From .academy to .zone: An Analysis of the New TLD Land Rush

Tristan Halvorson
trhalvorson@cs.ucsd.edu

Matthew F. Der
mfder@cs.ucsd.edu

Ian Foster
idfoster@cs.ucsd.edu

Stefan Savage
savage@cs.ucsd.edu

Lawrence K. Saul
saul@cs.ucsd.edu

Geoffrey M. Voelker
voelker@cs.ucsd.edu

Department of Computer Science and Engineering
University of California, San Diego

ABSTRACT

The `com`, `net`, and `org` TLDs contain roughly 150 million registered domains, and domain registrants often have a difficult time finding a desirable and available name. In 2013, ICANN began delegation of a new wave of TLDs into the Domain Name System with the goal of improving meaningful name choice for registrants. The new rollout resulted in over 500 new TLDs in the first 18 months, nearly tripling the number of TLDs. Previous rollouts of small numbers of new TLDs have resulted in a burst of defensive registrations as companies aggressively defend their trademarks to avoid consumer confusion. This paper analyzes the types of domain registrations in the new TLDs to determine registrant behavior in the brave new world of naming abundance. We also examine the cost structures and monetization models for the new TLDs to identify which registries are profitable. We gather DNS, Web, and WHOIS data for each new domain, and combine this with cost structure data from ICANN, the registries, and domain registrars to estimate the total cost of the new TLD program. We find that only 15% of domains in the new TLDs show characteristics consistent with primary registrations, while the rest are promotional, speculative, or defensive in nature; indeed, 16% of domains with NS records do not even resolve yet, and 32% are parked. Our financial analysis suggests only half of the registries have earned enough to cover their application fees, and 10% of current registries likely never will solely from registration revenue.

Categories and Subject Descriptors

C.2.m [**Computer Communication Networks**]: Miscellaneous; K.4.1 [**Computers and Society**]: Public Policy Issues

Keywords

Domain Name System; Top-Level Domains; Registration Intent; Internet Economics

IMC'15, October 28–30, 2015, Tokyo, Japan.

http://dx.doi.org/10.1145/2815675.2815696.

1. INTRODUCTION

Nearly any successful company today needs a good Internet presence, and most see a memorable domain name as a key part of that presence. Though a nearly infinite set of possible domain names exist, any given name is unique, and memorable names often change hands for thousands or even millions of dollars. The Domain Name System (DNS) originally included only a small handful of top-level domains (TLDs), and ICANN has kept that number low until recent times. The benefits of a new TLD seem obvious at first glance: simple and memorable strings, long since taken in the older TLDs, become available again under a new namespace. However, many registrations in new TLDs go towards defensive registrations, brand or trademark owners trying to protect their names.

Starting in 2013, delegation began of a whole new wave of TLDs. Whereas ICANN debated the inclusion of previous TLDs independently and over the course of multiple ICANN board meetings, TLDs in the new program go through a standard application process which does not include ICANN-wide attention. The new expansion has resulted in a swift expansion of the TLD namespace: on October 1, 2013, shortly before the beginning of the program, the root zone contained 318 TLDs, mostly country code TLDs (ccTLDs). As of April 15, 2015, the root zone contained 897 TLDs, an expansion of 579 TLDs in less than two years.

This paper identifies the impact of the New gTLD Program on the domain name ecosystem. Previous TLD additions, such as `biz` [12] and `xxx` [11], caused widespread speculation and defensive registrations, but this larger expansion could discourage both. With hundreds of new TLDs, we expect many smaller companies to find it infeasible to defend their name in each. Additionally, such a sharp increase in simple-word second-level domains could make it difficult for speculators to resell even desirable names. The program's success also depends on how Internet users view the new domains. Do consumers see TLDs as interchangeable, or will new TLDs discourage users from visiting the associated domain? To answer these questions, we make the following contributions:

- We **classify registration intent** with a methodology derived from our work on `xxx` [11]. The application of the methodology to the New gTLD Program presented additional scaling difficulties, most notably requiring further automation of domain analysis. Our main contribution is the timely result of this methodology applied to the TLD landscape during its current period of swift expansion.

