Flexible Traffic and Host Profiling via DNS Rendezvous

David Plonka
University of Wisconsin-Madison
Email: plonka@cs.wisc.edu

Paul Barford
University of Wisconsin-Madison
Qualys, Inc.
Email: pb@cs.wisc.edu

***Abstract*—The ability to accurately classify network traffic and to perform timely detection of the presence of unwanted classes of traffic has important implications for network operations and security. In recent years, classification has become more challenging due to applications that use ports that are not well-known, that overload or masquerade with other applications' well-known ports, and that may encrypt or otherwise obfuscate their payload. The goal of our work is to develop a method for traffic classification that is *flexible, i.e.,* that can be used to create arbitrary organizations of traffic from coarse to fine-grained groups, and can identify encrypted traffic as well as new applications. In this paper, we present a novel method for classification based on analyzing *rendezvous traffic* (*i.e.,* the traffic preamble in which a given host determines the remote IP address of a peer host or service) that usually precedes application traffic. Our approach exploits the most widely used rendezvous service, the Domain Name System (DNS). Specifically, through careful tracking of client IP addresses, alpha-numeric domain names, and answer IP addresses in rendezvous traffic, we apply classification labels to end-hosts and their traffic reported by flow-export data. Additionally, we present the notion of *host profiling* as a method for expanding traffic classification in cases where there is not a direct match between rendezvous traffic and application traffic. To assess the feasibility of our method, we perform a focused case study on one day in the lives of two drastically different user end-host populations: office and residential. Our results demonstrate the efficacy and capability of a DNS rendezvous-based method of classification that performs well even in situations where application payload is encrypted (or unavailable) or when application traffic is monitored by packet sampling.**

I. Introduction

The past decade has seen an explosion of new network applications such as peer-to-peer (P2P) file sharing, online social networks, gaming, and VoIP. Each application requires certain network resources so that users have a satisfactory experience. The past decade has also seen many forms of malicious network use and abuse such as denial-of-service attack bot nets, phishing scams, and thefts due to compromised host or protocol insecurity. Both the ability to discriminate between application types in live traffic streams and to identify suspicious hosts is critical in order to ensure the desired level of application performance and reliability in an enterprise.

Traffic classification was originally based solely on port numbers in IP packet headers. When the scope of applications was almost entirely limited to well-known ports, this approach was effective for identifying a large proportion of traffic. However, now there are many applications that operate on a wide variety of ports, and some ports, such as 80, are frequently overloaded since they are rarely blocked, which further complicates classification. This led to development of statistical classification methods that consider properties extracted from flow-export data and behavior-based classification methods that consider host behavior on several levels. Modern traffic classification methods are challenged to accurately identify nascent applications on high-performance networks carrying intentionally obscured or encrypted traffic. While statistical and prior behavior-based methods can be effective at placing traffic into coarse-grained groups such as P2P or WWW, they are limited in their flexibility, *i.e.,* their ability to accurately assign traffic to arbitrarily specified groups; they can falter with insufficient knowledge due to practicalities such as routing asymmetries and packet sampling on high-capacity links.

In this paper, we describe a new method for traffic classification that taps a traffic-independent source of information and enables flexible organization of traffic types into arbitrary groups. Our classification methodology is based on monitoring and analysis of the traffic generated by rendezvous services that are used by Internet applications. A rendezvous service is typically operated independently of its clients and enables a client application to identify the IP addresses that are the target for communication and are known to the user by an alpha-numeric name. The canonical example of a rendezvous service is the Domain Name System (DNS), which is used by most Internet applications. The first step in our method is to build and maintain a table of active local clients by IP addresses and their respective target remote IP addresses along with the associated alpha-numeric names extracted from the locally-observed DNS traffic.

Next, we search a target database of flow-export records collected in the same enterprise for local and remote IP address pairs that match entries in that table populated with the preceding rendezvous information. If there is a match, the corresponding alpha-numeric (DNS) name is the basis for classification.

Both the type of rendezvous mechanism employed (*e.g.,* DNS, static, DHCP, algorithmic, etc.) and its intrinsic characteristics offer opportunities for detailed classification at a level that has not been possible with prior methods. For example, the simple fact that a given host employed the DNS to rendezvous

with "www.example.com" via HTTP, may allow both that client host and the exchanged traffic to be classified as WWW and exclude it's misclassification as P2P. And, since the DNS uses names that follow a well-known domain hierarchy, the DNS name hierarchy can serve as one way to organize and aggregate resultant traffic, but others are possible as well. For example, names can be organized into categories that are user defined, come from a standard source (*e.g.,* [4]), or are based on application type (*e.g.,* WWW, FTP, online social networks, or streaming) or by subject (*e.g.,* weather or sports). This ability to create arbitrary traffic groups may offer network operators significant flexibility in how they manage traffic going well beyond what port-based methods offer.

The key technical challenge in developing a DNS rendezvous-based classifier is that it must monitor host rendezvous traffic and, in a timely fashion, link the information gleaned with corresponding observations of the application traffic to be classified. For many an institution or enterprise, the typical scenario involves a set of client hosts that utilize a locally-designated recursive DNS service (often located near their LAN) with those hosts' application traffic passing through some interesting observation point within a network element such as a high-capacity switch or border router. The observation point of the DNS rendezvous traffic need not be the same as that of the *target* traffic to be classified. With this model, we develop a software classifier that can accommodate parallel traces from multiple observation points.

