An Elemental Decomposition of DNS Name-to-IP Graphs

Alex Anderson*, Aadi Swadipto Mondal*, Paul Barford*, Mark Crovella†, Joel Sommers‡
*University of Wisconsin, Madison, WI USA †Boston University, Boston, MA USA ‡Colgate University, Hamilton, NY USA
aanderson65@wisc.edu, amondal5@wisc.edu, pb@cs.wisc.edu, crovella@cs.bu.edu, jsommers@colgate.edu

***Abstract*—The Domain Name System (DNS) is a critical piece of Internet infrastructure with remarkably complex properties and uses, and accordingly has been extensively studied. In this study we contribute to that body of work by organizing and analyzing records maintained within the DNS as a bipartite graph. We find that relating names and addresses in this way uncovers a surprisingly rich structure. In order to characterize that structure, we introduce a new graph decomposition for DNS name-to-IP mappings, which we term *elemental* decomposition. In particular, we argue that (approximately) decomposing this graph into bicliques — maximally connected components — exposes this rich structure. We utilize large-scale censuses of the DNS to investigate the characteristics of the resulting decomposition, and illustrate how the exposed structure sheds new light on a number of questions about how the DNS is used in practice and suggests several new directions for future research.**

***Index Terms*—Domain Name System, Graph Analysis, Domain parking, IP address parking.**

I. Introduction

The Domain Name System (DNS) is a critical element of the Internet's infrastructure. It serves as the basic translator of names to addresses, but this simple description belies its incredible complexity, and characterizing the DNS is accordingly quite challenging. One challenge in characterizing the DNS is its essentially distributed ownership: a plethora of agents are empowered to place entries into the DNS, and they do so for a wide range of reasons. The distributed ownership and control of the DNS also means that a full accounting of the *contents* of DNS (considered as a database) is quite difficult. Further, the highly dynamic nature of the DNS, and the many uses to which the DNS is put in practice, mean that a complete picture of DNS contents at any time is elusive.

Despite these challenges, a number of efforts have been made to characterize the DNS, in terms of observed traffic, client and server installations, registration patterns and other contents (reviewed in the next section). However, relatively little work has taken a macroscopic view of the contents of the DNS, considered as a graph relating names to addresses (the 'DNS graph'). This is a notable gap in understanding, since the name-to-address graph can be considered to be at the core of the information content of the DNS.

Understanding the DNS graph is important for a number of reasons. First, an understanding of the structure present within the DNS graph is a foundation for building models ((*e.g.,* synthetic workload generators, domain name generators, etc.)) and can improve how we think about the design of DNS systems. Second, the structures present within the DNS graph can shed light on management practices (both of names and addresses) and inform our understanding of how the DNS is configured and used (as we will show below). Finally, considering the DNS as a graph allows us to use graph structures to infer relationships *between* entities. Examples of how relationship inference can be useful include classifying domain names [1] and inferring public cloud use [2]. However, despite these evident benefits, the macroscopic nature of the DNS graph, the structures it contains, and its statistical properties are still largely unexplored.

This paper seeks to fill this gap in understanding in two ways: first, from a characterization standpoint, we describe properties of the DNS graph in terms of the structures present and their statistical properties; and second, from an operations standpoint, we relate those statistical properties to use cases observed in the DNS. In doing so, we are able to quantify known use cases as well as identify previously unappreciated use cases.

To meet these goals, we focus on characterizing a macroscopic view of DNS contents, and we propose both concepts and methods to address this challenge. We work with large-scale censuses of the DNS provided by Rapid7 [3] focusing specifically on *A records*, which map Fully-Qualified Domain Names (FQDNs, *e.g.,* `www.google.com`) to IPv4 addresses. Each census comprises approximately 1.7B records. We note that A records make up the vast majority of data stored in the DNS. Each census defines a bipartite graph with billions of edges.

Starting from this rich collection of datasets, we make four contributions:

1) We provide the first *statistical characterization* of the DNS graph. We show novel statistical properties, including the presence of subgraph sizes that are *simultaneously Zipf* in two different parameters.
2) We describe a method for decomposing the DNS graph into a collection of disjoint subgraphs while retaining most of the edges in the original graph, and show that our method scales to graphs with more than a billion edges. This decomposition of the DNS graph that we obtain is remarkably concise, and as a result we can get a valuable statistical characterization of the DNS graph by focusing on its subgraphs.

3) Using our method, we demonstrate that most of the information in the DNS graph can be captured by describing the DNS graph *as a collection of bicliques* (fully connected bipartite subgraphs). We refer to this as an *elemental decomposition* of the DNS graph. More specifically, we show that by removing only a small fraction of the edges in the DNS graph (approximately 5%–12%), what remains is exclusively a set of bicliques.
4) We show that our decomposition of the DNS graph exposes various use cases for the DNS, and identifies new, previously undocumented use cases.

Specifically: a key finding that emerges from our analysis is that the sizes of DNS bicliques are well-modeled as a bivariate distribution that exhibits Zipf-type scaling properties in both dimensions simultaneously. While Zipf's law has been applied to the DNS in single dimensions in the past, we believe this two-dimensional Zipfian characterization to be new.

We consider DNS elemental decompositions over a three-year timespan and show that the broad characteristics of the DNS graph are relatively stable. At the same time, we show that elemental decomposition can provide new and valuable insight on the nature of churn in the DNS. We find that churn over the course of months and years is most pronounced in instances of a single FQDN that maps to a single IPv4 address, which we connect to activities of cloud and service providers.

We next relate the decomposition to various use cases for the DNS. Within the elemental decomposition of the DNS graph are many diverse kinds of bicliques, reflecting distinct ways of using the DNS — for example, when many names are mapped to a few addresses, or a few names are mapped to many addresses. In fact, at both ends of this spectrum ('many names per address' versus 'many addresses per name') we find characteristic 'parking'-type behaviors. We relate the notion of 'domain parking' to many-name, single-address components in the graph that are managed by registrars. We also identify a new use case that we term 'address parking,' which occurs characteristically within many-address, single-name components. We are unaware of prior studies that have described this phenomenon, which we associate with IP address management practices in large networks.

