long-lived URL population, how is the volume of traffic distributed over time? This question is addressed by Figure 3. This figure is based on URLs that were requested at least 100,000 times during our year of observation. Each axis is associated with a label, $p\%$ and it displays the fraction of URLs that required at least d days to reach $p\%$ of their total volume. To illustrate this concretely, consider the axis $p = 75\%$. On this axis, it is shown that a significant fraction of popular-and-long-lived URLs require about 250 days to reach 75% of their total annual volume. This low-flat-distribution of volume is interesting in light of "flash-crowd" events, where the bulk of traffic occurs during the initial days of the URL's appearance, and trails off very quickly.

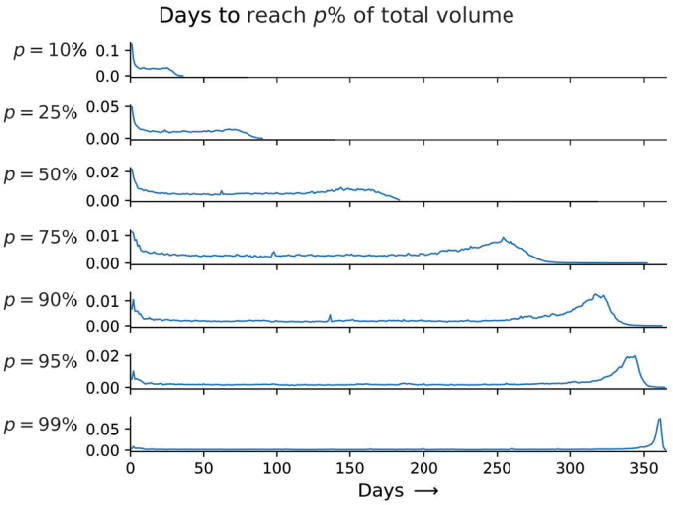

Fig. 3. Within each axis, the line height displays the fraction of traffic that required at least d days to reach $p\%$ of total volume. The area under each line sums to 1.

This first look at invoked lifetime, which is well-defined on longitudinal snapshots of URL data, demonstrates that it exhibits straightforward distributional characteristics. The invoked lifetime of a URL is relevant for infrastructure management and caching. URL lifetimes also serve as a proxy measure of URL churn, where shorter lifetimes indicate greater dynamism and complexity. Finally, a slightly more accurate model would include $t = 0$ in the domain. Such functions have the form $t \mapsto d_N \delta_0(t) + g(t)$, where d_N is the volume of URLs with 0-day invoked lifetime and δ_0 is supported at 0.

B. Revisiting the Sparse Web Graph

In this analysis, we address the question: how relevant is the Web Graph to users' browsing behavior in general? This was addressed in 2004 by Boldi and Vigna (BV) in [10] who showed empirically that the Web Graph, which is derived from page link structure, has a sparse representation. The efficiency of the BV representation is based on the observation that URLs on pages that link to each other are often lexicographically close. Our hypothesis is that, since that time, user habits, technology and the Web itself have change dramatically, and as a result, the characteristics that were originally observed have also changed.

To test this, we revisit the original analysis, but from the vantage point of a user's browsing behavior rather than page link structure. The early Web was dominated by static content, and the early model of navigation was consequently that users traversed the Web via hyperlinks. More formally, a user's

browsing activity could be represented as a connected path within the larger Web Graph. However, search, recommendations, bookmarks, URL shorteners, ephemeral URLs, social media, *etc.*, have changed how people navigate the Web.

Before describing our analysis, we sketch the key results of BV. The Web Graph is a directed graph where nodes are URLs and edges are defined by hyperlinks on pages. If $u, v \in \mathcal{U}$ are URLs (*i.e.*, nodes in the graph) and v appears on the page that u represents, then there is a directed edge in the graph, $u \to v$. In the BV analysis, each URL node in $\mathcal{U}$ is mapped to an integer according to its placement in the lexicographically ordered list of URLs in $\mathcal{U}$. Let $\#u$ denote this integer. A key finding in BV find is the "gaps" between hyperlinks follow a power law. More specifically,

$$|\{\#u - \#v = j : u, v \in \mathcal{U}, u \to v\}| = O\left(|j|^{-1.21}\right).$$

The consequence of this is that the Web Graph, despite having millions of nodes and billions of edges, has a sparse representation and it can be efficiently represented.