To assess and evaluate the capabilities and effectiveness of our method, we collect DNS query-response traffic and flow-export records from a campus network infrastructure for over a year. Analysis of this data exposes many interesting features such as well known diurnal behavior, frequent spikes in DNS traffic, and a qualitatively different DNS behavior for subgroups within the user population in a case study we present that considers traffic for a typical day. We separate two distinct user populations: a large office/staff group and a large residential/student group. We characterize and contrast the DNS and wide-area traffic of each group showing that, while the general types are similar, the quantity of each type is dramatically different. In particular, over 90% of the office traffic is classified by domain name. Less of the residential traffic can be classified by name, ostensibly due to the use of P2P and other applications that do not rendezvous based solely on the DNS. Serendipitously, however, we find that any DNS rendezvous classification discriminates traditional client-server application from P2P application traffic.

This work makes the following contributions. *(1)* We introduce the idea of rendezvous-based traffic classification. *(2)* We demonstrate the feasibility and capabilities of DNS rendezvous-based traffic classification by developing a tool and analyzing traffic for two diverse populations of users. *(3)* We show how DNS rendezvous-based classification complements and improves upon port-based classification.

II. Related Work

Internet traffic classification methods have been proposed and evaluated in a number of prior studies. To the best of our knowledge, none have proposed a DNS rendezvous-based approach.

Trestian *et al.* [22] perform traffic classification by first classifying end-hosts based on results from the Google search engine. They utilize a database of information that can be queried publicly on the web, with the hope that it contains correct and timely information about end-hosts of interest. The inspiration for their work is similar to ours in that in order for Internet communication to progress, an end-host must somehow discover the remote IP address with which to communicate. Their classification based on matching words in domain names could be applied to create aggregates for our DNS rendezvous-based approach.

The recent work by Kim *et al.* [13] provides a thorough overview and performance comparison of popular traffic classification methods and implementations from the literature and from practice. We utilize their port-based classification in presenting portions of our results. However, our method's labels and classes differ, so we report how our classes compliment theirs. We refer the reader to [13] for a survey of other prior work involving traffic classification based upon port [16], [1], payload [20], host-behavior [12], or flow-features [14].

Karagiannis *et al.* [12] introduced BLINC, a classification method based on host-behavior. Our method is similar to BLINC in that we do not rely directly on ports nor target traffic payload and is also similar in that our host classification employs a kind of "social" profile of each host. However, our traffic classes are based on innumerable domain names rather than a small, fixed set of application groups. Also, our method neither employs heuristics nor requires tuning based on previously observed behaviors. It has been found [22], [13] that BLINC's graphlet approach experiences problems when the target traffic is sampled ("1 in n" packets) or when the target traffic is not observed symmetrically at a gateway near the end-hosts. Thus, BLINC has more stringent requirements for deployment and operation than our method and so we use the port-based method to compare and contrast with our results (Section V). Our technique is robust in the face of sampling because it gleans social behavior of hosts from separable, low-volume DNS rendezvous traffic rather than from the aggregated, high-volume target traffic.

Both Cho *et al.* [9] and Estan *et al.* [10] describe and implement traffic measurement systems that use the hierarchical IP address space to profile or classify traffic in aggregates. Somewhat similarly, our work employs a hierarchy, but instead uses the hierarchical domain name space to form aggregates classes; domain names have advantages in terms of readability and persistence over IP addresses. Additionally, our classification groups hosts with similar profiles, and thus bears some similarity to aggregation in these prior works, but without our having to rely on structural cues from the hosts' IP addresses.

Some commercial products perform traffic classification and

filtering using identifiers that often contain domain names. Products such as Websense [6] and SmartFilter [5] inspect application traffic payload for identifiers such as URLs (and may optionally perform reverse DNS lookups). Our method differs in that it observes the content of the participating clients' DNS rendezvous traffic and thus can be effective in environments when it is infeasible to inspect the target traffic (*e.g.,* due to traffic volume, encryption, or policy). Alexa Internet [4] provides web traffic metrics labeled by domain name, such as top site lists and demographics. Their service is web-specific and observes Uniform Resource Locators (URLs), whereas our work considers all traffic and observes fully-qualified domain names (FQDNs). However, we advocate the use of Alexa's categories as a convenient basis for our operator-defined host classification.

While we focus on DNS-based rendezvous, prior work has described alternative rendezvous mechanisms. For example, Morris *et al.* propose a distributed hash table-based mechanism [15] and Walfish *et al.* propose replacing DNS with another mechanism for the World-wide Web in [23]. Baset and Schrulzrinne [8] and Rossi *et al.* [19] reverse engineer and infer Skype's application-specific rendezvous mechanism. There are standard rendezvous protocols other than DNS, *e.g.,* SIP [11], and P2P variant works in progress (*e.g.,* P2PSIP [3]).