II. Related Work

Active measurement of the DNS sends queries (possibly from different vantage points) using lists of domain names that are assembled using various techniques (*e.g.,* web crawling). Active measurement is useful for understanding how DNS responds to a wide range of queries but is limited by the coverage of the domain lists. Ongoing active measurement systems for DNS include OpenINTEL [4], RIPE Atlas [5] and Rapid7's Project Sonar [3], which is the source of our data. Alternatively, passive DNS measurement uses instrumentation deployed within the network or at servers to gather data (*e.g.,* [6]). An advantage of passive measurement is that it provides details on how the DNS infrastructure is being used in situ, but it is limited by the need to deploy traffic-capture capabilities in restricted locations. Both active and passive measurements of the DNS have been used to characterize its structure and behavior (*e.g.,* [7], [8]). These studies point to the vast scope of the DNS infrastructure which presents challenges in characterizing its details; one of the challenges that we address.

Also relevant to our work are prior studies on how domain names are registered (*e.g.,* [9]). Many have focused on domain names used for malicious purposes and abuse (*e.g.,* [10]–[13]). Other studies have investigated domain parking (*e.g.,* [14]) and typosquatting (*e.g.,* [15]). These studies provide important background for understanding the characteristics and dynamics of domain names that we analyze in our work.

Prior investigations of content delivery networks (CDN) and cloud providers also inform our work, as DNS plays a central role in determining how users interact with CDN and cloud services [16]–[18]. Building on these studies, our work provides insights on aspects of CDN and cloud infrastructure through analysis of how DNS names are mapped to addresses.

Our work takes a graph-centered view of the DNS. Previous efforts organizing name-to-IP address mappings as an undirected, bipartite graph have been applied mainly in the security domain, *e.g.,* to evaluate DNS agility (in terms of changes in name-to-IP address mappings) [19] and characterize malicious domains [20], and to study the interactions between DNS resolvers [21]. While basic characteristics of the name-to-IP address graph are reported for a limited set of transaction data in [21], we are not aware of any prior studies that examine the graph characteristics of a large-scale DNS census, nor any that propose a model for how to understand structures within the DNS, as we do in this paper.

Our methods rely on *community detection* techniques to break a graph into groups of densely connected nodes [22]–[24]). We use the Louvain method for community detection due to its efficiently and accurately when applied to very large graphs [25]. Finally, our elemental decomposition analysis falls within the scope of general techniques for graph mining and decomposition. Specifically, our focus on decomposition into bicliques is related to the problem of minimum biclique cover, which generalizes the clique cover problem on bipartite graphs [26]. Our method resembles the pivot algorithm in correlation clustering [27], motif-aware graph clustering [28], and near bliclique extraction [29].

III. Data

The primary source of data for our study is Rapid7's Project Sonar [3]. Rapid7 has been curating a running list of domains for several years. They collect domains for their list from several Internet wide scans that they perform on a weekly basis. They perform reverse DNS lookup on the IPv4 space to collect names of hosts. Also, they perform a TCP SYN scan across IPv4 on common ports, and then send an HTTP(S) GET request if a host responds positively to a TCP SYN port associated with a web server. The responses to the HTTP(S) GET requests are scraped for domains and these are added to their list. Additionally, if the server sends a TLS certificate along with the response to the HTTPS GET scan, the certificate is also scraped for any domains. These names are also combined with zone files from various TLDs and gTLDs to form an extensive list of domain and host names.

Rapid7 uses its list of host names and domains collected from scanning activities to actively probe the DNS from three Amazon Web Service (AWS) regions in the US. A DNS ANY request is sent for each name on the list and the records that are returned are then processed to transform the DNS responses into compressed JSON objects. This process is performed monthly by Rapid7 and the data files are made available online [3].

While Rapid7 data includes a variety of DNS record types, *our focus is on A records*, which provide FQDN-to-IPv4 address mappings. This choice recognizes that v4 remains the dominant version of IP in the Internet today. We evaluate a selection of monthly A record data sets to provide longitudinal context for our findings and to assess churn in these mappings.

Considerations regarding the R7 Data. As mentioned previously, Rapid7 performs their DNS scan from three different AWS regions in the US. As identified in prior work (*e.g.* [17]), CDNs and similar services may cause domains to be resolved to different IPs based upon the location of the DNS resolver due to geographical load balancing. As a result, the resolution of domains by Rapid7 will be skewed towards US-based hosts.

To assess the impact of Rapid7's US-only vantage points, we measured variations in name resolution across a set of vantage points having greater geographic diversity. We chose three locations in the US (N. Virginia, N. California, Oregon) and three outside the US (Singapore, Sydney, London). Choosing a set of 5,456,693 FQDNs at random from the Rapid7 data, we resolved each FQDN at each site. We find that 87% (4,729,473) of FQDNs have a consistent mapping across all six vantage points. Further, the number of additional mappings found among the non-US vantage points is 7.1% (388,506). Thus we estimate that data missing due to Rapid7's US-based vantage points is at a level that we believe does not significantly impact our high-level results.

IV. Methods

A. Connected Component Extraction

The first step of our analysis is to process a Rapid7 A Record DNS dataset to construct a set of bipartite graph edges and compute *connected components* (bipartite subgraphs). We first read the raw Rapid7 compressed JSON records, removing and counting any invalid record types (non-A record), invalid (non-globally routable) IPv4 addresses, and any FQDNs that do not conform to DNS RFCs 1034, 1035 and 3696 [30]–[32]. We then externally sort the records and assign a compact identifier to each unique FQDN. Following that, we separate singleton records from non-singletons; a singleton is defined as a FQDN that maps to one IP address, which itself maps to only that FQDN. Connected components are then computed on the non-singleton records, the result of which is a new file containing a name ID, an IP address, and a connected component ID.

The result of the above data processing is a set of connected bipartite subgraphs. Each of these subgraphs can be classified into one of 4 categories:

1:1 A single FQDN and a single IPv4 address (*singletons*).

1:M A single FQDN and more than one IPv4 address. Note that the number of edges is necessarily M.

N:1 More than one FQDN and a single IPv4 address. Again, the number of edges in a N:1 subgraph is necessarily N.

N:M More than one FQDN and more than one IPv4 address. While the previous three elements are already fully connected bicliques, at this stage the N:M elements may or may not be maximally connected.

Code to perform the above steps is written largely in Go,[1] with some components in bash (to coordinate use of UNIX sort, etc.) and Ruby (to coordinate parallel merging of files).