We now describe our analysis. As before, we order all URLs in $\mathcal{U}$ lexicographically, and let $\#u$ denote the index of $u \in \mathcal{U}$ in this list. Let i index a user, let $\left(u_0^i, u_1^i, u_2^i, \ldots\right)$ denote the time-ordered set of URLs in $\mathcal{U}$ that user i visited, and let the gaps in a user's browsing trace be the sequence, $\left(\#u_k^i - \#u_{k+1}^i : k \geq 0\right)$. Using BV's result that pages that link to each other are close lexicographically, if users navigate primarily by following hyperlinks, then we expect the distribution of gaps in user browsing behavior, in the aggregate, to be distributed according to a power law. However, if users generally navigate using some other method, for example with search or via links shared on social media, then we should not expect to see such a distribution.

Figure 4 shows the scatterplot of URL gaps appearing in the initial 5000 URLs visited by a random sample of 100,000 distinct panelists over one day. To compare this against the results of BV, also shown is the power law model that BV describe. Figure 5 displays the same analysis, except the domains are restricted to the top 500 domains reported by Moz [18]. To summarize, neither representation suggests that browsing behavior exhibits a power-law distribution. A linear estimator provides a better fit, but the fit is quite loose and the efficiency that was achieved by Boldi and Vigna is unlikely to be achieved here. Consequently, the early mental model of a user traversing the Web Graph by following a connected path does not apply. One might also expect that the widespread deployment of standard website templates (*e.g.*, WordPress) and development frameworks might also lead to small lexicographic gaps in browsing behavior. This does not appear to be the case.

This result helps to form a new basis for understanding a typical Web user's browsing behavior. The canonical model of a Web user, *viz.* the "Random Surfer," concerns the actions of a hypothetical user who traverses the Web primarily by following hyperlinks between pages [19]. However, our analysis demonstrates that this is no longer an accurate representation of modern browsing. This shift in behavior also highlights the power that search engines and social media platforms have to influence the content that modern users are exposed to, and by extension, what information users are most likely going to access.

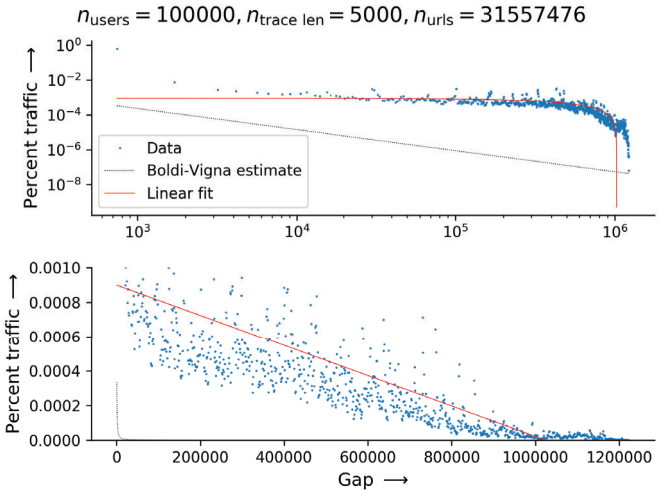

Fig. 4. The analysis of URL gaps within web browser traffic with log-log (top) and linear-linear (bottom) scales. The BV power-law model (black) and a least square linear fit (red) are indicated. The latter has slope and intercept -8.7×10^{-10} and 9.0×10^{-4}, respectively.

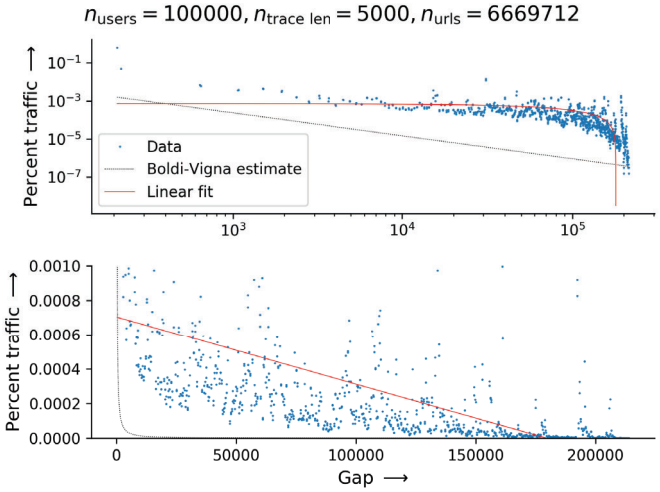

Fig. 5. The same analysis as appears in Figure 4, except that domains appear in the Moz top 500 list. The slope and intercept are -3.9×10^{-9} and 7.0×10^{-4}, respectively.