Finally, there are tools and visualizations that are related to our work. Wessels *et al.* [24], [25] provide a tool (*dnstop*) to measure DNS traffic by volume per client. Based on that tool, Plonka *et al.* [17] introduced *TreeTop*, a tool that implements domain name-based traffic measurement in aggregate. Our work improves TreeTop to track and report individual client's DNS activity and our results differ in that we apply that DNS information to label both traffic for traditional client-server applications (*e.g.,* World Wide Web and Streaming) and peer-to-peer traffic (*e.g.,* BitTorrent, Skype, and Massively Multi-player Online Gaming). Shneiderman [21] originated the treemap visualization that we employ to represent hierarchical data.

III. Empirical Data sets

In this work we are interested in applying information gleaned from DNS queries and corresponding replies, exchanged between end-hosts and their trusted recursive name servers within an enterprise, to the task of classifying that enterprise's wide-area traffic. To this end, we monitor a campus' traffic at two observation points: *(1)* the campus clients' name servers, and *(2)* one of the campus border routers that handles much of wide-area traffic including that for the commodity Internet. We perform full packet capture at the campus domain name servers, and collect packet-sampled flow data at the border router. [1] Thus, the payload of the DNS traffic is recorded, but the application traffic payload is not. Our interest is in the "canonical" DNS traffic, *i.e.,* the standard DNS traffic expected to precede application traffic that consists of a query by fully-qualified domain name (FQDN) and an answer containing one or more IP addresses associated with the query name.

Prior work has shown that traffic classification results can vary widely based on the trace traffic mix and observation point [13]. As such, while we monitor traffic for a single institution, we select two of its end-host/client populations that have very different characteristics, namely an *office* and a *residential* population. To expose the details within the limited space available here, we present results for a single representative day. (Classification results from other days are consistent with the results reported here.)

Table I summarizes the characteristics of the data sets. We studied, in detail, the traffic on one typical day selected at random: April 17, 2009. Both the office and residential data sets consist of *(1)* all the recursive DNS traffic between end-hosts and the campus DNS service and *(2)* the packet-sampled flow records collected at the campus border that represent wide-area traffic (see also Figure 4); only flow records involving campus hosts for which we've seen recursive DNS traffic involving the trusted campus DNS server are considered.

Data Set	Clients	Unique FQDNs	DNS Reply Pkts	DNS Reply Volume (ave. bps)	Wide-Area Out / In Volume (ave. bps)
Office	614	19.4 K	560K	12.2K	753K / 5.66M
Residential	9,819		15.7M	360K	244M / 276M
(subpop.)	(5,583)	(143 K)			

TABLE I: Characteristics of 24-hour data sets analyzed. The average wide-area traffic volume is estimated from packet-sampled flows. The parenthesized values are for a residential subpopulation that was used for the TreeTop-based results in Section V. From the inbound and outbound volume values, we see that the office population primarily consumes wide-area Internet content, whereas the residential population both consumes and provides a significant amount of content.

A. Office Traffic

The "office traffic" involves a group of staff employees on the campus. The office users are bound by the campus Appropriate Use Policy for information technology resources (that tolerates incidental personal use) and their end-hosts are typically owned by the university and located in campus offices with wired Ethernet connections. During the course of the day under study, we observed 614 end-hosts with an average (over 24 hours) of 180 active hosts performing DNS queries per 5 minutes. The office wide-area traffic and DNS traffic volume and rate values are shown in Tables I.

B. Residential Traffic

The "residential traffic" involves a subset of the students living in residence halls on a campus. The residential users are bound by the same Appropriate Use Policy as the office

[1]The flow data is based on a 1 in 1024 packet sampling rate using the "cflowd" feature on a Juniper router with 10-gigabit Ethernet interfaces; we report all our target traffic volume measurements by bits or bytes (approximated by multiplying sampled values by 1024).

users, but their end-hosts are privately-owned and located in private residences that have wired Ethernet connections. During the course of the day under study, we observed 9,819 end-hosts with an average (over 24 hours) of 1,886 active hosts performing DNS queries per 5 minutes. The residential wide-area traffic and DNS traffic volume and rate values are shown in Tables I.

IV. Analysis Method

In this work we analyze and classify the DNS and wide-area (application) traffic using an improved version of the TreeTop tool [17]. Specifically, we've enhanced TreeTop to track and report the relationship between IP addresses and domain names on a *per-client* basis.

In short, TreeTop processes pcap traces of combined DNS and application traffic, requiring the payload of DNS packets but only the transport header information of other traffic to be classified. It observes all DNS replies to each client and, when there is a successful response (*i.e.,* NOERROR code) to a DNS query for an IP address (*i.e.,* type A or AAAA), TreeTop *(a)* stores the query name in a central *domain tree* (an n-ary prefix search tree), *(b)* stores the IPv4 and/or IPv6 address answers in a client-specific *address tree* (a binary prefix search tree), and *(c)* links nodes in the client's address tree to their corresponding nodes in the domain tree. Thus, these data structures store per-client *DNS rendezvous state information* as to which remote IP addresses are known by domain name. Subsequently, when TreeTop observes application traffic (*e.g.,* the wide-area traffic at a network's border router), it uses the rendezvous state information to label the client traffic as either "unnamed" or as "named," and accumulates per-client traffic counters (in bytes or packets) for those meta-categories as well as for hierarchical sub-categories by domain name.

