Can Web Pages Be Classified Using Anonymized TCP/IP Headers?

Sean Sanders and Jasleen Kaur
University of North Carolina at Chapel Hill
Email:{ssanders, jasleen}@cs.unc.edu

***Abstract*—Web page classification is useful in many domains—including ad targeting, traffic modeling, and intrusion detection. In this paper, we investigate whether learning-based techniques can be used to classify web pages based only on anonymized TCP/IP headers of traffic generated when a web page is visited. We do this in three steps. First, we select informative TCP/IP features for a given downloaded web page, and study which of these remain stable over time and are also consistent across client browser platforms. Second, we use the selected features to evaluate four different labeling schemes and learning-based classification methods for web page classification. Lastly, we empirically study the effectiveness of the classification methods for real-world applications.**

***Index Terms*—Traffic Classification, Web Page Measurement**

I. Introduction

Why Classify Web Pages? The World Wide Web is the most popular application on the Internet and HTTP accounts for 80% of Internet traffic [1], [2]. Studying what *types* of web pages are being downloaded by clients has tremendous utility in several domains—we give four specific examples below.

- *Profiling the content type of a web page:* The content of web pages can typically be classified/categorized into genres such as Finance, Shopping, News, Education, Automobiles, etc [3], [4]. Knowledge of the genre of web pages downloaded by a given user can be used for gauging user interest, which is invaluable for delivering personalized content and targeted advertisements [5]. For instance, service providers rely on deep packet inspection to assess what type of content consumers are interested in, for the purpose of delivering ads [6].
- *Profiling the usage of video streaming (application type):* Video streaming is now reported to occupy nearly 50% of network bandwidth, and consumption is expected to grow [7], [8]. The ability to distinguish between bandwidth-hungry video and non-video streams at critical traffic aggregation points, can help facilitate better planning and control. For instance, a campus network manager may be able to prevent network abuse and/or rate-limit video streams destined for student dorms; researchers may want to build profiles of enterprise video traffic to facilitate traffic modeling and forecasting studies; or perhaps Internet service providers may even want to limit resources per business interests [9].
- *Profiling the usage of mobile devices:* The average number of devices per Internet user is estimated to grow to 5 by 2017 [10]—most of these are mobile devices. The ability to identify web page downloads targeted for mobile devices can help in: (i) building profiles of mobile web usage within an enterprise (for capacity planning, modeling, and forecasting purposes), and (ii) delivering personalized content and advertisements that are customized for constrained displays, power, and connectivity.
- *Profiling the navigation-style for web browsing:* The way users navigate through web pages can be classified as accessing either a landing page (home page), clickable content (non landing pages), or with the increasing use of search and recommendation engines, search results. Such a navigation-based classification can be useful for identifying network misuse. For instance, web crawlers are misused for purposes of web page scraping [11]. Recent studies have shown that the pattern of web page navigation from a given end-point can help identify the corresponding malicious bots [12], [13].

In this paper, we ask the question: can web page downloads be classified along dimensions such as the above, using *only* anonymized TCP/IP headers that appear in the corresponding network traffic?

Why Consider Only Anonymized TCP/IP Headers? Entities that do not have direct access to client-side or server-side end-points—such as Internet service providers or enterprise network administrators—have to instead monitor access links and conduct deep-packet inspection of the *network traffic* generated by client-server interactions [14], [6], [15]. For instance, signature- and keyword-based approaches that scan the HTTP headers and payload, can be used for identifying each of the above web page types—video streams, navigation patterns, content types, as well as mobile-targeted web pages [14], [4].

Recent studies have shown, however, that nearly 86% of traffic today is obfuscated (either encrypted or compressed) [16], which makes deep packet inspection practically infeasible—this situation is likely to worsen with the advent of default encryption in HTTP 2.0 [17]. Furthermore, strong Internet privacy legislation dictates that Internet Service Providers and network managers may access no more than *anonymized*[1] TCP/IP headers in the traffic [18]. This is a sig-

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144081, as well as by NSF under Grant No. CNS-1018596 and OCI-1127413

[1]Anonymized headers refer to those in which the IP addresses of the clients are changed to sequences that can not be traced back to the originals.

nificant challenge because most current monitoring approaches rely on application-specific information.[2] It is, consequently, imperative to understand: can anonymized TCP/IP headers be *mined* to gather information for the purpose of classifying the corresponding web page downloads? In this paper, we conduct the *first* study that relies on *learning-based classification* to address the above question, in the context of the four labeling dimensions mentioned above.

Our Contributions. Prior work on traffic classification primarily focuses on application/protocol classification (e.g., FTP, HTTP, peer to peer, mail, etc) [19], [20], [21], [22]. However, classifying traffic based on the *type* of web pages has never been considered before—this is a challenging domain due to the tremendous diversity of web pages and browsers as well as the significant complexity of multi-flow web page downloads [23]. More fundamentally, it is not even clear if the *type* of a web page would influence the TCP/IP features at all. We study this issue by making the following key contributions:

1) *Data collection:* We use five different modern browser platforms to conduct and analyze downloads of 3345 web pages, all belonging to the top-250 most popular web sites. Overall, we analyze more than 100,000 web page downloads. For each download, we process TCP/IP data as well as collect the ground-truth about the *type* of the corresponding web page, based on the four classification schemes.
2) *Feature extraction and selection:* We process the TCP/IP headers to derive 216 features, including temporal and multi-flow features as well as their statistical derivatives. We then conduct a systematic analysis of these features to identify robust and discriminatory features—to the best of our knowledge, this is the first work that argues for, and explicitly considers *consistency* (across different browser platforms) and *stability* (over time) while selecting robust features.
3) *Web page classification:* Using the selected robust features, we then evaluate how effectively can these help classify web page downloads according to each of the four diverse labeling schemes. We find that while mobile-targeted and video downloads can be KNN-classified with more than 90% accuracy, the genre- and navigation-based categories can be classified with a somewhat lower accuracy.
4) *Applicability of classification:* We then evaluate the impact of our work on two application domains. The first is that of *traffic modeling*, in which we study the distributions of (i) traffic modeling parameters, as well as (ii) properties of the generated traffic—we find that these are *statistically indistinguishable* from distributions derived using ground-truth labels.
   The second application is that of building user browsing profiles—we find that the genre-preference of a random synthetic Internet user can be reconstructed with more than 80% accuracy, based on the classified labels.