B. Community Detection via Louvain Modularity

Of the four categories of subgraphs extracted in the previous section, the first three are already bicliques. The N:M non-bicliques, *i.e.*, N:M components which are *not* fully connected, are typically quite large and difficult to comprehend without further decomposition. For example, in the data sets we consider, the largest of these N:M non-bicliques has hundreds of millions of FQDNs and edges, and hundreds of thousands of IPv4 addresses. The key to unlocking the structure in these large N:M components is the observation that, by only deleting a small fraction of edges, each N:M component can be decomposed into a *collection of bicliques*.

To obtain this decomposition, we seek to identify the densely-connected subgraphs within each N:M component. This is a hard problem in general, and particularly so at the scale we are working. However, we find that we can obtain good results using a recursive edge-deletion process, in which densely connected subgraphs are identified by the Louvain modularity community detection [25]. Specifically, *modularity* is a measure of the density of connections in a subgraph (community). Our strategy is iterative, alternating between detected dense subgraphs and pruning the *bridge* edges connecting those subgraphs. As we show below, our process only removes a small fraction of edges (5–12%), but exposes the constituent bicliques. Each exposed biclique is a fully connected instance of one of the four subgraph types; we refer to these as *elemental motifs*. Moreover, we find that the method converges very quickly and only requires 4 to 6 iterations of Louvain detection and edge pruning to converge on the data sets discussed in the results. We observe that over 90% of the nodes from N:M non-bicliques are decomposed into elemental motifs in the first iteration of our method.This process allows us to analyze, approximately, the nature of

[1] We use the golang.org/x/net/publicsuffix package for computing suffix and suffix+1 from an FQDN as needed for some processing. We also repurpose some source code from the https://golang.org/src/net/dnsclient.go package to identify invalid FQDNs according to RFCs 1035 and 3696.

the non-biclique components and the operational practices represented in them by using the behaviors inferred from their elemental motifs (constituent bicliques).

V. Results

Our results take three forms: (a) characteristics of the DNS bipartite graph and its connected components; (b) characteristics of the elemental decomposition of the DNS graph; and (c) analyses of the connected components within each class of the decomposition, including the operational relevance and implications of each class.

A. Rapid7 Dataset Characteristics

As described in § IV-A, the first step in processing Rapid7 A record data removes invalid FQDN's and IP addresses. These are most likely caused by aspects of Rapid7's data harvesting process that may indiscriminately include invalid information. *Invalid names* refer to FQDNs that contain invalid characters according to RFC 1034 [30]. *Invalid IPs* refers to IP addresses that are either malformed or that are not globally routable. On manual inspection, we found that many of the invalid IPs were actually FQDNs incorrectly stored as IPs in the Rapid7 record. Also, even though we explicitly use the Rapid7 data sets advertised as containing DNS A records, there are indeed non-A records present in the data. After removal of invalid records, there were about 1.7B records in the 2020 months we considered, with about 1.6B valid records in the 2019 month we considered, and about 1.3B valid records in 2018.

B. Summary Characteristics of Connected Components

Figure 1 shows summary characteristics of connected components for all 5 data sets considered. In the figure, we show log-log complementary CDFs of the number of unique FQDNs (left), the number of unique IPv4 addresses (center) and the number of edges (right) per connected component. We observe that each of these characteristics is heavy-tailed, as evidenced by the linear profile over a range of scales [33]. We also observe a great deal of consistency in connected component characteristics over these five data sets.

In Table I we show the numbers of different types of connected components. We see that the vast majority of connected components are fully-connected bicliques and that a large fraction of these are 1:1 (single-name, single-address) components. We also see that there are more than 100x N:1 (multi-name) components than 1:M (multi-address) components. Moreover, the table shows that of the N:M components, most ($\approx 95\%$) are not fully connected.

To present a distributional view of the connected components, we use the representation shown in Figure 2. This is one of the primary representations we use to characterize the DNS A record graph in terms of its structures, here and below. It is a two-dimensional histogram, in which position (x, y) represents the components having x addresses and y names. The size of the marker at each point encodes the count of how many components are present of that size. Note that there is a log scale on the x and y axes, and also on marker sizes.

Like Table I, this figure depicts the DNS graph components prior to decomposition. In this plot (and in plots below), bicliques are shown in blue, while non-bicliques are shown in red. Hence all 1:1 (singletons) are at the origin, 1:M bicliques (multiple IP addresses connected to a single FQDN) lie along the x axis at $\log(y) = 0$, and that N:1 bicliques (a single IP address connected to multiple FQDNs) lie along the y axis at $\log(x) = 0$. We see that while there are many fully-connected N:M components, there are also a large number of not-fully-connected (non-biclique) N:M components.

Of the non-biclique (red) points in Figure 2 we see some interesting features. First, there are a high number of components along the line x=y indicating that a common operational use of the DNS is to have roughly similar numbers of names and addresses that map to one another. We also observe a vertical line around x=256 indicating that a common operational configuration is to map multiple names to a subset of addresses within an IPv4 /24 prefix. We also observe some clustering around other powers of 2, *e.g.*, 16, indicating a more general operational pattern of assigning a collection of names to a subset of addresses within a well-defined IPv4 prefix. Finally, as previously noted, there are many large non-bicliques — some with hundreds to thousands of IPv4 addresses and with thousands to *millions* of FQDNs. The largest (in the upper-right corner) has 270,553,077 FQDNs, 235,437 IPv4 addresses, and 344,707,222 edges.

C. Characterization of the Elemental Decomposition

Although Table I shows that there are relatively few non-biclique N:M connected components, Figure 2 indicates that these components can be *extremely* large. To gain insight and to better understand the operational practices that lead to components like these, we use our elemental decomposition methods as described in Section IV-B.

Table II provides a high level view of how edges are pruned when decomposing the DNS A record graph into smaller communities (*i.e.*, mostly bicliques). The table shows the fraction of edges that are in bicliques before decomposition, the fraction of edges cut during decomposition, and the total fraction of edges that are associated with a biclique after decomposition. The table shows that the decomposition is accomplished by removing a relatively small number of edges from the original bipartite graph, leaving most of the original graph intact. We see in the table that the graph can be decomposed by removing 5–12% of its edges. These results show that for most edges in the original graph (88–95%), the edge can be associated with a set of names and addresses in which each name is connected to each address.

Table III shows the number of edges and bicliques in our data *after* elemental decomposition over a three year period and for the three months we consider in 2020. It shows that in some cases the high level statistics of the decomposition are relatively stable over both long and shorter time periods (*e.g.,* singletons and 1:M), while in other cases changes are more substantial (*e.g.,* increases in both edges and bicliques for N:1s and a fairly large decrease in edges for N:Ms between '19 and '20).