- We determine the program's **impact on the old TLDs**, both on registration rates and on the types of registrations.
- We examine **registry profitability** to learn where the registration money goes and what kinds of TLDs get the most registrations.

Taken together, our findings suggest that the new gTLDs have yet to provide value to the Internet community in the same way as legacy TLDs. Although the new TLDs greatly expand the domain name space, overall we find that speculative and defensive registrations dominate the growth of registrations in new TLDs. For domains that resolve with some kind of content, over 45% are speculative in nature, nearly 40% are defensive, and less than 15% host primary Web content. Users also visit new domains in the new TLDs less frequently than in the old, and new TLD domains are more than twice as likely to appear on a blacklist within the first month of registration. Finally, we also find that the new TLDs have yet to have significant impact on the old TLDs. Registrations in the new TLDs generally increase the number of total registrations, and `com` continues to dominate Internet domain name registration activity overall.

2. BACKGROUND

The Domain Name System (DNS) is the Internet service that maps human-readable names to machine addresses. In the Internet today, the DNS is overseen by the Internet Corporation for Assigned Names and Numbers (ICANN), which holds the authority for establishing new top-level domains (TLDs). After a number of minor TLD additions in the decade previous (e.g., `biz`, `info`, `mobi`, `xxx`), ICANN initiated a new process in 2008 to normalize the policies for creating new gTLDs. In late 2013, the first new gTLD was delegated. The length of this process reflects the significant complexity involved, the range of stakeholders and the significant potential for conflict. In this section, we provide background on how the domain name ecosystem works, how companies apply for a TLD in the new program, and who the significant players are.

There are three key actors in the DNS ecosystem:

Registries operate TLDs and have a contract with ICANN for each one.

Registrars sell domain names, typically in many different TLDs, and also have an ICANN accreditation.

Registrants are entities that buy domain names.

and our goal in this paper is to explore how these actors have reacted to the rapid expansion of the DNS name space.

2.1 The Delegation Process

In preparation for the expansion, ICANN formalized a detailed application process for those seeking to sponsor new TLDs (well over 300 pages in English) [19]. Each applicant prepared an extensive submission covering business, technical and operational issues and paid a USD 185,000 evaluation fee for the initial evaluation. These applications in turn were open to public comment and for review by government interests and interested stakeholders. In such cases, the TLD might undergo extended evaluation, dispute resolution, or a contention period when multiple applications pursue the same TLD (and in such cases the fees could increase considerably). With the addition of legal fees, drafting fees, data escrow fees, auctions for contested names and operational costs, applications for a new gTLD require significant capital and thus favor those large organizations (e.g., Google, Amazon, Donuts) who would amortize these expenses across many such applications.

Those applicants whose submission survived evaluation transitioned to a phase called "delegation" (when the TLD is entered into the zones of the root DNS servers) subject to a series of contractual obligations (e.g., a registry agreement with ICANN covering dispute resolution, fees, technical standards, etc.) and technical tests. Delegation marks the time when end users can first resolve domains under the new TLD and is thus a major milestone for any registry. Due to capacity constraints inside ICANN and changes in applicant business goals, there can be considerable delay between evaluation and delegation.

2.2 TLD Rollout

After delegation, the TLD life cycle depends on the registry. TLDs intended for public use have a sunrise phase, a period of time during which only trademark holders may register. This phase gives brand holders a first chance to defend their names. Most TLDs follow with a "land rush" phase where registrants can get an earlier chance at any domain name for a price premium, usually on the order of a few hundred dollars. Finally, public TLDs will have a general availability phase, where registrations become first-come first-served, and registrants just pay the standard yearly registration rate for most names. Though ICANN has some minimum length standards for the sunrise phase, the registry chooses the exact length of sunrise, domain pricing and promotions, and all of the details about the other introductory phases.