To prepare the data sets for TreeTop, we synthesize pcap files from the flow data (with a modified flow-export utility [18], [2]) and merge them with the DNS pcap data (using mergecap) to form one coherent input data set. Note that, in general, it is sufficient for the DNS pcap records to be observed before the application traffic pcap records (from the flow data); so, for off-line studies, we can perform a single batch analysis for an entire day using TreeTop by first reading all DNS traffic data then the application traffic data. By contrast, performing an online analysis (at one observation point) obviates the need to carefully interleave the DNS and target traffic records based on their packet arrival times because the DNS responses are interspersed in the trace with the target traffic (to be measured) and would be observed before the subsequent associated target traffic.

A. Traffic Labels

1) Direct Classes: Direct DNS rendezvous-based traffic classification involves at least two sorts of traffic classes. The first, are the "named" and "unnamed" traffic classes, which simply indicate whether a client end-host knows the traffic's remote IP address by a domain name as the result of a canonical "forward" DNS query to translate that name to an address. The second and more challenging traffic classes are the domain names themselves. To deal with the innumerable fully-qualified-domain-names (FQDNs) that may exist in the world-wide DNS, we treat them hierarchically. For instance, traffic involving the FQDN "www.example.com" is in the "com" class, the "example.com" class, and "www.example.com" class, and thus can be presented at a number of levels of granularity. One can imagine categorizing domain names by common owner (*e.g.,* "facebook.com" and "fbcdn.net"), similar purpose (*e.g.,* weather or sports content), or even application groups such as WWW, FTP, Streaming, etc. We leave such classification by policy or operator objectives for future work by using readily available references [4].

2) Indirect Classes: Our indirect DNS rendezvous classification utilizes host profiles that are defined by configurable sets of domain names. We defined three such profiles for P2P clients. The P2P profiles are: "Torrent" (BitTorrent client applications, directories, and trackers), "Talk" (Skype and Google Chat applications), and "Game" (Massively Multi-player Online Games). For instance, a client end-host that issued a DNS query trailing with "bittorrent.com" or "utorrent.com" will be profiled as a "Torrent" client because of its ostensible interest in a popular BitTorrent client application. Likewise, a client end-host that issued a DNS query trailing with "thepiratebay.org" will be profiled as a "Torrent" client because of its interest in this popular BitTorrent tracker site. These host profiles then, are used to label traffic classes. For example, the "Torrent" label would be used for traffic exchanged by a host having only the "Torrent" client profile; the "Talk+Game" label would be used for traffic involving a host having both the "Talk" and "Game" (but not "Torrent") profile. Note that we do not claim that "Torrent"-labeled traffic is necessarily BitTorrent traffic; instead, we claim that it certainly involves an end-host that matched the Torrent profile and is thus (at least indirectly) associated with this P2P application. Each profile is defined by a set of domains that were assembled from readily available references [7], [4]. The Torrent domains (31) are popular BitTorrent Clients from Wikipedia and from Alexa's top "Torrent Directories and Tracker Domains." The Talk domains (2) are from observed behavior of the Skype application and Google Chat. The Game domains (35) consist of a well-known online game domain and Alexa's top "Massive Multiplayer Online Domains."

3) Port-based Classes: In addition to our rendezvous-based labels, we use traditional port-based application labels from an existing classifier [1] that has been used in prior work. [13] These are: "WWW," "P2P," "FTP," "Net. oper.," "Mail/News," "Streaming," "Encryption," "Games" (distinct from "Game" which is an indirect host profile-based label), "Chat," "Login," "Tunnels," "Other," and "UNKNOWN." While many services can be uniquely identified solely by that service's FQDN, port-based classes offer the advantage of familiarity and of distinguishing amongst multiple services that happen to be identified by a single (unspecific) FQDN.

4) Classification Order: In this study, we take a pragmatic approach based on flexible classification that emphasizes the

complementary strengths of each method. First, we label the traffic with the direct DNS rendezvous-based classes (named and unnamed). Next, for results that involve port-based classification, we label the traffic using port-based classes, nested within "named." Finally, we label unnamed traffic with indirect labels based on profiling hosts by DNS rendezvous. This initial choice of order, we argue, is from the least to most speculative. In the general case, a complementary method, such as port-based, could be performed at any point; other orderings provide opportunities to explore how the rendezvous classes overlap with other classification schemes.

V. Results

In this section we report the results of a "day in the life" of an office and a residential user population, in terms of their DNS and wide-area application traffic.

A. DNS Traffic Analysis

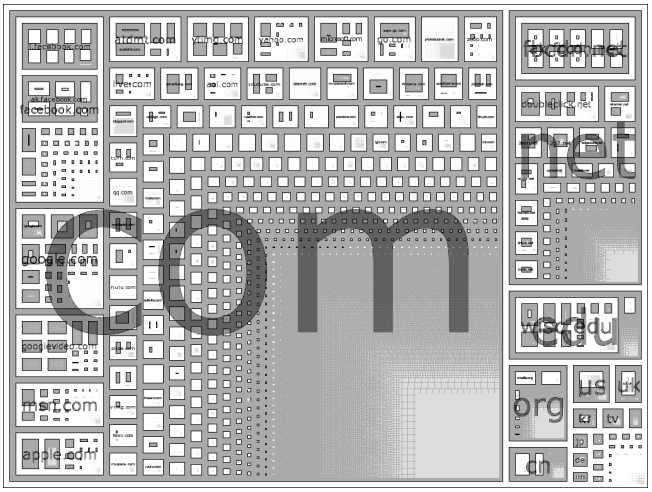


Fig. 1: A treemap of domain popularity for all domain names queried that were answered with IP addresses during one day. This treemap for the residential population represents 142,594 unique FQDNs. The relative size of the rectangles indicate the domain names' relative popularity based on the number of IP address answers that a client knows as being associated with that name. The more clients that knew IP addresses associated with a given domain name, the more prominently it is shown.