In the rest of this paper, we present our data collection methodology in Section II; feature selection, classification, and applicability study in Sections III-V; related work in Section VI; and our conclusions in Section VII.

II. Data Collection Methodology

Learning-based classification requires correctly-labeled ground-truth data for *training* the classifiers. Below we describe our methodology for selecting, labeling, and downloading web pages, as well as extracting TCP/IP features.[3]

Which Web Pages? We focus on the top 250 web sites from [3]—studies suggest that nearly 99% of web traffic originates from just these 250 web sites. We browse each web site to collect a list of URLs for their landing pages, as well as non-landing pages, including search results and media content.[4] Overall, we include a list of 3345 web pages.

Ground-truth Labels We also assign labels to each web page (Table I), according to the four labeling schemes as follows. AGL: A *content-genre* based label is assigned to each web page, using the top-level Alexa genre for the corresponding web site (the 4 most common labels are listed in Table I). WNL: A *navigation-based* label is assigned based on whether the web page was the landing page, a search-result page (obtained by entering random keywords in a search box), or a clickable content page (including news articles, video content, and social networking pages).[5] VSL: The *video-streaming* (vs. non-video streaming) label is assigned to web pages where a video has played—this includes samples from top video streaming providers like netflix, youtube, and hulu. The non-video category also includes traffic sources that are fairly bandwidth-intensive, including radio sites (soundcloud and pandora), and file transfer sites with large files (dropbox and thepiratebay). TDL: The final set of labels correspond to *mobile-optimized* or traditional pages. We only include mobile web pages that also have a traditional web page that serves the same content—e.g., a superbowl article on bleacherreport.com that also appears on its mobile web site.

Trace Collection The TCP/IP trace generated by the download of a given web page may differ across client browser platforms [27]. In order to make our classification robust to the browser platform, we load each of the 3345 pages using 5 different modern browsers, and study the consistency of each TCP/IP feature across these. The 5 browsers—Internet

[2]Note that the content in anonymized TCP/IP headers directly reveals nothing more about the application, other than the fact that it is using HTTP! Even the HTTP headers themselves are unavailable!

[3]It is important to note that web page classification using TCP/IP traces will first require us to identify which set of TCP flows correspond to a given web page download—such *web page boundary detection* has received prominent interest in recent literature [24], [25], [26] and is beyond the scope of this paper. We ask: once web page boundaries have been detected, how effectively can the anonymized TCP/IP headers be used to classify the type of the web page?

[4]Our methodology does not capture the fact that some websites present different landing pages to users who are logged in (e.g., facebook.com)—study of such "personalized" web pages is left for future work.

[5]There may be several homepages per web site—e.g., www.yahoo.com and www.finance.yahoo.com. We classify each of these as landing pages.

TABLE I
DISTRIBUTION OF CLASS LABELS

Labeling Scheme	Class Names	# Web Pages
Video Streaming	Video page	169 (5.05%)
(2 Classes)	Non-Video page	3176 (94.95%)
Targeted Device	Traditional page	2481 (74.17%)
(2 Classes)	Mobile optimized page	864 (25.83%)
Alexa Genres	Computers	821 (24.54%)
(18 Classes)	Shopping	375 (11.21%)
	Business	363 (10.82%)
	News	320 (9.57%)
	Other 14 classes	1466 (43.86%)
Web page	Clickable content page	1505 (44.99%)
Navigation	Search result page	1226 (36.65%)
(3 Classes)	Landing page	614 (18.36%)

Explorer (IE) v 9.0.8112.16502, Firefox v 23.0.1, Google Chrome v 29.01547.66m, Safari v 5.1.7, and Opera v 12.16—are run on a modern Windows 7 desktop.[6]

Web pages are also updated over time [25]—in order to study which TCP/IP features remain stable over time for a given web page, we also repeat the above 3345×5 downloads 6 times each, over a period of 20 weeks (Mar 10 - July 24, 2014). Overall, this results in 100,350 web downloads. TCP/IP traces are *automatically* collected for each download as:

1) Start packet capture tool
2) Start a browser with a web page URL as an argument
3) Close the browser and packet capture tool after 120 seconds
4) Clear the local DNS resolver and browser cache
5) Go to Step 1 using a new URL

Quantitative Feature Extraction Access to the TCP/IP traffic traces allows us to extract many bidirectional traffic features—such as the number of PUSH flags or the size of HTTP objects transmitted in a TCP connection—that are not available from other sources such as netflow logs.[7] Our methodology also allows us to add *multi-flow* features that span the multiple TCP transfers characterizing a given page download—for instance, the number of TCP connections, number of distinct IP-pairs used, flow inter-arrival times, and total number of packets and bytes transmitted. We also include fine-scale temporal features such as round-trip time (RTT) and inter-epoch (inter-object) arrival times. We also include statistical derivatives—such as the minimum, maximum, and several percentiles—of the occurrence of a given feature. In total, we extract 216 quantitative features for processing (listed in [29]).

III. FEATURE SELECTION

The success of classification models relies critically on the selection of *informative*, *uncorrelated*, and *robust* features [30]. Prior traffic classification studies have focused on the first two properties by using automated correlation-based feature selection algorithms (e.g., [19])—robustness of features has not been considered though. Given the diversity and dynamism present in the Internet (and especially in the World Wide Web), this is a rather serious issue [31], [32], [33]! Specific to our goal, it is important to consider the impact of at least two factors:[8]

- Time: Modern web pages may change several times a day [31]. It is important to study how this impacts the stability over time of the TCP/IP features generated when the page is downloaded—indeed, classifiers that are trained on features that are stable over time are more likely to perform well on unseen data and do not need to be retrained often.
- Browsers: Client browser platforms differ in their configurations and may generate different TCP/IP features when downloading the same web page (e.g., depending on the extent to which they use pipelining). It is important to study which features are consistent (similar) across different browsers—else, classifiers trained on one browser will not perform well on unseen data that may have been generated by a different browser.