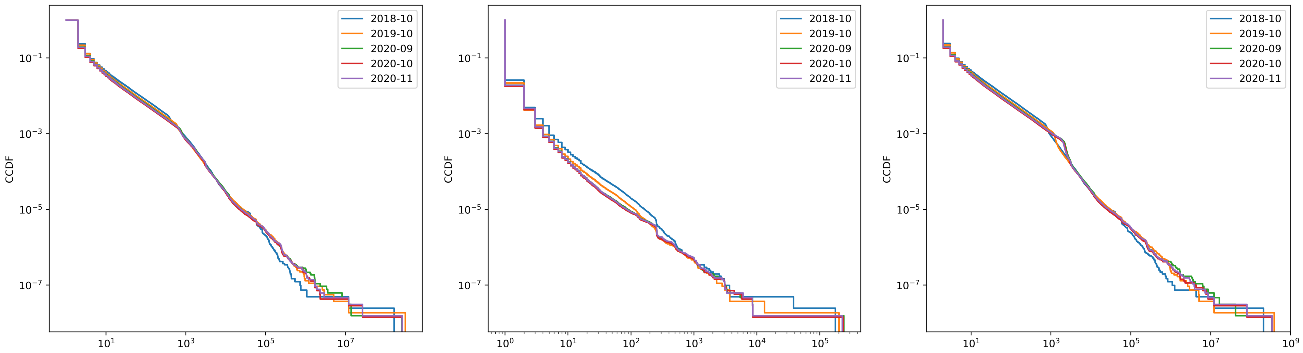

Figure 1. Log-log complementary CDFs of the number of names (left), IP addresses (center), and edges (right) per connected component.

Table I
NUMBER OF CONNECTED COMPONENT MOTIF TYPES FOR CONNECTED COMPONENTS COMPUTED FROM EACH RAPID7 DNS A RECORD DATASETS.

Dataset	1:1	1:M	N:1	N:M bicliques	N:M non-bicliques
2018-10	772,549,944	371,845	39,630,092	68,666	617,368
2019-10	795,116,436	336,966	52,602,684	72,066	764,626
2020-09	762,495,293	348,915	63,140,326	56,893	822,018
2020-10	735,682,036	350,263	68,794,603	56,966	833,439
2020-11	727,606,285	347,255	63,459,342	56,223	823,058
2021-04	705,458,448	405,504	64,033,969	55,769	861,164

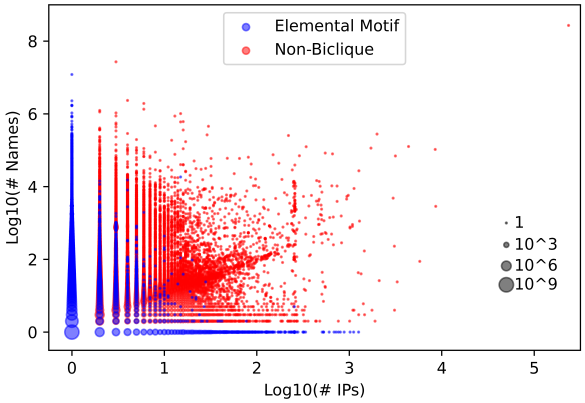

Figure 2. Connected Component Distribution Oct. 2020 (Pre-Decomposition)

Table II
EDGES REMOVED BY DECOMPOSITION

Year	% Clique Component Edges	% Edges Cut	% Clique Edges
2018	79.93%	4.59%	95.41%
2019	71.06%	6.34%	93.66%
2020	66.64%	11.91%	88.09%

The statistics in Table III do not expose the extreme heterogeneity of sizes found among the elemental motifs. To illustrate that, in Figure 3 we show histogram characterizations of the complete decompositions of the 2018, 2019, and 2020 data sets. Each of the plots in this figure represents a high-level characterization of the entire DNS A record graph. Each plot can be seen as characterizing the data in terms of a distribution of elemental motif sizes. This is a bivariate distribution in N and M. The figure confirms the Zipfian laws that apply to names and addresses individually, and further suggests that those laws extend to the N and M parameters of all the elemental motifs.

We provide further evidence that the distribution of elemental motif sizes is a bivariate Zipf (or power-law) distribution in Figure 4. This figure shows value-conditional 'slices' through the Figure 3 histogram for the 2020 dataset. Each straight line on log-log scale shows that for a particular value of M (resp. N) the density conditioned on that value is Zipfian (power-law). There are some notable deviations from overall power-law behavior: the distributions tend to deviate when $N = 1$ or $M = 1$, suggesting that N:1 and 1:M motifs are special cases; and there is a significant 'bump' around 200:1 to 400:1 (left hand plot) which suggests that mapping a set of around 200-400 names to a single address is another special case. However we conclude that our results suggest that the collection of A records in the DNS can be thought of — *approximately* — as a collection of N:M bicliques in which the N and M values form a two-dimensional Zipf distribution. This Zipfian structure has implications for DNS operations and workload generation, which we plan to investigate in future work.

Figure 5 shows summary characteristics of connected and decomposed components, similar to the plot shown in Figure 1 for connected components prior to decomposition. In the figure, we show log-log complementary CDFs of the number of unique FQDNs (left), the number of unique IPv4 addresses (center) and the number of edges (right) per connected component. We observe in these plots that even after decomposing

Table III
SUMMARY OF ELEMENTAL DECOMPOSITION OVER 3 YEAR PERIOD.

Month,Year	Type	Singletons	N:1	1:M	N:M	Bridge
October 2018	Edge	773,370,621	506,204,126	1,346,050	13,166,357	62,323,786
October 2018	Bicliques	773,370,621	40,463,017	412,749	70,772	N/A
October 2019	Edge	795,975,739	760,988,260	1,120,212	19,396,861	106,782,775
October 2019	Bicliques	795,975,739	53,800,420	368,146	74,231	N/A
September 2020	Edge	763,432,351	805,149,052	1,109,375	2,120,227	197,810,698
September 2020	Bicliques	763,432,351	64,528,682	381,296	59,733	N/A
October 2020	Edge	736,633,229	829,071,322	1,114,219	2,395,808	212,225,000
October 2020	Bicliques	736,633,229	70,197,556	382,890	59,619	N/A
November 2020	Edge	728,539,217	802,998,590	1,103,697	2,144,884	212,982,170
November 2020	Bicliques	728,539,217	64,844,076	379,653	59,151	N/A
April 2021	Edge	705,458,448	803,848,303	1,163,107	2,119,340	199,631,058
April 2021	Bicliques	705,458,448	64,033,969	405,504	58,379	N/A