C. On Query Length and Path Depth

A typical web page renders in a browser via dozens, often hundreds of individual HTTP(S) requests that are issued to many different web servers, and most of this activity occurs out of sight of the user. Therefore, the record of a user's Web browsing session, whether assembled from packet traces or through in-browser methods will include this unseen activity. Accurate labeling of a collection of such HTTP requests into the "foreground" or the "background" categories is essential to an accurate reporting on publisher page views, for example.

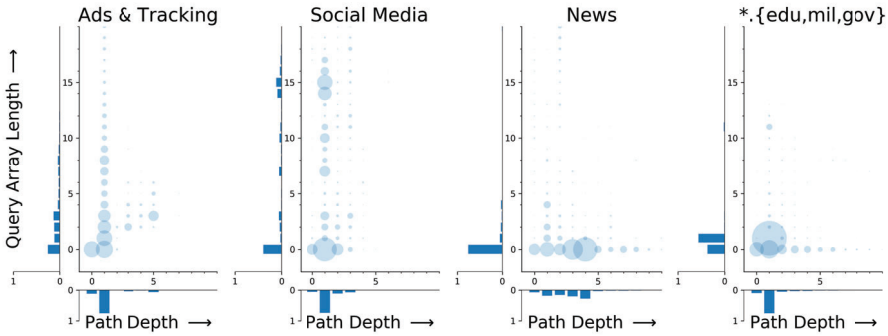

Fig. 6. The distribution of path-length versus query-array-length across site class. Within a class, let d_{pq} denote the fraction of URLs with path-depth p and query-array-length q. The radius of each disc in the scatterplot is proportional to d_{pq}. The marginal distributions, namely $\sum_q d_{pq}$ and $\sum_p d_{pq}$ are represented as histograms.

An heuristic approach to this sort of categorization may rely on curated keyword lists and lists of well-known domains that are associated with particular roles. It is generally, albeit informally, recognized that the use of path hierarchy and query fields within the URL depend upon the nature of the resource requested. In this section, our aim is to add rigor and quantify the conventional wisdom about how URLs are structured to serve different purposes. Specifically, we report that different classes of Web sites structure their URLs on the path-depth-versus-query-length plane in distinct ways.

The URL-to-resource relationship is that of a key-value pair: the characters within a URL are irrelevant as long as the URL properly serves as an identifier of its corresponding resource. After the domain name, the main features of the URL are its path and query fields. Web sites typically use the path portion to organize resources into a hierarchy, much as filesystems do. In contrast, the query parameters form an unordered list of key-value pairs.

One risk associated with a deep path hierarchy is that the labels within the hierarchy grow irrelevant or become a burden to maintain over time. This was recognized early in the development of the Web [20]. In contrast, a site that structures its URLs exclusively using query parameters has an extensible and flexible structure to work within while maintaining a manageable hierarchy. Site developers are free to structure their URLs, and thereby organize the resources that live on their sites, in any manner that they wish.

Figure 6 concerns the URLs associated with four categories of web site: social media, advertising, news, and URLs whose domain matches `*.{edu,mil,gov}`. The domains used to represent each category were manually selected from a handful of the most visited sites in the category. Each scatterplot displays path depth against the number of query fields of the URLs that were observed in panel data for each category. The sites with relatively static content (*e.g.*, `*.{edu,mil,gov}`) exhibit a preference for path-centric URLs. Conversely, services that customize content for users such as social media and advertising exhibit a preference for query parameters. The mean number of query field parameters and path depths are displayed in Table II.

TABLE II
MEAN VALUES OF PATH-DEPTH AND QUERY-LENGTH FOR EACH CLASS OF WEB SERVICE.

	Ads	Social	News	`*.{edu,mil,gov}`
$\mu_{\text{path depth}}$	1.3	1.2	2.9	1.3
$\mu_{\text{query length}}$	7.8	6.5	3.0	1.5

Query strings, through the inclusion of ad auction bid prices, referrer strings, and other information, provide insight into Web infrastructure. One issue that arises in especially complex strings is the inclusion of URLs as values of query parameters in other URLs. When embedding like this occurs, accurately parsing a URL becomes more complicated, or even impossible.