In order to incorporate the above aspects, we use a 3-step process for feature selection: (i) identify a set of most *informative* features for web page classification; (ii) group the most informative features into subsets of highly *correlated* features; and (iii) select the most *stable* (over time) and *consistent* (across browsers) features from each of the above subsets.

We elaborate on these steps below. In what follows, for each feature i, let $M^i_{n,b,t}$ represent an $N \times B \times T$ matrix populated with the measurements of feature i across the N (= 3345) web pages, B (= 5) browsers, and T (= 6) repeated web page downloads over time.

Identifying and Grouping Informative Features For selecting informative features, we first minimize noise due to browser selection or time of measurement by computing the *average* of the $B \times T$ measurements of a feature for a given web page. We then use the RELIEF method [30] to rank the 216 averaged features according to their ability to classify the 3345 web pages. We select the top 40 ($\sim$ top 20%) most informative features for *each* of the four labeling schemes—of the total 160 features, we find that only 63 are unique (many features were informative for multiple labeling schemes).

We then group the 63 features into correlated subsets. For this, we use the pearson correlation, ρ, to identify 10 groups of highly correlated features (listed in [29]). The features within each group have $\rho \geq 0.75$, whereas the correlation between features from different groups is typically less than 0.3.

Feature Stability Next, we quantify how stable these 63 features are over time. For each feature i, to control for the effect of different client browsers, we define an $N \times T$ matrix: $S^i_{n,t} = \sum_b w(b) M^i_{n,b,t}$, where $w(b) \in [0, 1]$ represents the usage fraction for browser b (obtained from [36]), which helps ensure that our analysis is representative of real-world web

[6]Apple does not support Safari on Windows. Thus, the version of Safari used for this study is outdated compared to the version used on OSX.

[7]We use the method from [28] to identify application data units in TCP/IP traces—these generally correspond to objects.

[8]In this work, we do not consider the impact of client location. Some recent studies show that location does not significantly impact basic web page features (for Firefox) [34], or that TCP/IP features that are informative for application protocol classification do not vary significantly across location [35]—we leave a comprehensive evaluation of the impact of this factor for web page classification for future work.

traces.

We then estimate the *stability* over time for each feature i and web page n by computing the average percent deviation metric defined as:

$$DS_n^i = 100 \cdot \frac{\sum_{t=1}^{T} \left| S_{n,t}^i - \mu_n^i \right|}{T \mu_n^i} \quad (1)$$

where, $\mu_n^i = \frac{\sum_t S_{n,t}^i}{T}$. For each feature i, we then extract the median, 10- and 90-percentile values of the deviation DS_n^i observed across the 3345 web pages—these values are plotted in Fig 1(a). The features are first grouped according to the 10 correlated subsets, and then sorted according to the median value of DS_n^i. [29] lists these features in the same order.

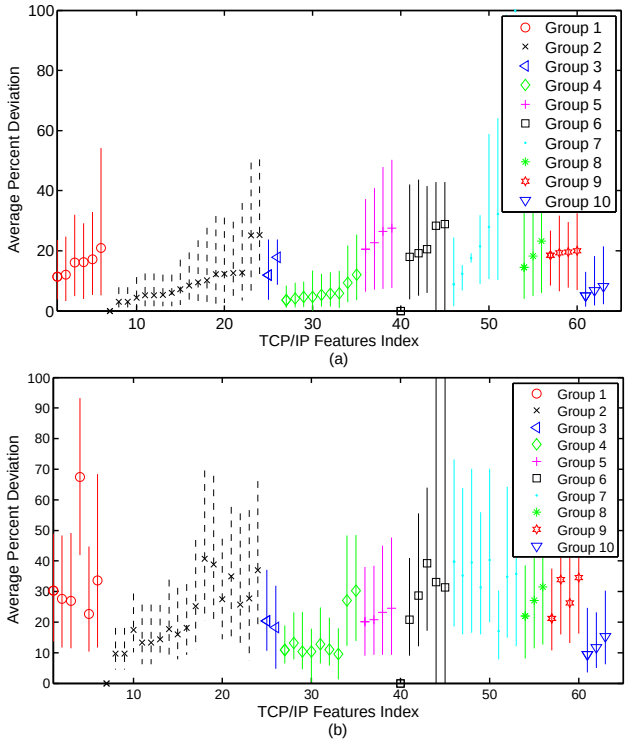

Fig. 1. Stability (a) and Consistency (b) of Features in Groups 1-10

Feature Consistency We use a similar formulation to estimate the *consistency* across browsers for each feature i and web page n by computing a corresponding average percent deviation as:

$$DC_n^i = 100 \cdot \frac{\sum_{b=1}^{B} w(b) \left| C_{n,b}^i - \nu_n^i \right|}{\nu_n^i} \quad (2)$$

where, $C_{n,b}^i = \frac{\sum_t M_{n,b,t}^i}{T}$ is an $N \times B$ matrix, each element of which represents the *average* measurement of feature i when browser b downloads web page n repeatedly; and $\nu_n^i = \sum_b w(b) C_{n,b}^i$. Fig 1(b) plots the median, 10- and 90-percentile values of DC_n^i observed across the 3345 web pages—the x-axis uses the same feature index as Fig 1(a).

Selection of Robust Features We use Fig 1 to select the most time agnostic and browser agnostic features from each of the 10 groups of correlated features [29]. By comparing these two plots we find that the median, 10- and 90-percentile deviations for nearly all features in the feature consistency plot are larger than the corresponding values in the feature stability plot. This clearly implies that the *TCP/IP features generated by the download of a given web page vary more across client browser platforms than over time*. A deeper analysis reveals that the traffic variation across browsers is primarily caused by ad and tracking services, which are browser specific [23].