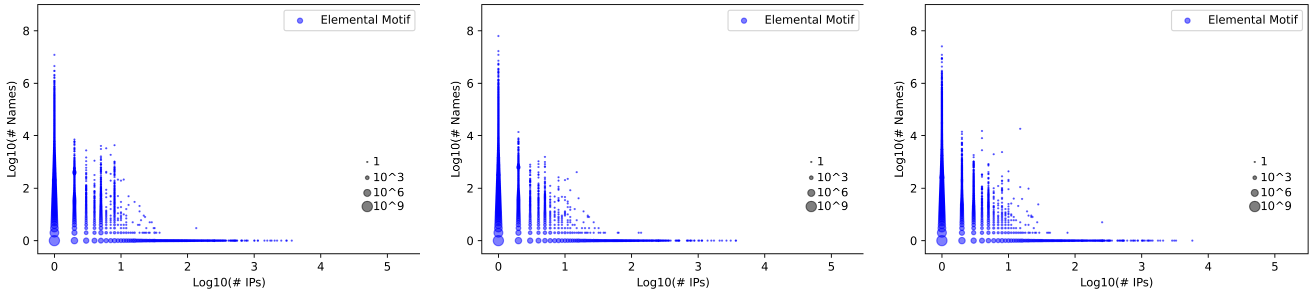

Figure 3. Fully decomposed motifs for 2018 (left), 2019 (center), and 2020 (right) data sets.

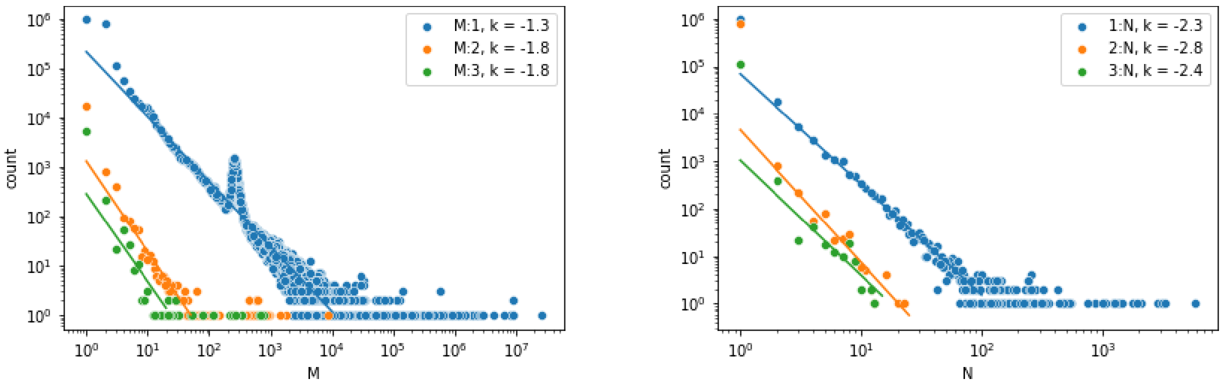

Figure 4. 2020 Decomposition: Bivariate Zipf distribution of $M : N$ motifs for fixed values of M and N. Listed k values are estimated power-law exponents from shown fitted lines.

non-biclique components into elemental motifs we continue to see heavy-tailed characteristics as evidenced by the linear profile over a range of scales. We again observe a great deal of consistency over the three years considered. These results imply that the communities resulting from decomposition have similar features as the aggregate collection of components from the original graph.

D. Dynamics of the Elemental Decomposition

Each of our results so far gives a picture of the A records in the DNS at a single point in time. In order to understand the state of the DNS more thoroughly, we now characterize its dynamic evolution from the perspective of elemental motifs. We study changes occurring between the three months of September, October, and November 2020. Note that, as discussed in Section III, while the Rapid7 measurement process is not fully known, the set of FQDNs queried generally grows from month to month. So the dynamics we observe could be biased by the measurement process, but we expect this bias to be small and that the results should be primarily driven by actual changes in the DNS.

We characterize the evolution of the DNS by accounting for

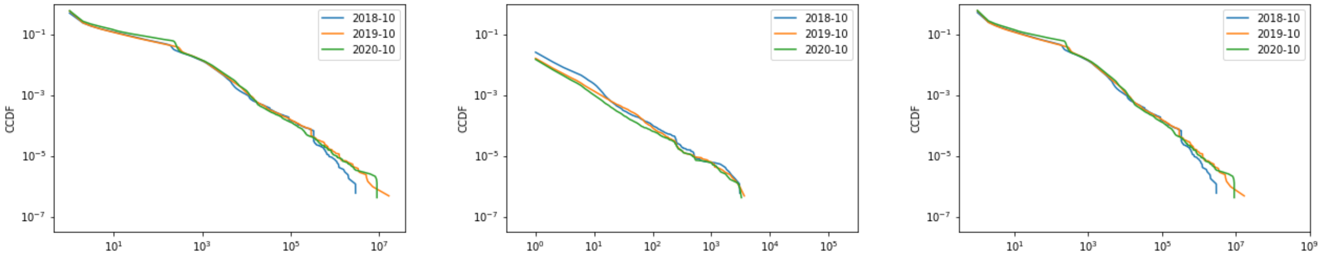

Figure 5. Log-log complementary CDFs of the number of names (left), IP addresses (center), and edges (right) per community from decomposition.

edges (A records), looking at how edges move between motif types. To make the accounting complete, we need to include two additional categories of edges: those that appear/disappear in any given snapshot, and bridge nodes (those that are pruned by our methods, and so do not participate in elemental motifs).

We present the results of this analysis in the form of a Sankey plot of monthly dynamics in Figure 6. The figure illustrates a number of insights into the dynamics of DNS A record mappings. First, we note there are about as many edges in N:1 components as in 1:1 components. Hence, the use of DNS mappings to map multiple names to a single host is quite common. This is in contrast to the relatively small number of edges in 1:M components. This disparity can be understood in light of the fact that names (FQDNs) are an unlimited resource while addresses are drawn from a finite set.