To conclude this section, our analysis shows that it is possible to make nontrivial inferences about the nature of an HTTP request by relying on very coarse metrics, *viz.*, path depth and query length. In other words, the form of a URL reflects the use and purpose of the URL.

IV. Privacy and Lexical Scope

We now discuss two issues that concern URL use within the modern Web: privacy and lexical scope. We argue that neither is effectively addressed by the URL standard [6].

The modern Web does not generally leverage the URL standard's proposed method to pass credentials from client to server. One reflection of this is that the dominant web browser (*i.e.*, Chrome) no longer implements this part of the standard [12]. Indeed, credentials are usually passed from client to server in a different manner and in a way that does not utilize the URL at all. Several reasons not to include sensitive information in the URL are that URLs are stored in browser history, they get stored in web server logs, and URLs are routinely passed around as values of query parameters in other URLs. The historical context for this is that URLs are envisioned as public documents.

Despite this, we find that personal and sensitive information are routinely included in URLs. To support this statement, we

report on a very simple set of *ad-hoc* analyses performed on a 1-day snapshot of desktop browser request data. To illustrate the flavor of the analyses performed, one analysis searched the corpus for literal strings such as "username=", "password=" and a handful of similar variants. Multiple instances of well-known brands pass credential login information using clear text that is embedded in URL query strings. We list the nature of other sensitive information that was found in this snapshot:

- the number of children in one's family (hotel rental)
- financial account status and credentials (financial site)
- location information in the form of latitude and longitude
- referrer information with medical search terms (ad request)

The diverse nature of information suggests that an automated detection methodology is unlikely to be comprehensive. It can be very challenging to prevent the leakage of sensitive information, and the cost in time and effort to reengineer a solution can be very high [21]. As already mentioned, large bodies of URL data are routinely bought and sold in a very opaque manner. Based on our experience within the industry, we report that data vendors do attempt to limit the sharing of personally identifiable information (PII) but in practice, these efforts are incomplete and as a result, this sort of information can be shared widely. Effective removal of sensitive information from the URL has no broadly-accepted solution, and no established norms exist to signal compliance with current (*e.g.*, GDPR and CCPA) or proposed policies (*e.g.,* [22]). In contrast to the failed DNT flag [23], which is now unsupported by at least one major browser [24], the emerging legal environment incentivizes market participants to actively embrace solutions that restrict the transmission of private information.

The second issue, namely lexical scope, arises in the following common scenario: if a URL has been properly "%"-encoded (per the standard), it becomes admissible to include as a value of another URL's query field. When this happens several times, the provenance of information can be impossible to trace accurately. As demostrated in the above analyses, long and complex URLs are now quite common, and many of these URLs contain URLs.

To address both these issues, we now state two proposals:

1) **Privacy** A syntax to label sensitive information within the URL. Examples of such information include IP address, geolocation coordinates, unique identifiers, human names, referrer, *etc*. Explicit privacy labels would provide visibility for industry participants and users on what sensitive information is being included in URLs. This is important because many users on the Web unknowingly and freely leak sensitive personal information that has monetary value to third parties. Such information is passed around by third-party data providers who have little to no visibility to the end user and face low risk of a consequence for sharing or monetizing this type of information. Included in this class of data is historical content consumption, search terms on medical conditions, and interpersonal relationships.
   If information were tagged within URL as "private" in a standardized manner, it would allow web servers to automatically scrub this information, it would allow browsers to quarantine this information, it would provide a trivial technical means for data providers in the marketplace to meet contractual guarantees and meet public normative expectations that information is not carelessly passed around, and it could provide non-technical policymakers a meaningful handle that can be used to discuss the management and transmission of private information.
2) **Lexical Scope** A syntax that formalizes initial and terminal tokens for query field parameters. This addresses problems that arise when a URL is passed as query parameter within another URL. In practice, URL encoding is inconsistently applied, and as a result it can be impossible to know which namespace "owns" a particular query parameter. A well-defined lexical scope would resolve the problem of assigning provenance to query parameters. In panelist data, we observe that as URLs are passed along from one entity to the next via the URL, the flat structure of the query parameters grows complex. Inevitably, every party that touches a URL and passes it along to another party appends their own metadata to the URL. This complicates the task of untangling the sequence of events that built up a compound URL.