Although most of our selected features vary significantly across browsers, some features vary much more across browsers than others. For example, feature 4, Number of Bidirectional Reset flags, is relatively stable over time like other features within its group (Group 1). However, this feature changes much more dramatically across browsers.

Based on the above, we select 10 features (one from each group) as robust and informative—these are discussed below.

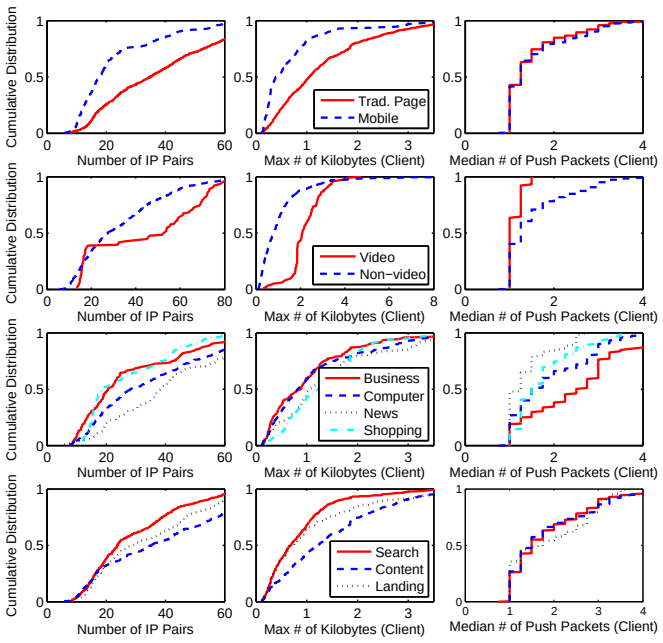

Fig. 2. Discriminatory Power of Some Non-temporal Features

A. What Features Are Informative?

Number of servers contacted The *total number of distinct servers contacted* for downloading a web page is discriminatory for several classes across the two labeling schemes (Fig 2). We find that mobile optimized web pages contact significantly less number of servers—this is presumably because they are designed for devices with constrained resources.

We also find that video pages contact more servers than non-video pages—these extra servers correspond to increased number of ads, images, and comment boxes. This is especially true for Youtube pages, which establish around 400 TCP connections (whereas Netflix uses 60 connections). We also find that search results generally display less content from multiple servers than do clickable content pages.

In the genre-based category, we find that News pages contact significantly more number of servers than other classes—this was also previously observed in [34], and is presumably because News sites tend to summarize on the same page different types of topics (sports, weather, finance, etc).

Number of PUSH flags per TCP connection The *maximum number of push packets sent in a TCP connection* for a given web page is also an informative feature for all labeling schemes. Previous studies show that the PUSH flag corresponds to a HTTP object[24]—our data also yields a high correlation between the two. The *maximum* number of objects sent in a TCP connection quantifies the prevalence of HTTP pipelining or connection reuse—we find that this is more popular for traditional (non-mobile) web pages and video pages as compared to their counterparts.

The *median number of push packets sent by the client per TCP connection* seems to be informative for the genre and navigation-based labels (Fig 2 confirms this). Clients that view landing pages use PUSH flags slightly more often than non-landing or search pages. This is presumably because landing pages are likely to collect many more objects summarizing the web site; these objects are co-located on a small number of servers—this may be done to help reduce the load time for the "entry page" of the web site by using persistent connections and by contacting less number of servers. We have also observed that the distinguishable categories using this feature (i.e., genre and content) also tend to include more javascript objects than other pages.

Total number of bytes transferred The number of *bidirectional packets* is a feature that roughly approximates the amount of data transferred to render a web page. As expected, this feature can be used to identify mobile and video pages. The *minimum number of bytes transmitted by the client per TCP connection* is informative for classifying mobile web pages—this makes sense because mobile sites are designed to be more efficient than traditional web pages.

Object size in largest TCP connection The object size distribution for the largest TCP connection (i.e., the TCP with the most bytes transmitted) is a valuable feature for discerning video traffic and is not useful for any other labeling scheme. It is common for video objects to be larger than 200 KB, which is rare for other types of traffic.

Temporal features Features in Group 8-10 are temporal features. We find that these are informative mostly for the video labels. The three selected video features include, the *75 percentile inter-connection arrival time*, *75 percentile inter-object arrival time*, and the *average RTT*. Two of these features, the inter-connection arrival time and inter-object arrival time, make sense for video classification because variable bit rate algorithms may request objects at irregular times to reduce transmitting video content that depends on a user's interest [37]. The average RTT is a surprising feature—it perhaps reflects that video servers take longer to respond to requests than other servers.

Are port numbers and first few packets helpful? Prior work on traffic classification (identifying the application layer protocol) has found port numbers and the sizes of the first few packets to be the most informative features [19]. Our analysis of web page classification, which also incorporates features that span multiple flows finds a completely different set of informative features—while port numbers are not even an applicable feature for web-only traffic, we find that even the first few packet sizes does *not* help distinguish between different types of web downloads. We believe that this is the case because the first few packets may capture *handshaking* mechanisms that are application protocol/application specific, but do not capture the differences between different *types* of web pages which are transmitted over the same application.

TABLE II
WEB PAGE CLASSIFICATION ACCURACY (KNN)

Classification Model	VSL	TDL	AGL	WNL
Stable Tcpdump features	99.1%	90.2%	73.0%	82.2%
Unstable Tcpdump features	98.3%	84.1%	62.3%	77.6%
Netflow features	98.4%	87.6%	67.2%	77.7%
Stable Tcpdump: different browsers	98.4%	84.4%	58.1%	72.5%
Stable Tcpdump: different time	99.1%	88.7%	70.4%	79.9%

IV. WEB PAGE CLASSIFICATION

Evaluation Methodology We use classification methods similar to the ones used widely for traffic classification [20]. This includes the non-parametric methods—that do not make assumptions about the distribution of features—of K-Nearest Neighbors (KNN) and Classification Trees (CT), as well as the parametric methods of Naive Bayes (NB) and Linear Discriminant Analysis (LDA). Our results show that non-parametric methods perform significantly better—we include results only for KNN here; more details can be found in [29].