With respect to dynamics of A records, we observe that there is a high rate of change overall. However change is not equally distributed across motif categories. The highest amount of change, both appearance and disappearance, takes place among singletons; the second most dynamic category are the bridge edges. On a monthly basis, edges in N:1 components are overall much more stable than those in singletons, suggesting a different role for N:1s compared to singletons (which we elaborate below). For instance, with the monthly dynamics we note that about 1/3 of the edges that disappear in October re-appear in November (*i.e.,* are churning). Again, these are almost exclusively singleton edges. Finally, we note that month-to-month movement of edges between the bridge category and motifs is relatively rare, suggesting our decomposition is fairly stable over time.

E. Analysis of Elemental Motifs

We now turn to examining characteristics of the elemental motifs (bicliques) and the role that they play in the DNS.

1) Singletons: Elemental motifs that map a single name to a single address (1:1) comprise a very large fraction of all elemental motifs, as discussed above and shown in Table I. The set of singletons includes all hosts for which a single name has been assigned for management and access, *e.g.*, for remote login, web hosting, infrastructural purposes, etc. As seen in Table III, nearly half of all A records in our data fall into this category.

On closer examination and somewhat surprisingly, it appears that singleton mappings are used primarily for network infrastructure management. For example, the suffixes that appear most often among the set of singletons correspond to Amazon AWS, Comcast, SBC, Verizon, and Roadrunner (in that order). Inspection of examples of the FQDNs contained in singleton mappings from these providers shows that most names map to infrastructure interfaces and customer premises devices. This is evident because in most cases the location or role of the device or interface is encoded in some way in the FQDN [34]. The largest single use of singleton mappings by providers concerns what appear to be customer premises equipment for broadband access.

These observations add nuance to an understanding of the modern role of the DNS. While the traditionally stated role of the DNS is to assign names to network resources [30], [31], *e.g.*, assigning names to addresses, it appears that currently, one of the most common uses of those DNS mappings is in infrastructure management.

2) N:1 Elemental Motifs and Domain Parking: Next, we consider elemental motifs consisting of more than one name mapped to a single address (N:1). As shown in Table III, starting in 2020, this set of motifs contains more A records than the singletons; however, as also shown in Table III, there are less than 1/10 as many N:1 bicliques as singletons.

Table IV shows the top 5 Autonomous Systems and associated organizations for N:1 bicliques. The table is dominated by access networks and cloud providers. Beyond the top 5 are many additional canonical web- and virtual machine-hosting organizations within the top 50, *e.g.*, Digital Ocean, Linode, Amazon, Google, etc. The N:1 motif is the most commonly occurring operational arrangement within the DNS, as it provides the commonly-used name aliasing capability of DNS. Moreover, as we now discuss, it is used for *domain parking*, where we observe occurrences of millions of names mapped to a single address.

When examining the largest N:1 communities we find evidence that our elemental decomposition exposes a collection of parked domains. A parked domain is defined as a name that is registered in the DNS but not hosting an actively maintained web site. In the meantime it is typically used to generate advertising revenue via a parking service [14]. To identify whether an N:1 component represents a set of parked domains, we compare the number of FQDNs in a biclique with the number of unique suffix+1s. The intuition for this approach is

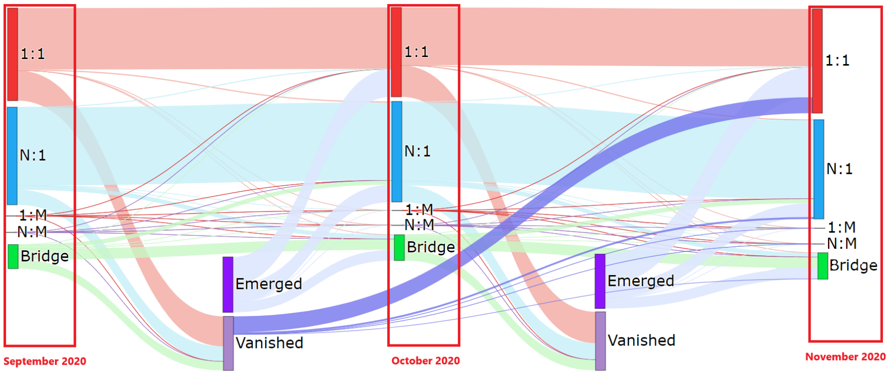

Figure 6. Monthly Dynamics by Elemental Motif. Each red box in the figure corresponds to a single census comprising about 1.8B edges (A records).

Table IV
Top 5 most frequently occurring ASes for N:1 motifs, October 2020.

ASN	Org	Motif Count
20115	CHARTER-20115, US	8,847,365
3462	HINET Data Comm. Group, TW	8,807,258
5089	NTL, GB	6,120,802
6805	TDDE-ASN1, DE	4,858,137
16509	AMAZON-02, US	1,855,579

that we expect a set of domains parked on a single IP address to exhibit a wide variety of suffix+1s. This is in contrast to N:1 bicliques that may be used to provide several aliases for the *same host* (and which may also be used in the context of IP Anycast), which we argue are more likely to exhibit some commonality in suffix+1. Indeed, we see many N:1 bicliques that follow this pattern. As an example, the largest N:1 component contained 25,492,695 unique FQDNs mapped to 34.102.136.180 which is an IP address owned by Google. The FQDNs represent 25,489,294 unique suffix+1s (99.98% of the total). The fact that nearly all FQDNs are also unique suffix+1s suggests that these names represent a large number of disparate organizational entities. By manual inspection of a sample of these domains, we confirmed that they are parked domains managed by GoDaddy.

To support our domain parking hypothesis, we adapt a technique identified by Gowda *et al.* [35]. The intuition with this approach is that web pages for parked domains often exhibit a canonical structure, *e.g.*, an "under construction" banner or indication that the domain is for sale, etc. Our approach focuses on the DOM hierarchy and extracting a descriptor from it based on computing hashes of any `style`, `script`, or `link` tags within a page, along with a hash of the DOM structure itself. We consider two pages to have the same underlying structure if they have identical descriptors.

Considering the same large N:1 component, we performed HTTP GET requests to a random selection of its FQDNs until the ratio of unique descriptors found to the total number of samples is less than 0.01. In total, 3,989 FQDNs were sampled, 2,439 of which were discarded mainly due to expired domains (a small number were network timeouts). Of the remaining 1,550 FQDNs that were successfully accessed and analyzed, 1,450 had the same descriptor. These domains used the same parked domain template from GoDaddy. While these results are limited, we view them as providing strong support that this large N:1 biclique is involved in domain parking.