A subset of TLDs are never made available for public registration. For these private TLDs, the only intended registrant is the registry itself, frequently to protect a brand mark. For example, the TLD `aramco` is closed to the public and only Saudi Aramco and its affiliates can operate domains under this TLD.

2.3 Examples

The data for this paper comes from hundreds of new TLDs, many of them managed by unique registries. We cannot describe all actors due to space limitations, but this section describes some of the larger registries and their TLDs.

2.3.1 Donuts

Though most registries run one or a small handful of new TLDs, Donuts Inc. is the largest and manages hundreds. Their TLDs largely consist of topical English words, such as "singles", "digital", and "coffee". The company's founders each have years of experience in the domain name industry, and the company secured over USD 100 million in venture capital [9]. Another large registry, Rightside, runs the technical infrastructure for Donuts TLDs. In return, Donuts gives some of its TLDs to Rightside after they reach delegation.

2.3.2 The `xyz` *TLD*

The `xyz` TLD is a generic alternative to `com` and is the largest in the new program. In the middle of 2014, Network Solutions, a large registrar, began offering `xyz` domains for free to some of their customers on an opt-out basis (e.g., the owner of `example.com` would find the domain `example.xyz` had appeared in their account). While registrants received these domains for free, Network Solutions still paid the registry full price for each domain [16, 26], although documents released as part of a lawsuit by Verisign suggest Network Solutions may have paid for these domains with advertising credit [8].

Due to this promotion, the number of registered domains in `xyz` rose by thousands per day in its earliest days until early August, when the number of registrations slowed to around 428,806 domains. Since then, `xyz` registrations appear at a much lower rate:

the number of domains finally doubled on April 13, 2015, taking over eight months to register a number of domains that originally took only two.

Registrants, however, appear to have only limited interest in these free domains. In our data set, 351,457 `xyz` domains (46% of `xyz`) remain unused and display a standard Network Solutions registration page when visited in a Web browser. Upon further analysis, we find that 351,440 of these domains appeared in the zone file in its first two months and still showed the unused Network Solutions template six months later. In fact, 82% of the 428,806 `xyz` domains in the August 2, 2014 zone file originated from this promotion and remained unclaimed as of early February 3, 2015. According to the monthly reports, Network Solutions had acted as registrar for only 360,683 `xyz` domains at the end of July, 2015, so registrants from this promotion claimed fewer than 10,000 free domains in the first six months.

2.3.3 *The* `science` *TLD*

The `science` TLD allows generic registrations, but targets the scientific community. Starting with general availability on February 24, 2015, the AlpNames registrar offered `science` TLDs for free. Similar to `xyz`, this promotion appears to have significantly impacted the number of `science` registrations: within only a few days, the TLD boasted 36,952 unique domains. The promotion has since ended, but the AlpNames registrar still sells `science` domains for $0.50, making it one of the cheapest TLDs. Two months after the start of general availability it had 174,403 registrations. Even though general availability started after our cutoff date, `science` is already the third largest TLD.

2.3.4 *The* `realtor` *TLD*

The National Association of Realtors owns the `realtor` TLD and targets accredited realtors, but also requires all registrants to prove that they are members of their association [22]. The registry provides the first year of registration for free to anyone that provides their NAR membership information. The promotion only applies to a single domain per NAR membership number. 46,920 `realtor` domains (51%) still show the default Web template provided by the registrar.

3. DATA AND INFRASTRUCTURE

We use data from many sources in our analysis, including zone files and several reports from ICANN. We actively crawl Web and DNS for each domain, and compare our findings with Alexa rankings and various blacklists. In this section we describe our data sources and data collection infrastructure.

3.1 Zone Files

When a registrant purchases a new domain from a registrar, the registrar sends a request to the registry with the domain and name server information. Once the domain goes live, it will appear in that TLD's *zone file*. At a high level, a zone file reflects a snapshot of a DNS server's anticipated answers to DNS queries. For a domain to resolve, it must have name server information in the zone file.