To ensure that the dataset used in this section is consistent with prior knowledge of browser usage in real-world traces, we randomly sample our 100,350 captured web page downloads by browser (using weights from [36]). This data is then used to evaluate web page classification using the 4 independent labeling schemes (VSL, TDL, AGL, and WNL). We conduct 10 independent 5-fold cross validation trials (80% of the dataset is used for training and 20% is used for testing) and report the mean classification accuracy across these trials.

Classification Results Table II summarizes the mean classification accuracy of KNN, using our selected features for each labeling scheme. We find that the accuracy depends on the labeling scheme—web pages with streamed video can be identified with 99% accuracy, mobile-targeted pages with 90%, navigation-based labels with 82%, and genres with 73% accuracy. We believe these numbers are highly encouraging—while the numbers for video and mobile-targeted pages are expected to be high due to the presence of highly discriminatory features in these classes, surprisingly, we find that even the *content-genre* and the *navigation-type* of a web page can be inferred with relatively high accuracy using just information seen in TCP/IP headers of the corresponding traffic![9]

How important is feature stability? We next study whether the features selected in Section III actually outperform features that are less robust. Specifically, we compare classification accuracy when the most *unstable* (over time) features are selected from each of the 10 feature groups in Section III,

[9]Related performance metrics of recall and precision are included in [29].

instead of the most stable ones. Recall that all features in each group are fairly informative (for classification) and are highly correlated with each other. The results are summarized in Table II, which show that accuracy with unstable features can be up to 10% lower than with stable features. Thus, we conclude that it is important to include not just informative features for classification (as most prior work on traffic classification does), but to also consider the stability of a feature.

Would Netflow-derived features suffice? Our results above are obtained with classification performed based on fine-grained features derived from per-packet TCP/IP headers. Sometimes, access to such packet traces may be infeasible or costly. We next ask: what accuracy can be achieved if only coarse-grained features that are obtainable from Netflow logs, are used for classification? For this, we consider those (stable) features from each group that can be derived from Netflow logs. For instance, instead of the maximum number of PUSH packets sent by the client (Group 2), we include the maximum number of bytes sent by the client per TCP connection. None of the features in Group 6 and 7 qualify, though.

Table II shows that while video-streams can still be identified with high accuracy, Netflow-derived features yield lower classification accuracy by up to 8% for the other classes. It is important to note that the performance with even coarse-grained Netflow features is *better* than with unstable tcpdump features—this further underscores the importance of considering stability in selecting fine-grained features.

Sensitivity to Time and Browser Our dataset includes 6 repeated downloads of each web page, using 5 different browsers for each. While we have explicitly identified features that are the most robust across time and browsers, it is important to understand the impact of training on one portion of a dataset and testing on another. We first consider the impact of time on classification performance, controlling for browser. Table II shows that this hardly impacts classification performance at all. This result is promising, because it implies that classifiers do not have to be trained on data every day. In fact, our dataset includes measurements for over a period of nearly 20 weeks!

We next consider the impact of browser on classification performance, controlling for time. Table II shows that while video streams can still be identified with the same accuracy, the accuracy for the mobile-targeted and navigation-labels reduces by about 6-10%. The most significant impact, however, is on the genre-based labels, which can be classified only with 58% accuracy! These results imply that traffic classification performance is much more browser-dependent than time-dependent—our analysis of repeatability and consistency of traffic features in Section III supports this observation. We conclude that it is important to train models on data that is representative of browser mixes found in real-world traces.

Can classification be done with less packets? Some applications—such as rate-limiting or intrusion detection—benefit if a web page can be classified *while* it is being downloaded. Since most of our 10 selected features are multi-flow features, they may not be estimated accurately before a transfer has completed. In order to analyze how many packets need to be observed before web page classification can be performed accurately, we next analyze the first N (varied from 10 to 1000) TCP/IP packets in a traffic trace and derive our stable tcpdump features only from this limited information.

TABLE III
CLASSIFICATION ACCURACY (KNN) WITH FIRST N PACKETS

Classification Model	VSL	TDL	AGL	WNL
10 Packets	75%	75%	35%	50%
100 Packets	94%	79%	42%	60%
500 Packets	97%	83%	63%	67%
1000 Packets	99%	88%	65%	72%

Table III shows that the classification performance is proportional to the number of packets used, N. For our dataset, the average number of packets transmitted per web page is 2200. We find that while there is a performance hit for using only 25-50% of the packets downloaded for classification, classification accuracy is still high for video traffic and mobile-optimized web pages, and reasonable for navigation-type.

Risks of Our Classification Approach The success of our classification framework depends on the degree to which web pages of a particular category yield similar traffic features. Our evaluation above suggests that this does happen to a surprising extent on the current Internet. For some labeling schemes, the association with traffic is to be expected—e.g., mobile web sites are designed to be more resource concious than traditional ones; and video streaming transfers large volumes of data. For other labeling schemes, though, the association seems to result from similar design decisions by web site designers—e.g., modern search engines include similar search options such as web search, image search, and news search; and most News web sites have similar templates and layout. The risk with our approach is that the association of such categories with traffic features may change as web site designs evolve over time. In order for our framework to remain effective, it is fairly important to: (i) strategically sample web sites that are likely to be included in a real traffic trace (we focus on popular web sites in this paper), and (ii) periodically retrain our classifiers.

V. HOW APPLICABLE IS OUR WORK?

In Section I, we provide several motivating examples for web page classification using TCP/IP headers. Here, we focus on two of these and quantify the impact of our work.

A. Application: Building Traffic Models for Forecasting

We first consider the application of *forecasting traffic growth*, which can help with better capacity planning. Specifically, in this section we study the trend in the growth in mobile web usage. Our aim is to study whether traffic predictions based on this trend, look any different when they rely on classified labels instead of ground-truth labels.