Figures 1 is a treemap of domain names based on their popularity for the residential population. (The treemap for the office population was visually similar and thus omitted.) For the day under study, the residential population resolved roughly 7 times more unique FQDNs than the office population (142.6K vs. 19.4K) in DNS queries from about 9 times as many client end-hosts (5,583 residential vs. 614 office clients). From values in Table I we can see that there are roughly 32 and 26 unique FQDNs per office and residential client end-host, respectively, on this day. Many of the most popular domains are common between the office and residential populations, including "google.com", "facebook.com" and the associated Content Distribution Network (CDN) "fbcdn.net", "yahoo.com", "apple.com", "microsoft.com" or "msn.com", and the local campus' domain. The least popular domains, such as those that only a single host might know, are minuscule in the treemaps, and thus form the light gray fields in the lower right of the rectangles.

To further explore the popularity of FQDNs amongst these populations, Figure 2 shows the unique FQDNs known, ordered by popularity, *i.e.*, the FQDN numbered 1 on the horizontal axis is the most popular and that numbered 10 is the tenth most popular within the given user population. This figure clearly shows that most FQDNs are known by only a small percentage of the population. Specifically, only those FQDNs in the top 1000 are known by more than 5% of the hosts. The raw data from TreeTop shows that more than 68% of the FQDNs were known to only a single client end-host during this day. This underscores the need to aggregate the numerous FQDNs in some fashion, and here we do so hierarchically, beginning with TLDs, then second-level domains, and so on.

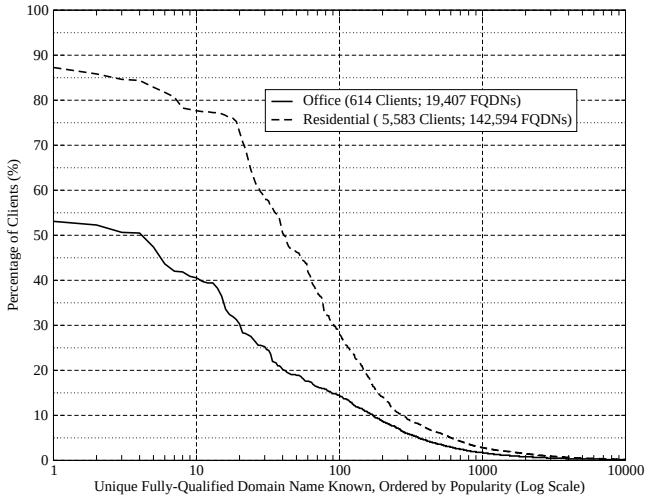


Fig. 2: Popularity of FQDNs by client end-hosts during one day. Here we see that the most popular 10 and 100 FQDNs are known in common to more than 50% and 10% of clients, respectively. Note that the popularity ranking for office and residential populations were determined independently, thus it is unlikely that they share the same FQDN at a given rank.

We also examined the distribution of clients based on the number of unique remote IP address answers known by domain name to the client. For these data sets, we find that 95% of the office and residential hosts learned (via DNS answer replies) of fewer than 1000 unique remote IP addresses by a domain name, and that more than 99% of all the hosts learned fewer than 2000 unique remote IP addresses by domain name throughout the entire course of this day.

B. Traffic Classification Results

Because our DNS domain name-based classification approach uses drastically different labels than prior classification work, we do not have a straightforward means of comparing performance. However, because our direct DNS rendezvous approach classifies based on domain names and IP address answers observed in each client's DNS traffic, it can be considered tacit ground truth. That is, we are certain that the client end-host had the opportunity to know the remote IP address by that name. However, we are guided by finding 1 of Kim *et al.,* [13]:

> [The] port-based approach still accurately identifies most legacy applications [...] this suggests that ports still possess significant discriminative power in classifying certain types of traffic.

Our DNS rendezvous-based approach and a port-based approach are similar in that both of them label traffic based solely on easily-observed traffic elements, instead of labeling using heuristics and tunable thresholds.

1) Port-based Classification: We first classify the inbound and outbound traffic for our two populations using a port-based approach to set a baseline for comparison. Specifically, traffic identified by well-know ports is labeled either as one of 12 pre-defined application groups or UNKNOWN.