3) 1:M Elemental Motifs and Address Parking: Next, we consider elemental motifs consisting of one name mapped to multiple IPv4 addresses (1:M). Our first observation is that in the vast majority of cases, the addresses in a 1:M bicliques are all owned by the same organization. Using WHOIS Regional Internet Registry (RIR) data from team-cymru.com, we examined the AS ownership of the prefixes in 1:M bicliques. We find that 99.8% of 1:M bicliques contain addresses that all lie within the same AS.

Next, in Table V we show the top 5 most frequently occurring suffix+1's for the one FQDN occurring in 1:M components. For these components, in which a single name is mapped to more than one IPv4 address, the naming patterns used provide hints as to how different organizations may use their allocated address space and manage hosts in their networks. For `cloudflare.net`, for example, the names follow a pattern of including a customer hostname with `cdn.cloudflare.net` suffix, such as `www.tieto.cz.cdn.cloudflare.net` or `www.levi.jp.cdn.cloudflare.net`. Since Cloudflare is known to use IP Anycast to perform load balancing [36], it is unsurprising that we observe a reasonably large number of 1:M elemental components with a suffix+1 of `cloudflare.net` . As another example, we observe more than 6,000 components with `outlook.com` suffix+1's, suggesting that 1:M arrangements are commonly used for load-balancing mail relay servers.

For other suffixes, the 1:M pattern appears to be used in order to treat multiple addresses as a single unit for infrastructure management purposes. For example, with `urlatt.net`, many of the names refer to aggregation devices for network access, *e.g.*, `200.chicago-09rh15-16rt.il.dial-access.att.net`. We hypothesize that in many of these infrastructural-related 1:M motifs, the IP addresses are assigned to M interfaces of a single device, *e.g.*, a router, and

Table V
Top 5 most frequently occurring suffix+1s for 1:M motifs, October 2020.

Suffix+1	Motif Count
awsglobalaccelerator.com	22,811
alcatel.com	7,053
cloudflare.net	6,838
outlook.com	6,818
att.net	6,693

Table VI
Top 5 classifications of 1:M components using Enemieslist.

Classification	Count	Addresses
static (infrastructure)	8202	84398
dynamic (e.g., DHCP-assigned)	2534	19992
legitimate mail source	1771	16644
unassigned	30	5067
NAT/proxy	267	2506

the single name is used to refer to the device as a whole. We observe similar patterns and naming conventions for other providers, *e.g.*, Telstra; these findings are consistent with prior work [34], but provide some additional insight due to the observed relationship between these names and IPv4 addresses through the DNS.

Address Parking. Analogous to the notion of a parked domain, we observe 1:M components in which the name suggests that the associated IP addresses are not in use and not currently assigned. For example, we observe FQDNs with words like "reserved", *e.g.*, `reserved.102net.gantep.edu.tr` or "unused", *e.g.*, `unused.vmb-rostov.ru`. We refer to the practice of associating one or more IP addresses with a name that indicates that those addresses are not in use as *Address Parking*; we are unaware of other research that has previously described this phenomenon. We hypothesize that Address Parking is a manifestation of a particular style or method of IP address management (IPAM). IPAM systems are prevalent in large organizations and typically work hand-in-hand with management and configuration of a DNS zone database to maintain consistency between addresses and how/whether they are in use, as well as names with which they are associated.

To gain some insight into components with apparent parked IP addresses as well as other 1:M components, we used the Enemieslist domain classification service, which classifies FQDNs for policy-related purposes [37] such as spam filtering. We focus on 1:M components with 4 or more IP addresses associated with a single name. Table VI shows the top 5 classifications by number of addresses. We found that most components have names that are associated with infrastructural-type roles such as routers, mail servers, DHCP pools, NAT devices, cloud computing servers, etc. The "unassigned" category, with over 5,000 addresses in 30 components, refers to parked addresses which are currently unused.

4) N:M Elemental Motifs: Finally, the N:M elemental components exist as collections of name aliases mapped to a pool of IP addresses. As with other elemental motifs, due to these components being fully-connected bicliques the names and addresses are *functionally equivalent*. These components are most often associated with cloud infrastructure providers like Cloudflare and Amazon Web Services which typically provide a set of aliased FQDNs to their customers for accessing a service that is load-balanced across a set of IP addresses. While this finding should not be surprising — it is common operational practice — the scale of some components we observe (millions of names, thousands of IP addresses; *cf.* Figure 3) has not been previously reported. For example, we observe one component with 24M FQDNs mapped over 3 Cloudflare-owned addresses, and another component with 88 FQDNs and 5790 IP addresses, where the suffix+1 is `amazonaws.com`.

VI. Conclusion

In this paper we have presented a first look at the graph defined by the IPv4 mappings of the DNS, macroscopically and in total. Although some prior studies have looked at small samples of the DNS contents as a graph, to the best of our knowledge ours is the first large-scale study that attempts to examine a large portion of the DNS contents (A records) as a graph. We showed that a useful way to decompose, and therefore think about the DNS graph is as a collection of bicliques, connected by a relatively small subset of non-biclique edges. This suggests that bicliques are an important organizing principle for the DNS. We identify four different types of bicliques found in the DNS graph that we define as *elemental*. We present high-level depictions in the form of statistical characterizations of the DNS A record graph using this elemental decomposition. We show that beyond Zipf laws for names and addresses, we can think of the A record name-to-IP mappings as a collection of bicliques whose sizes are drawn from a two-dimensional Zipf-type distribution. And finally, we examine details of the elemental motifs including their churn over time and illustrate diferences in elemental motif function within the DNS ecosystem.

Our work suggests a number of directions for future study. First, we plan to expand our graph-based analysis of DNS contents to include other record types, *e.g.*, AAAA and CNAME records, and higher levels of domain name aggregation. Next, we plan to investigate how elemental motifs can reveal malicious activity and inform security monitoring. Finally, we will consider how operational practices related to DNS record management (including IPAMs) can be improved using the elemental motif perspective.

Acknowledgements

The authors thank the Rapid7 Research Team and Steve Champeon from Enemieslist for their support on this project. This material is based upon work supported by the National Science Foundation under Grant No. CNS-2106517, CNS-2312709 and CNS-2319367. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] I. Khalil, T. Yu, and B. Guan, "Discovering Malicious Domains through Passive DNS Data Graph Analysis," in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security (ASIA CCS)*, 2016.

[2] Q. Jacquemart, C. Pigout, and G. Urvoy-Keller, "Inferring the Deployment of Top Domains over Public Clouds using DNS Data," in *Proceedings of the Network Traffic Measurement and Analysis Conference (TMA)*, 2019.