V. Related Work

The structure of the Web Graph, the manner in which users traverse the graph and how information resides within the graph have each received a great deal of attention. An early investigation into the Web Graph's structure finds that the distribution of node degrees follows a power-law [25]. The authors of that work use URL lifetime as one component of their first principles analysis. In [26] the authors report that, although the graph's node degree distribution has a heavy tail, it may not be Zipfian.

A misalignment between hyperlink structure and user behavior was noted as early as 1997 [27]. More current research [28] reports link structure may no longer be useful for understanding Web use. The results in Section III-B strengthen this argument.

Other early efforts to understand Web browsing behavior include [29]. This work was quickly applied to problems such as Web caching [30], protocol analysis [31] and server analysis [32]. Demographic traits such as gender, income and ethnicity have also been associated with browsing patterns [33]. Recent studies have focused on tracking and privacy issues [34].

In [35], [36], hyperlink structure is used to infer the authority of information on the Web. Information provenance and tracking misinformation across the Web is still being investigated [37]. Early studies of the Graph informed our understanding of how hyperlinks connect information. They also influenced crawling strategies and PageRank for Google [38], [19].

In addition to the DNT flag, which was already discussed, other (failed) initiatives to protect or restrict the general user's privacy include a Google-created plugin for Chrome that allows a user to opt out of Google Analytics [39] and a Google-created

plugin that allows a user to opt out of [40] interest-based ads. As of this writing, neither plugin has seen wide adoption and neither has been updated since 2014.

VI. Summary and Conclusions

In this paper, we report results of an initial analysis of over 1 trillion URLs requested by a 2 million person user panel over a period of 12 months. The goal of this work is to report findings that provide perspective on prior studies and reveal new characteristics of Web structure and dynamics. We analyze URL lifetime using a metric we call invoked lifetime and find that URLs in today's Web are dynamic and well-modeled by a gamma distribution. We also analyze URL-traversal patterns and find that there are distinct differences between browsing behaviors and link connectivity, which indicates significant differences from early studies of the Web. Next, we divide the URL into its path and the query components and examine their differences in each based on website genres. We conclude that URL structure can broadly classify the kind of resource that a URL references, which provides a measure of rigor to the conventional wisdom that in spite of functional equivalence, paths and query strings are applied differently across application domains. Finally, our examination of URLs leads to a series of suggested enhancements to the URL standard that could benefit the broader Web community.

References

[1] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The Web as a Graph: Measurements, Models, and Methods," in *Proceedings of the 5th Annual International Conference on Computing and Combinatorics (COCOON)*. Tokyo, Japan: Springer, July 1999.

[2] P. Barford and M. Crovella, "A Performance Evaluation of Hyper Text Transfer Protocols," in *Proceedings of ACM SIGMETRICS*. Atlanta, Georgia: ACM, May 1999, pp. 188–197.

[3] C. A. Cunha, A. Bestavros, and M. E. Crovella, "Characteristics of WWW client-based traces," Boston University Department of Computer Science, Tech. Rep. TR-95-010, Apr. 1995, revised July 18, 1995. [Online]. Available: http://www.cs.bu.edu/techreports/pdf/1995-010-www-client-traces.pdf

[4] General Data Protection Regulation, "General Data Protection Regulation," https://gdpr-info.eu, 2018.

[5] California Consumer Privacy Act (CCPA), "California Consumer Privacy Act (CCPA)," https://oag.ca.gov/privacy/ccpa, 2018.

[6] T. Berners-Lee, R. Fielding, and L. Masinter, "RFC 3986 Uniform Resource Identifiers (URI): Generic Syntax," United States, 2005.

[7] Comscore, "Privacy Policy," https://www.comscore.com/About-comScore/Privacy-Policy, 2018.

[8] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the Self-similar Nature of Ethernet Traffic," in *Conference Proceedings on Communications Architectures, Protocols and Applications*, ser. SIGCOMM '93. New York, NY, USA: ACM, 1993, pp. 183–193. [Online]. Available: http://doi.acm.org/10.1145/166237.166255

[9] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec 1997.

[10] P. Boldi and S. Vigna, "The Webgraph Framework I: Compression Techniques," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. New York, NY, USA: ACM, 2004, pp. 595–602. [Online]. Available: http://doi.acm.org/10.1145/988672.988752

[11] W3C, "Good Practices for Capability URLs," https://www.w3.org/TR/capability-urls/, 2014.