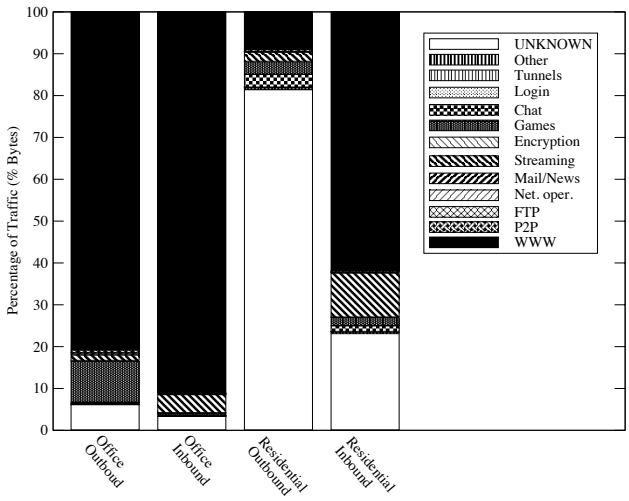


Fig. 3: Port-based classification of traffic (bytes) for the office and residential populations during one day. While 93.9% (outbound) and 96.6% (inbound) of the office traffic is labeled (*i.e.,* not UNKNOWN), only 18.6% (outbound) and 76.9% (inbound) of the residential traffic is labeled due to the different application traffic mixes. Note the coarse labeling as only the WWW, Games, and Streaming applications represent 10% or more of the traffic by volume.

Figure 3 shows the classification of traffic using the simple port-based method. [2] Here we see a stark difference between the office and residential traffic; most of the office traffic is classified, but much less of the residential traffic is classified. Furthermore, the application mix differs greatly in these two populations, with over 80% of the office traffic being labeled as WWW and only about 5% unknown, whereas less that 10% of the outbound residential traffic is labeled WWW, and more than 80% being left unidentified.

2) Direct Rendezvous-based Classification: We now classify the same office and residential traffic by our direct DNS rendezvous-based approach using labels as described in Section IV-A1. We consider two broad rendezvous-based classes, "named" and "unnamed," then detailed sub-classes by domain name.

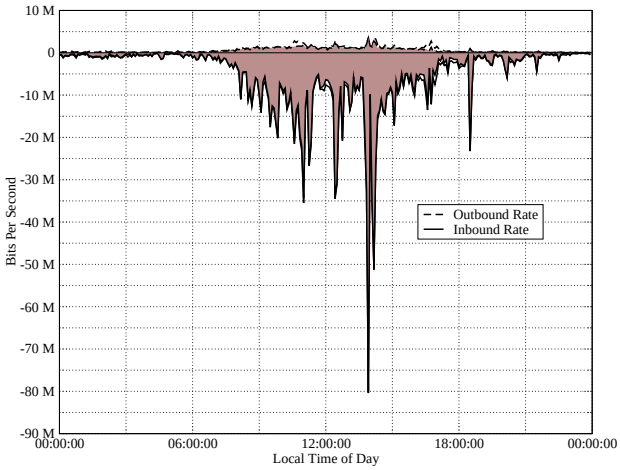
Figure 4 shows the time series volume for the named and unnamed portions of the office and residential traffic throughout the day under study. We see that nearly all office traffic involves DNS rendezvous and can be named. While a significant amount of inbound residential traffic can also be named, 32.1% (inbound) and 93.3% (outbound) is unnamed and, therefore, apparently does not employ the DNS for rendezvous. Also, note the correspondence between the portion of named traffic identified here by our method and that labeled by the port-based method shown in Figure 3; this suggests that DNS-named traffic very often uses well-known ports, *e.g.,* traditional client-server applications.

While we have omitted the traffic volume detailed by specific domain names due to space limitations, we represent these by treemap as in Figure 1, so that the domains involving the highest traffic volume are largest. For instance, of the named residential inbound traffic, *i.e.,* from source IP addresses that the clients know by domain name, the following are amongst the most significant: "facebook.com", "googlevideo.com", and "edgefcs.net". The prominence of this last domain led us to discover that the majority of traffic that is named by our method yet UNKNOWN to the port-based method is associated with the "edgefcs.net" domain. This domain hosts streaming content (presumably on Macromedia Flash Communication Servers, hence the name "fcs") atop the Akamai CDN. These servers deliver content by the proprietary Real Time Messaging Protocol (RTMP, port 1935) or by tunneling via HTTP (port 80) and HTTPS (port 443). Thus informed, we updated the well-known ports database so that RTMP traffic is properly classified as Streaming in Figure 3. This example illustrates how the DNS rendezvous-based classification, by examining the "forward" domain name by which the clients accessed this service, can assist in nascent protocol identification leading to an improved port-based method.

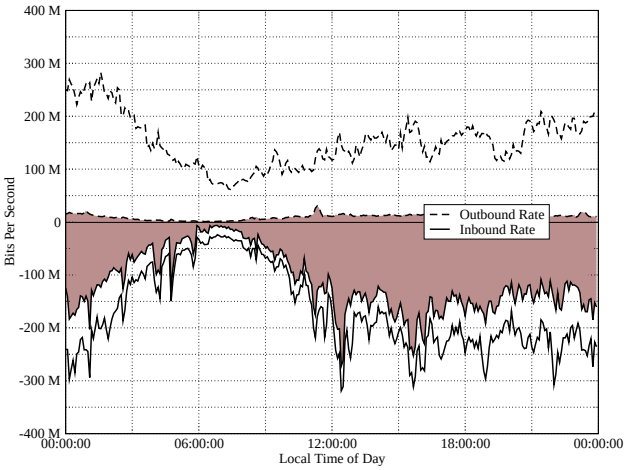
3) Host Profiling and Classification Results: We now apply our indirect DNS rendezvous-based approach, using labels as described in Section IV-A2.