[3] "Rapid7 Project Sonar," https://www.rapid7.com/research/project-sonar/, 2024.

[4] "OpenINTEL," https://openintel.nl/, 2024.

[5] "RIPE Atlas DNSMON," https://atlas.ripe.net/dnsmon/, 2021.

[6] "DomainTools: Farsight DNSDB," https://www.domaintools.com/products/farsight-dnsdb/, 2024.

[7] P. Foremski, O. Gasser, and G. Moura, "DNS Observatory: The Big Picture of the DNS," in *Proceedings of the ACM Internet Measurement Conference*, 2019.

[8] K. Schomp, T. Callahan, M. Rabinovich, and M. Allman, "On Measuring the Client-Side DNS Infrastructure," in *Proceedings of the ACM Internet Measurement Conference*, 2013.

[9] T. Halvorson, M. Der, I. Foster, S. Savage, L. Saul, and G. Voelker, "From .academy to .zone: An Analysis of the New TLD Land Rush," in *Proceedings of the ACM Internet Measurement Conference*, 2015.

[10] M. Korczynski, M. Wullink, S. Tajalizadehkhoob, G. Moura, A. Noroozian, D. Bagley, and C. Hesselman, "Cybercrime After the Sunrise: A Statistical Analysis of DNS Abuse in New gTLDs," in *Proceedings of the Asia Conference on Computer and Communications*, 2018.

[11] Y. Zhauniarovich, I. Khalil, T. Yu, and M. Dacier, "A Survey on Malicious Domains Detection through DNS Data Analysis," *ACM Computing Surveys*, vol. 51, no. 4, 2018.

[12] S. Yadav, A. Reddy, N. Reddy, and S. Ranjan, "Detecting Algorithmically Generated Malicious Domain Names," in *Proceedings of the ACM Internet Measurement Conference*, 2010.

[13] M. Mowbray and J. Hagen, "Finding Domain-Generation Algorithms by Looking at Length Distribution," in *Proceedings of the IEEE International Symposium on Software Reliability Engineering Workshop*, 2014.

[14] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking Sensors: Analyzing and Detecting Parked Domains," in *Proceedings of the Network and Distributed System Security Symposium*, 2015.

[15] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, "Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse," in *Proceedings of the Network and Distributed System Security Symposium*, 2015.

[16] J. Otto, M. Sanchez, J. Rula, and F. Bustamante, "Content Delivery and the Natural Evolution of DNS: Remote DNS Trends, Performance Issues and Alternative Solutions," in *Proceedings of the ACM Internet Measurement Conference*, 2010.

[17] W. Scott, T. Anderson, T. Kohno, and A. Krishnamurthy, "Satellite: Joint Analysis of CDNs and Network-Level Interference," in *Proceedings of the USENIX Annual Technical Conference*, 2016.

[18] G. Moura, S. Castro, W. Hardaker, M. Wullink, and C. Hesselman, "Clouding up the Internet:How Centralized is DNS Traffic Becoming?" in *Proceedings of the ACM Internet Measurement Conference*, 2020.

[19] J. Ruohonen and V. Leppänen, "Investigating the Agility Bias in DNS Graph Mining," in *Proceedings of the IEEE International Conference on Computer and Information Technology (CIT)*, 2017.

[20] M. Kuhrer, C. Rossow, and T. Holz, "Paint it Black: Evaluating the Effectiveness of Malware Blacklists," in *Proceedings of the 17th International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, 2014.

[21] L. Deri, S. Mainardi, M. Martinelli, and E. Gregori, "Graph Theoretical Models of DNS Traffic," in *Proceedings of 9th International Wireless Communications and Mobile Computing Conference*, 2013.

[22] D. Spielman and S. Tang, "Nearly-linear Time Algorithms for Graph Partitioning, Graph Sparsification, and Solving Linear Systems," in *Proceedings of ACM Symposium on Theory of Computing*, 2004.

[23] F. Murtagh, "A Survey of Recent Advances in Hierarchical Clustering Algorithms," *Computer Journal*, vol. 26, no. 4, 1983.

[24] U. von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, no. 4, 2007.

[25] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10, 2008.

[26] J. Amilhastre, M. Vilarem, and P. Janssen, "Complexity of Minimum Biclique Cover and Minimum Biclique Decomposition for Bipartite Domino-free Graphs," *Discrete Applied Mathematics*, vol. 86, no. 2-3, 1998.

[27] F. Bonchi, D. Garcia-Soriano, and E. Liberty, "Correlation Clustering: from Theory to Practice," in *Proceedings of ACM SIGKDD Conference*, 2014.

[28] C. Tsourakakis, J. Pachocki, and M. Mitzenmacher, "Scalable Motif-aware Graph Clustering," in *Proceedings of the WWW Conference*, 2017.

[29] M. Mitzenmacher, J. Pachocki, R. Peng, C. Tsourakakis, and S.-C. Xu, "Scalable Large Near-Clique Detection in Large-Scale Networks via Sampling," in *Proceedings of ACM SIGKDD Conference*, 2015.

[30] P. Mockapetris, "RFC 1034: Domain Names—Concepts and Facilities," https://datatracker.ietf.org/doc/html/rfc1034, November 1987.

[31] ——, "RFC 1035: Domain Names—Implementation and Specification," https://datatracker.ietf.org/doc/html/rfc1035, November 1987.

[32] J. Klensin, "RFC 3696: Application Techniques for Checking and Transformation of Names," https://datatracker.ietf.org/doc/html/rfc3696, February 2004.

[33] M. E. Crovella, M. S. Taqqu, and A. Bestavros, "Heavy-Tailed Probability Distributions in the World Wide Web," in *A Practical Guide To Heavy Tails*, R. J. Adler, R. E. Feldman, and M. S. Taqqu, Eds. New York: Chapman and Hall, 1998.

[34] J. Chabarek and P. Barford, "What's in a name? Decoding Router Interface Names," in *Proceedings of the 5th ACM HotPlanet Workshop*, 2013, pp. 3–8.

[35] T. Gowda and C. A. Mattmann, "Clustering Web Pages Based on Structure and Style Similarity (Application Paper)," in *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, 2016.

[36] M. Prince, "Load Balancing without Load Balancers," https://blog.cloudflare.com/cloudflares-architecture-eliminating-single-p/, June 2013.

[37] "Enemieslist: Security and Antispam," https://http://enemieslist.com/, 2024.