[12] Google, "Drop support for embedded credentials in subresource requests." https://www.chromestatus.com/feature/5669008342777856, 2017.

[13] E. Waite, "Google wants to kill the URL," https://www.wired.com/story/google-wants-to-kill-the-url/, 2018.

[14] Linkedin, "Sharing news on linkedin just got easier," https://blog.linkedin.com/2010/04/21/linkedin-sharing-news, 2010.

[15] Google, "The AMP project," https://www.ampproject.org, 2015.

[16] S. Lawrence, D. M. Pennock, G. W. Flake, R. Krovetz, F. M. Coetzee, E. Glover, F. A. Nielsen, A. Kruger, and C. L. Giles, "Persistence of Web references in scientific research," *Computer*, vol. 34, no. 2, 2001.

[17] D. Spinellis, "The Decay and Failures of Web References," *Commun. ACM*, vol. 46, no. 1, pp. 71–77, Jan. 2003. [Online]. Available: http://doi.acm.org/10.1145/602421.602422

[18] Moz, "The Moz Top 500," https://moz.com/top500/domains, 2018.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab, Tech. Rep., 1999.

[20] T. Berners-Lee, "Cool URIs don't change," URI Style, W3C 1998. [Online]. Available: https://www.w3.org/Provider/Style/URI.html

[21] F. Contact, "Never Put Secrets in URLs and Query Parameters," https://www.fullcontact.com/blog/never-put-secrets-urls-query-parameters/, 2016.

[22] J. Daniels, "California governor proposes 'new data dividend' that could call on Facebook and Google to pay users," https://www.cnbc.com/2019/02/12/california-gov-newsom-calls-for-new-data-dividend-for-consumers.html, 2019.

[23] W3C, "Tracking Preference Expression (DNT)," https://w3c.github.io/dnt/drafts/tracking-dnt.html, 2019.

[24] Macrumors, "Apple Removes Useless 'Do Not Track' Feature From Latest Beta Versions of Safari," https://www.macrumors.com/2019/02/06/apple-removes-safari-do-not-track-option/, February 2019.

[25] L. A. Adamic and B. A. Huberman, "Power-law distribution of the World Wide Web," *Science*, vol. 287, no. 5461, pp. 2115–2115, 2000.

[26] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer, "Graph structure in the Web revisited: a trick of the heavy tail," in *Proceedings of the 23rd international conference on World Wide Web*. ACM, 2014, pp. 427–432.

[27] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: Information and pattern discovery on the world wide web," in *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*. IEEE, 1997, pp. 558–567.

[28] H. Taneja, "Mapping an audience-centric World Wide Web: A departure from hyperlink analysis," *New Media & Society*, vol. 19, no. 9, pp. 1331–1348, 2017.

[29] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Transactions on networking*, vol. 5, no. 6, pp. 835–846, 1997.

[30] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1. IEEE, 1999, pp. 126–134.

[31] P. Barford and M. Crovella, "A performance evaluation of hyper text transfer protocols," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 27, no. 1. ACM, 1999, pp. 188–197.

[32] J. M. Almeida, V. Almeida, and D. J. Yates, "Measuring the behavior of a world wide web server," in *High Performance Networking VII*. Springer, 1997, pp. 57–72.

[33] S. Goel, J. M. Hofman, and M. I. Sirer, "Who Does What on the Web: A Large-Scale Study of Browsing Behavior," in *ICWSM*, 2012.

[34] A. Cahn, S. Alfeld, P. Barford, and S. Muthukrishnan, "An empirical study of Web cookies," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 891–901.

[35] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[36] J. Kleinberg and S. Lawrence, "The structure of the Web," *Science*, vol. 294, no. 5548, pp. 1849–1850, 2001.

[37] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Hoaxy: A platform for tracking online misinformation," in *Proceedings of the 25th international conference companion on world wide web*. International World Wide Web Conferences Steering Committee, 2016, pp. 745–750.

[38] R. Meusel, P. Mika, and R. Blanco, "Focused crawling for structured data," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 1039–1048.

[39] Google, "Google Analytics opt-out browser add-on," https://support.google.com/analytics/answer/181881?hl=en, 2014.

[40] ——, "IBA Opt-out (by Google)," https://chrome.google.com/webstore/detail/iba-opt-out-by-google/gbiekjoijknlhijdjbaadobpkdhmoebb?hl=en, 2013.