As shown in Figure 4b, our direct DNS rendezvous-based classification method determined that only 6.7% of the out-

[2]Our application group classes are those identified by CoralReef [1], specifically, coral-3.8.4, and thus are equivalent to those used in the work of Kim *et al.* [13].



(a) Office



(b) Residential

Fig. 4: Wide-area traffic rate, as observed at campus border during one day. Outbound rate (from campus) is plotted above the horizontal axis and the corresponding inbound rate (to campus) is plotted below. Clearly the office population is primarily a consumer of wide-area Internet content, whereas the residential population is both a significant consumer and provider of content. The portion of "named" traffic (*i.e.*, by DNS rendezvous) is shaded; while 81.1% (outbound) and 93.2% (inbound) of the office traffic is named, only 6.7% (outbound) and 67.9% (inbound) of the residential traffic is named.

bound residential traffic was named, and, in Figure 3, we see the majority of this traffic is UNKNOWN by port. We expect this unnamed traffic might be dominated by P2P file transfer (*e.g.*, BitTorrent), game, and/or talk (*e.g.*, VoIP) traffic, *i.e.*, those groups of applications that do not typically use the DNS for rendezvous and also often use unreserved (not well-known) port numbers.

To classify this traffic, we employ the DNS rendezvous information *indirectly* by labeling local hosts according to P2P client *profiles* based on their DNS rendezvous activity. The resulting assignments are shown in Figure 5.

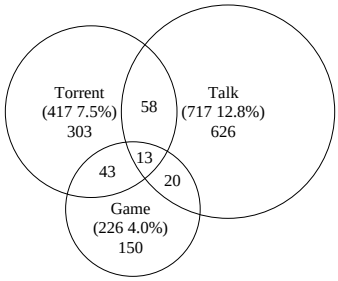


Fig. 5: Residential subpopulation host counts by P2P application type based on their DNS queries during one day. Here we see that 1,252 hosts (22.4% of 5,583 total) appear to run one or more P2P applications. (Parenthesized values are totals for that subpopulation's circle.)

Then, in Figure 6, we correspondingly label portions of the unnamed residential outbound traffic (93.3% unnamed, as seen in Figure 4b). That is, when traffic is classified as "unnamed," we determine if that traffic involved one of the 1,252 P2P

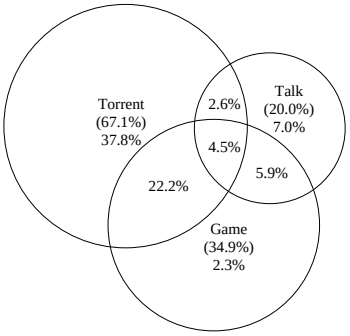


Fig. 6: Residential unnamed outbound traffic volume (bytes) by P2P client profile. Here we see that 67.1% of this unnamed outbound traffic and involved local hosts that were profiled as BitTorrent clients based on their DNS rendezvous activity. (Parenthesized values are totals for that subpopulation's circle.)

profiled residential hosts, and if so, we label that portion of the traffic by the given host's P2P profile name: "Torrent," "Talk," and/or "Game." For instance, clients running the Skype application are known to resolve "ui.skype.com", thus this is one of the domain names that causes it to fit the "Talk" P2P profile. While somewhat speculative, DNS rendezvous profiles are flexible and configurable; we find that our initial effort attributes 82.3% of the otherwise unlabeled traffic to the 22.4% of the hosts that fit a P2P profile, indicating the traffic was sourced from hosts that had resolved popular Torrent, Talk, or Game-related DNS domain names.

4) Results Summary: Table II summarizes the overall classification performance of the port-based method and ours.

Data Set	Port-known	DNS-named and Port-known	DNS-named	DNS-named and DNS-Profiled
Office Out	93.9%	80.5%	81.8%	91.9%
Office In	96.6%	91.8%	93.2%	95.4%
Residential Out	18.6%	6.2%	6.7%	83.5%
Residential In	76.9%	58.3%	67.9%	88.2%

TABLE II: Traffic classified (bytes) by each method: Port-known (by the port-based method), DNS-named (DNS rendezvous named), DNS-named and DNS-Profiled (DNS rendezvous named plus unnamed matching a P2P host profile).

The significant proportion of "DNS-named" traffic that also has "Port-known" for the office traffic (98%) suggests that one can be somewhat confident in the port-based method there. The lesser proportion for the residential traffic (86% outbound, 93% inbound) suggests that port-based result is suspect given that traffic mix. Lastly, for residential outbound traffic, we realize a 64.9% increase in volume classified by our DNS rendezvous method over the port-based method.

VI. Conclusion

In this paper we present a novel traffic classification method based on DNS rendezvous, *i.e.,* the domain names by which end-hosts present and discover IP addresses. Our rendezvous-based approach combines some of the best characteristics of prior methods: *(i)* port numbers are not implicitly trusted, *(ii)* deep packet inspection of the target traffic is not required, and *(iii)* packet sampling of the target traffic is not an obstacle. The goal of our work is to add flexibility in classification with high accuracy of classification in live operational deployments. This approach gleans information from the most common rendezvous method, the DNS, which is widely used and offers flexible options to both profile hosts and classify their traffic.