ICANN requires most registries to provide zone file access for a variety of purposes, including research. Some registries, such as most ccTLDs, do not need to provide access. For zones delegated prior to 2013, we gained access to their zone files by signing and faxing a paper contract to the TLD's registry, each of which gave us FTP credentials. We originally used this method to gain access to `aero`, `biz`, `com`, `info`, `name`, `net`, `org`, `us`, and `xxx`.

In anticipation for the rapid TLD expansion, ICANN developed a more scalable solution to zone file access requests, known as the Centralized Zone Data Service (CZDS). Registries and interested third parties can all apply for accounts on the service. After filling out their online profile with contact information and project details, requesting access to multiple zone files becomes straightforward. Registries still see multiple requests and can approve or deny them individually, but the process is much simpler. Once the registry provides access, the user can download the zone file through a simple API call up to once per day. Older TLDs can migrate to the new system for zone access, but progress has been slow; so far, only `museum`, `coop`, and `xxx` have migrated.

We have an account on CZDS, and manually refresh all new or expired approval requests almost once per day.[1] We have access to the zone files for hundreds of domains, most using the new CZDS system. We download a new snapshot of each daily, totaling 3.8 GB of gzipped text, more than half from `com`. For the analysis in this paper we simplify the zones and store all NS, A, and AAAA records on our HDFS cluster, and then store the raw zones on our archive server for future use.

3.2 ICANN Public Data

ICANN also requires each registry to provide a handful of summary reports. We have used most of them at some point in our methodology. The monthly transaction reports are particularly useful for our study. ICANN requires each registry to publish monthly summary statistics about the number of domains registered, transferred, expired, and renewed for each accredited registrar. We use the monthly summary reports to identify the number of registered domains that do not have any name server information and therefore do not appear in the zone file. We also use their breakdown of domains per registrar when estimating registration costs.

We also relied upon ICANN's New gTLD Current Application Status listing [20]. We used the data provided to determine TLD status and registry information as the new TLDs worked through the application system.

3.3 Our TLD Set

We include results for new TLDs that started general availability by the date of publication of ICANN's latest monthly registry reports on January 31, 2015, which altogether totals 502 new TLDs. Table 1 breaks down these new TLDs into various high-level categories, together with the total number of new domains registered in them at the time we crawled them.

We have focused our analysis on why registrants spend money on domains in the new TLD program. Some companies defensively register private TLDs, while others simply want a shorter domain name for their services. However, some companies in the latter category have not established their presence in their new TLDs yet, so we do not have a methodology to differentiate between these cases. Thus, we are more interested in public TLDs, where we can establish the registrant intent of individual domain names.

We differentiate public and private TLDs by checking public information about the start of general availability, as provided by several large domain registrars and nTLDStats [21], a Web site that tracks information on the new TLD program and is well-regarded in the domain community. Registries include their TLDs in these listings when they want public registrations, since the registrar collects this list in anticipation of selling domains in the TLD. This classification technique held up to the 15 randomly sampled private domains we verified manually. With this classification, 128 of the 502 new TLDs are private.

[1]We considered scripting our requests, but CZDS blocked obvious scripting attempts, so we did not pursue this further.

	TLDs	*Registered Domains*
Private	128	—
IDN	44	533,249
Public, Pre-GA	40	—
Public, Post-GA	290	3,657,848
Generic	259	3,061,416
Geographic	27	494,824
Community	4	101,608
Total	502	4,191,097

Table 1: The number of new TLDs in each category on February 3, 2015, and their sizes. For the three TLDs for which we had pending access requests, we used the size of the closest zone file.

In addition to the above, we found it difficult to learn substantial information about the new internationalized TLDs. In many cases, registrants can only purchase domains for them from international registrars. They tend to have rules for sunrise and general availability that we found unclear even with the help of a native speaker. As a result, we also do not include these 44 new TLDs in our analysis. Additionally, we focus on domains that reached general availability (GA) before our February 3, 2015 Web crawl so the set of registrants can include all prospective domain owners.