Do feature distributions for classified and ground-truth labels match? We first study if our classification results are useful in extracting the *true* distributions of traffic features within a given class—such distributions can then be used for traffic modeling and simulation studies. To quantify this,

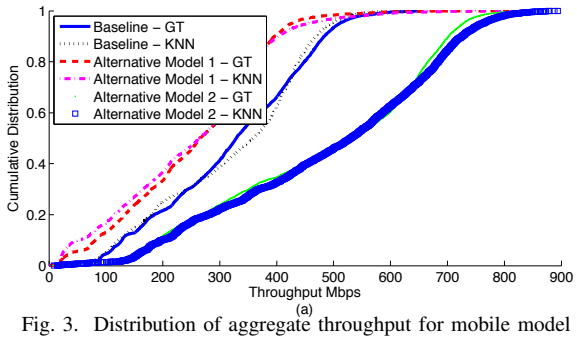

Fig. 3. Distribution of aggregate throughput for mobile model

we statistically compare the distributions of several features obtained from traffic with ground-truth labels to those resulting from traffic with classified labels.[10] We rely on two hypothesis testing approaches—the Wilcoxin sum ranked test and Kolmogorov-Smirnov test. We find that with the KNN classifier, both of these tests yield favorable p-values, larger than 0.05 for *all* classes for *each* feature tested—this is true even for the genre-based classes, that yielded a somewhat lower classification accuracy. We conclude that both the median value as well as the empirical distributions of these features are statistically the *same* across classes identified using either classified labels or ground truth labels. Detailed tabulations of the p-values can be found in [29].

Traffic Generation Methodology We next compare traffic generated by relying on ground-truth versus classified labels. We use the ns-2 network simulator to simulate the behavior of 400 active web users—all to/from traffic gets aggregated on a 1Gbps link. Each user behaves independently and randomly visits a web page. The inter-arrival time for web page downloads by a given user is gaussian distributed with a mean of 30 s and standard deviation of 15 s—this distribution is chosen for simplicity (and is adequate for our purpose of comparing the impact of classified and ground-truth labels).

The download of each web page itself is simulated using TMIX, which provides a source-level traffic generation interface in ns-2 [28]. Specifically, we provide this tool with the TCP/IP trace of a web page download (selected randomly from the 100,350 downloads we collect in Section II). TMIX then derives from the trace, application-level descriptors of the corresponding traffic sources—including request sizes, response sizes, user think times, and server processing times. It then generates a corresponding traffic in ns-2 by reproducing these source-level events. Thus, this tool allows us to faithfully produce realistic source-level behavior for each web page download. We use this traffic generation methodology in the context of the forecasting application below.

Modeling Growth in Mobile Web Usage We first construct a *baseline model*, in which each user visits a mobile-optimized web page 20% of the time and a traditional page 80% of the time—nearly 20% of current web traffic is considered mobile [38]. The TMIX input for each is obtained by randomly selecting a mobile (or traditional) page download from our set of 100,350 downloads—we conduct two experiments, in which the mobile or traditional pages are selected based on either ground-truth (GT) labels or KNN labels (ML). The throughput on the 1 Gbps aggregated link is observed every 1ms, and its distribution is plotted in Fig 3.

We next conduct two sets of experiments that incorporate growth in mobile traffic. In the first set, referred to as *alternate model 1*, we envision the scenario in which all users start abandoning their desktop and laptops, in favor of mobile devices—specifically, in this model, each user visits a mobile-optimized web page 50% of the time (labeled using either GT or ML). In the second set of experiments, we envision growth in the number of users that rely *solely* on mobile devices. In this model, referred to as *alternate model 2*, we retain the behavior of the 400 baseline users, but simulate an additional 200 users that browse only (GT or ML-identified) mobile-optimized web pages (100% of the time). The distribution of the aggregate throughput for each of these forecasting experiments is also plotted in Fig 3. We find that:

- First and foremost, the distributions yielded by the ground-truth (GT) and the classified (ML) labels are quite similar to each other. In fact, we run the hypothesis testing approaches mentioned earlier to confirm that the distributions are, in fact, statistically equivalent. This is true for the baseline traffic, as well as each of the forecasted alternative models. This confirms that *web page classification, based only on anonymized TCP/IP headers, can be used to effectively conduct traffic modeling studies involving mobile web traffic*.
- The distributions suggest that an enterprise that expects an increase in next-generation users that spend most of their time on mobile devices (alternate model 2), is likely to face capacity issues earlier than one in which most users simply choose to spend more time on mobile devices (alternate model 1).

We emphasize that our intention is *not* to make forecasting claims, but simply to illustrate that our classification work can very well facilitate such traffic modeling applications.

B. Application: Building User Browsing Profiles

Applications such as behavioral-ad targeting and clickbot detection rely heavily on building user browsing profiles [24], [11]. The ability to classify web page downloads according to their content-genres, for instance, can be used to profile users according to their interests, and target relevant ads towards them, without the need for deep packet inspection. A similar approach, that classifies based on navigation-type, can be helpful in detecting clickbots or automated webcrawlers as well as in determining their intent in security applications, such as web page scraping or search engine abuse [11]. In this section, we generate synthetic user browsing "sessions" to evaluate the efficacy of our work in recovering user browsing profiles based on these two labeling schemes.

Generating Synthetic User Browsing Sessions A user typically views several web pages in any given browsing "session".

[10]Examples of features we test for, include number of servers contacted for downloading a web page, number of bytes transmitted, etc. Plots comparing the distributions can be found in [29].

TABLE IV
ACCURACY IN IDENTIFYING MOST-VISITED GENRES

No. of Top Genres	$N = 20$	$N = 50$	$N = 200$
$K = 1$	86.0%	85.2%	89.2%
$K = 2$	89.2%	89.2%	89.2%
$K = 3$	83.2%	86.8%	86.8%
$K = 4$	84.8%	81.2%	81.6%
$K = 5$	80.4%	82.0%	92.0%

Alexa [3] includes statistics on the frequency of visiting a particular web site after being on a given (different) web site. We use this data to develop a simple Markovian model of user browsing sessions as follows [39]. Each of the 3345 web pages in our data set, is represented by a state, S_i. The transition probability between state S_i and S_j is assigned based on the transition frequency from Alexa (for the corresponding web sites). This simple model is used to generate 1000 user browsing sessions, of N clicks each—the starting state for each user session is selected randomly from the 3345 states. It is important to note that users who start browsing a particular genre (shopping, for instance), are likely to keep browsing in that genre, and the Alexa statistics used here will reflect that.