We demonstrate the feasibility and utility of rendezvous-based classification by implementing our method in the TreeTop tool and applying it to DNS traces and flow-export data gathered from a campus network, focusing on two starkly different user groups' traffic for a typical day. We show that a large proportion of the traffic from the office group is arranged via the DNS, enabling it to be directly classified by our method. In the residential group, where a significant amount of traffic is not is preceded by DNS queries, we implement two alternatives: *(i)* we apply the port-based method selectively, to just the named traffic, to minimize that method's false reports, and *(ii)* we infer labels for unnamed traffic by profiling the end-hosts involved, based on their DNS activity. These initial results demonstrate how a traffic classifier can make effective use of a hitherto untapped, independent source of information, *i.e.,* the Domain Name System.

Acknowledgments

This work was supported in part by NSF grants CNS-0716460, CNS-0831427 and CNS-0905186. Any opinions, findings, conclusions or other recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the NSF.

References

[1] CoralReef. http://www.caida.org/tools/measurement/coralreef/, 2008.
[2] flow-tools: Tool set for working with NetFlow data. http://code.google.com/p/flow-tools/, 2009.
[3] Peer-to-Peer Session Initiation Protocol (p2psip). http://www.ietf.org/dyn/wg/charter/p2psip-charter.html, 2009.
[4] Alexa Internet, Inc. http://www.alexa.com/topsites, 2010.
[5] McAfee SmartFilter. http://www.mcafee.com/us/enterprise/products/email_and_web_security/web/smartfilter.html, 2010.
[6] Websense, Inc. http://www.websense.com, 2010.
[7] Wikipedia: Comparison of BitTorrent clients. http://en.wikipedia.org/wiki/Comparison_of_BitTorrent_clients, 2010.
[8] S.A. Baset and H. Schulzrinne. An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol. In *Proceedings of IEEE INFOCOM '06*, Barcelona, Spain, April 2006.
[9] K. Cho, R. Kaizaki, and A. Kato. Aguri: An Aggregation-Based Traffic Profiler. In *Proceedings of the Workshop on Quality of Future Internet Services (QofIS '01)*, Coimbra, Portugal, September 2001.
[10] C. Estan, S. Savage, and G. Varghese. Automatically Inferring Patterns of Resource Consumption in Network Traffic. In *Proceedings of ACM SIGCOMM '03*, Karlsruhe, Germany, August 2003.
[11] G. Camarillo A. Johnston J. Peterson R. Sparks M. Handley E. Schooler J. Rosenberg, H. Schulzrinne. SIP: Session Initiation Protocol. IETF RFC 3261, June 2002.
[12] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel Traffic Classication in the Dark. In *Proceedings of ACM SIGCOMM '05*, Philadelphia, PA, August 2005.
[13] H. Kim, kc claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee. Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices. In *Proceedings of the 4th ACM International Conference on emerging Networking EXperiments and Technologies (ACM CoNEXT 2008)*, Madrid, Spain, December 2008.
[14] A.W. Moore and D. Zuev. Internet Traffic Classification Using Bayesian Analysis Techniques. *ACM SIGMETRICS Performance Evaluation Review*, 33(1):50–60, 2005.
[15] R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-To-Peer Lookup Service for Internet Applications. In *Proceedings of ACM SIGCOMM 2001*, San Diego, CA, August 2001.
[16] D. Plonka. FlowScan: A Network Traffic Flow Reporting and Visualization Tool. In *Proceedings of the USENIX Fourteenth System Administration Conference (LISA XIV)*, New Orleans, LA, December 2000.
[17] D. Plonka and P. Barford. Context-aware Clustering of DNS Query Traffic. In *Proceedings of the ACM SIGCOMM / USENIX Eighth Internet Measurement Conference (IMC 2008)*, Vouliagmeni, Greece, October 2008.
[18] S. Romig and M. Fullmer. The OSU Flow-tools Package and Cisco NetFlow Logs. In *Proceedings of the USENIX Fourteenth System Administration Conference LISA XIV*, New Orleans, LA, December 2000.
[19] Dario Rossi, Marco Mellia, and Michela Meo. Understanding Skype signaling. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 53(2):130–140, 2009.
[20] S. Sen, O. Spatscheck, and D. Wang. Accurate, Scalable In-network Identification of P2P Traffic Using Application Signatures. In *Proceedings of the 13th International Conference on World Wide Web*, pages 512–521. ACM New York, NY, USA, 2004.
[21] B. Shneiderman. Tree Visualization with Tree-Maps: 2-d Space-Filling Approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992.
[22] I. Trestian, S. Ranjan, A. Kuzmanovi, and A. Nucci. Unconstrained Endpoint Profiling (Googling the Internet). In *Proceedings of ACM SIGCOMM 2008*, Seattle, WA, August 2008.
[23] M. Walfish, H. Balakrishnan, and S. Shenker. Untangling the Web from DNS. In *Proceedings of the 1st USENIX/ACM Symposium on Networked Systems Design and Implementation (NSDI '04)*, San Francisco, CA, March 2004.
[24] D. Wessels. dnstop. http://dns.measurement-factory.com/tools/dnstop/, 2002.
[25] D. Wessels. Is Your Caching Resolver Polluting the Internet? In *Proceedings of the ACM SIGCOMM Workshop on Network Troubleshooting: Research, Theory and Operations Practice Meet Malfunctioning Reality*, Portland, OR, August 2004.