After removing private and internationalized TLDs from those that had already began general availability, we end with a set of 290 new public TLDs. The total set of TLDs includes generic words like `bike` and `academy` and geographical regions like `berlin` and `london`, both represented in Table 1. Additionally, four TLDs gate registrations to members of a particular community, such as the `realtor` TLD for accredited realtors. To give a sense of how many common word TLDs exist, our data set contains four synonyms for "picture": `photo` (12,933 domains), `photos` (17,500 domains), `pics` (6,506 domains), and `pictures` (4,633 domains). Table 2 gives an overview of the largest TLDs in our set, with some of the geographic TLDs featuring prominently. In the rest of this paper, we restrict our analyses to these 290 TLDs.

3.4 Active Web

For each domain in the zone file of a new gTLD, we visit the Web page hosted on port 80 of the domain with a crawler based on Firefox, an improved version of the crawler used in our previous study of `xxx` [11]. Our browser-based Web crawler executes JavaScript, loads Flash, and in general renders the page as close as possible to what an actual user would see. We also follow redirects of all kinds. After the browser loads all resources sent by the remote server, we capture the DOM and any JavaScript transformations it has made. We also fetch page headers, the response code, and the redirect chain.

Our primary data set for this paper is our Web crawl of all domains in the new TLDs on February 3, 2015. We chose this date due to its proximity to the timing of the latest ICANN reports, which reflect the number of registered domains in each TLD as of the end of January 2015.

3.5 Active DNS

Every time we Web crawl a domain, we also perform a DNS query using a DNS crawler developed for [15]. We follow CNAME and NS records and continue to query until we find an A or AAAA record, or determine that no such record exists. We save every record we find along the chain. We use DNS data to detect invalid NS records and to annotate each Web crawl with its CNAME chain.

GTLD	*Domains*	*Availability*
xyz	768,911	2014-06-02
club	166,072	2014-05-07
berlin	154,988	2014-03-18
wang	119,193	2014-06-29
realtor	91,372	2014-10-23
guru	79,892	2014-02-05
nyc	68,840	2014-10-08
ovh	57,349	2014-10-02
link	57,090	2014-04-15
london	54,144	2014-09-09

Table 2: The ten largest TLDs in our public set with their general availability dates.

3.6 Active WHOIS

Registry operators for most TLDs must publicly provide accurate domain ownership data using the WHOIS protocol. ICANN intends the use of WHOIS for "any lawful purpose except to enable marketing or spam, or to enable high volume, automated processes to query a registrar or registry's systems" [14]. In particular, ICANN encourages its use by consumers, registrars, and law enforcement, and discourages its use by spammers [29].

WHOIS server operators have leeway in how they achieve these goals. They typically rate limit requests, and responses do not need to conform to any standard format, which causes parsing difficulty even once records are properly fetched. We only query WHOIS for a small percentage of domains in the new gTLD program as an investigative step towards understanding ownership and intent.

3.7 Pricing Data

One dimension of our analysis focuses on the economic impact of the new TLD program, a task that requires domain pricing information. Unfortunately for our data collection purposes, registries do not sell domain names directly, but instead sell them through ICANN-accredited registrars. A registry can sell their domain names through any registrars they choose, but each must get similar wholesale prices and promotions [5].

We gathered pricing data for domains in the new gTLDs from a wide range of registrars. First, we collected data from the most common registrars for as many TLDs as possible. In some cases the registrar included a pricing table with information for many TLDs and we were able to automate the data collection process. Other registrars only showed pricing information after querying a domain name's availability, which required many separate queries. We made these queries manually. Some registrars made us solve a single captcha after five to ten requests.

Obtaining pricing information for the most common registrars simplifies the process and allows us to obtain a large number of (registrar, TLD) pairs in a short amount of time. However, we ultimately want to estimate pricing per TLD, so we would like to have registrar pricing data for many domain registrations in each TLD. Some TLDs do not sell well or are not available at the most common registrars (e.g., geographical TLDs for non-Western regions). We use the monthly registry reports to learn how many domains each registrar manages in each TLD, and we collect pricing information for the top five in each. Where possible, we also removed registry-owned domains from our analysis, since they did not cost anything. When registrars reported prices for non-standard time intervals or in foreign currencies, we used the current exchange rate to convert all prices to US dollars per year.