How accurate are classification-based browsing profiles? Applications such as behavioral ad targeting use browsing profiles to infer what the user is likely to be *most* interested in. For instance, if the user has been recently visiting car-listing sites, it may be a good idea to pop up automobile ads for him/her. Our goal here is to analyze the accuracy with which our classified labels can help build a useful browsing profile for the user browsing sessions we simulate above. We do that, by collecting statistics on the top-K (in terms of frequency) AGL genres that a user visits in a browsing session—we collect two sets of statistics, one based on ground-truth genre labels, and the other based on classified genre labels.

In Table IV, we list the percent of users for which the set of top-K genres based on classified-labels matches perfectly with that based on ground-truth labels. We find that even the top-5 genres that a user is interested in, can be estimated perfectly for more than 80% of the users! Further, the length (N) of the browsing sessions has little impact on the estimation accuracy. These numbers are highly encouraging and suggest that targeted ad delivery can significantly benefit from web page classification based on just anonymized TCP/IP headers.[11]

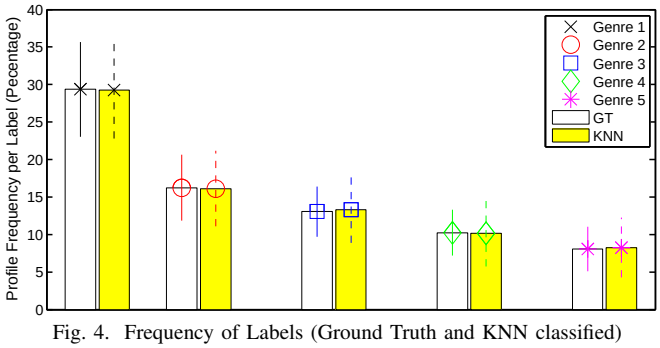

Fig. 4. Frequency of Labels (Ground Truth and KNN classified)

[11]It is important to note that even though the classification accuracy for the AGL class was only around 75% in Section IV, the user browsing profiles being constructed here are simply relying on a comparison of the *sets* of top-K genres (and not correct classification of each of the N web-page visited).

Another piece of useful information that may be needed from a user browsing profile is the *fraction* of time a user chooses to visit a particular genre. For instance, if a user visits the top genre 95% of the time, he/she is unlikely to be interested in ads related to any of the other top genres. We collect statistics on the fraction of time a user visits each of their respective top-5 genres (both based on ground-truth labels as well as classified labels). Fig 4 plots the median and 95% confidence intervals of these per-user fractions, for their top-5 genres. We find that the top-genre browsing frequencies yielded by classified labels align extremely well with those based on ground-truth labels. We conclude that web page classification is fairly well suited for building frequency-based user browsing profiles, even for content-genre based labels.[12]

VI. RELATED WORK

Traffic classification using just TCP/IP headers has received some attention and success recently—[19], [20] present an extensive summary of traffic classification methodology and applications including feature selection, and comparisons of different learning algorithms. Prior work focuses mostly on identifying the application and its protocol using features derived from a single flow [21]—web page classification, on the other hand, is a problem where the features must include information derived from multiple flows. [22] addresses a different HTTPS webmail classification problem by using Netflow data—however, it relies on features that include server IDs and server co-location which are not available in anonymized TCP/IP headers. The prime focus of this body of work, has been to map the headers to the *application and/or its protocol*, including HTTP, Mail (SMTP), Chat, etc.

[40], [41] also addressed the problem of web page classification, but used a rather small dataset ($\sim$ 50 web pages) to test their methods and focused on the problem of web page deanonymization, which focuses on identifying potential limitations of deanonymization techniques rather than the research and business potential of traffic classification. [24] attempts to go further than our goal, and even *identify* the exact web pages downloaded—their methodology, however, relies on comparing signatures of known web pages with those in traces. This is not scalable to traces collected in the wild.

VII. CONCLUSIONS

This paper advances the state of the art in traffic classification both methodologically as well as by offering new insights. Methodologically, we (i) establish the need for (and present metrics for) finding consistent (across browsers) and stable (over time) informative features, (ii) use features that span *multiple* TCP/IP flows, and (iii) use a statistical framework to study the applicability of classification results in the context of real-world applications. Our analysis leads to new insights on which multi-flow TCP/IP features are robust and informative

[12]It is important to note that identifying individual users via traffic analysis may be difficult due to network proxies and other technology. However, previous work shows that other methods for identifying malicious users behind proxies are still effective despite this limitation [11].

for web page classification, as well as what type of web page classes that can be successfully identified using these.

We are planning future work along several directions. First, we will need to address the problem of "web page boundary detection" before these results can be applied to real-world traces. Second, we will consider the impact of other factors (such as client location) on classification performance.

REFERENCES

[1] L. Popa, A. Ghodsi, and I. Stoica, "Http as the narrow waist of the future internet," in *Proc. 9th ACM Workshop on Hot Topics in Networks (Hotnets-IX)*, Oct 2010.

[2] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian, "Internet inter-domain traffic," in *Proc. ACM SIGCOMM*, Aug 2010.

[3] "Alexa," http://www.alexa.com, accessed: 2013-02-19.

[4] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying diverse usage behaviors of smartphone apps," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 329–344.

[5] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen, "How much can behavioral targeting help online advertising?" in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 261–270.

[6] "Phorm." [Online]. Available: http://www.phorm.com

[7] "Global internet phenomena report," https://www.sandvine.com/downloads/general/global-internet-phenomena/2013/2h-2013-global-internet-phenomena-report.pdf, accessed: 2014-05-04.

[8] "Half of internet traffic in north america is just to watch netflix and youtube," http://www.thewire.com/technology/2013/05/netflix-youtube-traffic/65210/, accessed: 2014-05-04.

[9] "Netflix performance on verizon and comcast has been dropping for months," http://arstechnica.com/information-technology/2014/02/netflix-performance-on-verizon-and-comcast-has-been-dropping-for-months/, accessed: 2014-05-04.

[10] "By 2017, we'll each have 5 internet devices(and more predictions from cisco)," http://www.businessinsider.com/cisco-predicts-mobile-2013-5?op=1, accessed: 2014-05-04.

[11] G. Jacob, E. Kirda, C. Kruegel, and G. Vigna, "Pubcrawl: Protecting users and businesses from crawlers," in *Presented as part of the 21st USENIX Security Symposium*. USENIX, 2012, pp. 507–522.

[12] F. Tegeler, X. Fu, G. Vigna, and C. Kruegel, "Botfinder: Finding bots in network traffic without deep packet inspection," in *Proceedings of the 8th international conference on Emerging networking experiments and technologies*. ACM, 2012, pp. 349–360.

[13] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream analysis for sybil detection," in *Presented as part of the 22nd USENIX Security Symposium*. USENIX, 2013, pp. 241–256.

[14] E. Baykan, M. Henzinger, L. Marian, and I. Weber, "Purely url-based topic classification," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 1109–1110.

[15] H. Asghari, M. Van Eeten, J. M. Bauer, and M. Mueller, "Deep packet inspection: Effects of regulation on its deployment by internet providers." TPRC, 2013.

[16] A. M. White, S. Krishnan, M. Bailey, F. Monrose, and P. Porras, "Clear and present data: Opaque traffic and its security implications for the future," *NDSS. The Internet Society*, pp. 24 096–1, 2013.

[17] "Mandatory http 2.0 encryption proposal sparks hot debate," http://www.theregister.co.uk/, accessed: 2014-05-04.

[18] "Interception and disclosure of wire, oral, or electronic communications prohibited." [Online]. Available: http://www.law.cornell.edu/uscode/text/18/2511

[19] Y.-s. Lim, H.-c. Kim, J. Jeong, C.-k. Kim, T. T. Kwon, and Y. Choi, "Internet traffic classification demystified: on the sources of the discriminative power," in *Proceedings of the 6th International COnference*. ACM, 2010, p. 9.

[20] H. Kim, K. C. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet traffic classification demystified: myths, caveats, and the best practices," in *Proceedings of the 2008 ACM CoNEXT conference*. ACM, 2008, p. 11.

[21] J. Erman, A. Mahanti, M. Arlitt, and C. Williamson, "Identifying and discriminating between web and peer-to-peer traffic in the network core," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 883–892.

[22] D. Schatzmann, W. Mühlbauer, T. Spyropoulos, and X. Dimitropoulos, "Digging into https: flow-based classification of webmail traffic," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pp. 322–327.

[23] S. Sanders and J. Kaur, "On the variation in webpage download traffic across different client types," in *Proceedings of the Ph.D. Forum, IEEE International Conference on Network Protocols (ICNP'14)*, 2014.

[24] G. Maciá-Fernández, Y. Wang, R. Rodríguez-Gómez, and A. Kuzmanovic, "Isp-enabled behavioral ad targeting without deep packet inspection," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[25] S. Ihm and V. S. Pai, "Towards understanding modern web traffic," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 295–312.

[26] B. Newton, K. Jeffay, and J. Aikat, "The continued evolution of the web," in *Modeling, Analysis and Simulation of Computer Telecommunications Systems, 2013. MASCOTS 2013. 11th IEEE/ACM International Symposium on*. IEEE, 2013.

[27] E. Gavaletz, D. Hamon, and J. Kaur, "Comparing in-browser methods of measuring resource load times," in *W3C Workshop on Web Performance 8*, 2012.

[28] M. C. Weigle, P. Adurthi, F. Hernández-Campos, K. Jeffay, and F. D. Smith, "Tmix: a tool for generating realistic tcp application workloads in ns-2," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 3, pp. 65–76, 2006.

[29] S. Sanders and J. Kaur, "Can web pages be classified using anonymized tcp/ip headers?" in *Technical Report in the Department of Computer Science at UNC-CH*, 2014.

[30] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[31] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, "A large-scale study of the evolution of web pages," in *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003, pp. 669–678.

[32] J. Cho and H. Garcia-Molina, "The evolution of the web and implications for an incremental crawler," 1999.

[33] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. C. Mogul, "Rate of change and other metrics: a live study of the world wide web." in *USENIX Symposium on Internet Technologies and Systems*, vol. 119, 1997.

[34] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Understanding website complexity: measurements, metrics, and implications," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 313–328.

[35] A. Este, F. Gringoli, and L. Salgarelli, "On the stability of the information carried by traffic flow features at the packet level," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 3, pp. 13–18, 2009.

[36] "Statcounter," http://gs.statcounter.com/, accessed: 2013-06-30.

[37] A. Rao, A. Legout, Y.-s. Lim, D. Towsley, C. Barakat, and W. Dabbous, "Network characteristics of video streaming traffic," in *Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*. ACM, 2011, p. 25.

[38] "Mashable," http://mashable.com/2013/08/20/mobile-web-traffic/, accessed: 2013-11-27.

[39] F. Chierichetti, R. Kumar, P. Raghavan, and T. Sarlós, "Are web users really markovian?" in *WWW*, 2012, pp. 609–618.

[40] T.-F. Yen, X. Huang, F. Monrose, and M. K. Reiter, "Browser fingerprinting from coarse traffic summaries: Techniques and implications," in *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2009, pp. 157–175.

[41] Q. Sun, D. R. Simon, Y.-M. Wang, W. Russell, V. N. Padmanabhan, and L. Qiu, "Statistical identification of encrypted web browsing traffic," in *Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on*. IEEE, 2002, pp. 19–30.