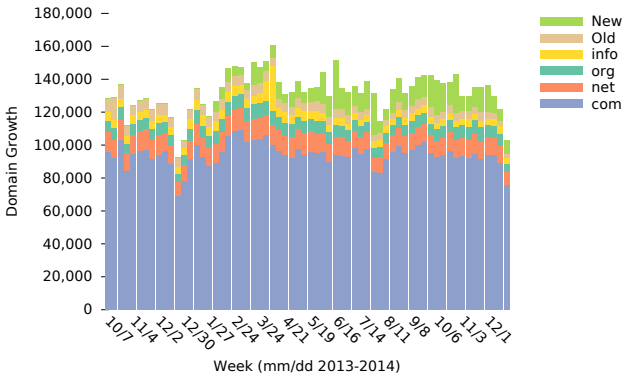

Figure 1: Number of new domains per day. Bars indicate the average rate for each week.

Registries reserve a set of strings which they sell for increased prices, known as *premium domain names*. For instance, GoDaddy sells normal `club` domains for $10 USD, but `universities.club` costs $5,000 USD, and this increase in price represents revenue to both the registry and registrar. These domains number in the thousands for any given TLD, and prices can vary per string. Our methodology treats premium domains as normal domains, thus underestimating registry and registrar revenue. Premium domain sales do not always correlate with wholesale revenue, and we do not see a scalable method to address this problem.

3.8 Alexa

We use the Alexa top million domains list to make an estimate of how often users visit domains in the new TLDs [1]. Alexa collects their data by allowing browser extensions to include their measurement code in exchange for providing domain analytics, and by allowing Web page operators to do the same. We use a domain's presence in the list as an indication that users visit it, but do not place any emphasis on domain rankings.

3.9 Blacklists

We also compare new domain registrations with URIBL, a publicly available domain blacklist, to see how the blacklist rate compares between old and new TLDs [27]. We use their high-volume rsync instance to download a new copy of the blacklist every hour. Though they provide many types of blacklists, we only use the standard and highest-volume blacklist, labeled "black", as the rest tend to be lower volume. This list represents the domains most likely to be malicious, while the other lists include domains detected through more experimental methodologies.

4. REGISTRATION VOLUME

We first look at the impact of the new TLDs on overall registration volume. The new TLDs represent new opportunities for registering domains. As registrants create new domains, one possibility is that they decide to create them in the new TLDs rather than the old, thereby displacing registration activity in the old TLDs (e.g., because names taken in `com` are available in the new TLDs). Another possibility is that the new opportunities motivate even more registrations, thereby growing total registration activity overall.

Figure 1 shows the number of new domain registrations per week broken down into various categories. Days for which we did not have access to the zone files resulted in slight drops in the graph.

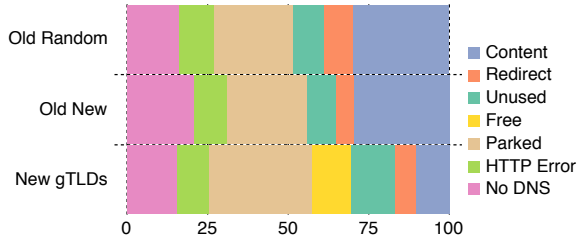

Figure 2: Classifications for all domains in the new TLDs, a random sample of the old TLDs, and a month of new domain registrations in the old TLDs.

We show the most active old TLDs individually, the remaining old TLDs grouped into "Old", and the new TLDs in "New".

Overall, the introduction of the new TLDs had only minimal impact in the rate of registration of the old TLDs. The new TLDs generally increase the total number of registrations rather than shift focus from old to new TLDs. However, the new TLDs see far fewer registrations than the old TLDs, largely because `com` continues to dominate.

5. CONTENT CATEGORIES

As a first step towards learning the intent of each domain's registrant, we classify the technical data each domain returns when queried by our DNS or HTTP infrastructure. We perform this classification with features of both crawls, including DNS CNAME records, Web headers, Web contents, and the NS records in the zone files.

Domains with invalid DNS or HTTP errors are straightforward to identify, but in many instances, we need to classify the domains based on the textual content they return to HTTP queries. We use a combination of automated machine learning techniques and manual inspection of Web pages hosted at these domains.

We assign each domain to one of the following seven categories:

No DNS domains do not successfully resolve DNS queries.

HTTP Error domains have valid DNS, but do not return an HTTP 200 when queried.

Parked domains are owned by an ad network or are for sale by their owners and typically return Web pages dominated by ads.

Unused domains return HTTP content that is not consumer-ready, including empty pages, default Web server templates, or PHP errors.

Free domains include domains given out as part of a promotion that still have the original template, as well as domains with registry-owned Web templates.

Defensive Redirect domains redirect through one of several technical mechanisms to a different domain name.

Content domains host valid Web content for users to visit.

We start by presenting high-level content categorizations, including domains in the older TLDs as a reference point. Then Section 5.2 provides more detail about our clustering methodology, and Section 5.3 describes the seven categories in more detail.

5.1 Content Summary

To place the new TLD results in context, we present domain classifications for three data sets. The first includes all domains in the new TLDs as of February 3, 2015. The second includes 3 mil-

Content Category	*Results*	
No DNS	567,390	15.6%
HTTP Error	362,727	10.0%
Parked	1,161,892	31.9%
Unused	504,928	13.9%
Free	432,323	11.9%
Defensive Redirect	236,380	6.5%
Content	372,569	10.2%
Total	3,638,209	100.0%

Table 3: Overall content classifications for all domains in the zone file for the new public TLDs.

lion domains from the old TLDs defined in Section 3.1 chosen uniformly at random. The third includes all domains in the same set of old TLDs that were newly registered during December 2014. (Delays in our `com` processing pipeline prevented us from using a more recent data set.) Figure 2 summarizes all three data sets. This paper focuses on the new TLDs, so we focus on those domains. Table 3 shows exact values for the 290 public English TLDs described in Section 3.3, minus `quebec`, `scot`, and `gal`, the TLDs for which we did not have zone file access at the time.

For most categories the classification breakdown is comparable among the three data sets: erroneous domains (No DNS and HTTP Error) account for about a quarter of all domains, another quarter utilizes domain parking, and roughly 20% of domains are either unused or redirect elsewhere. The old and new TLDs differ greatly in content and promotional domains: the new TLDs show a dearth of content, but make up for it with a high volume of free domains, which domain owners do not actively use yet.

Figure 3 shows our content classification for the 20 largest TLDs that allow public registrations. Most TLDs show a typical split between the major content categories, but other TLDs show very different registration types, especially those with free domains.

5.2 Content Clustering

Our goal is to cluster Web pages hosted at domains into one of the content categories. Two key challenges to classifying content are the sheer size of the data (millions of domains), and the lack of labeled data for training a classifier. With so many unlabeled Web pages, we must learn from scratch to classify the domains.

Our first step is to cluster Web pages with highly similar content. This procedure groups together duplicate and near-duplicate Web pages, which commonly arise when HTML is automatically generated using a fixed template. Prevalent examples include parked pages, and default placeholder pages served by a registrar before the registrant publishes any content.

To map Web pages to inputs for a clustering algorithm, we follow a conventional "bag-of-words" approach which extracts HTML features from the Web pages. In particular, we compose a dictionary of all terms that appear in the HTML source code, and for each Web page, we count the number of times that each term appears. In this way, each Web page is represented as a sparse, high-dimensional vector of feature counts. We implemented a custom bag-of-words feature extractor which forms tag-attribute-value triplets from HTML tags, as described in [7].

For reasons of computability and conciseness of results, we begin by clustering roughly one tenth of the crawled Web pages. We used the k-means clustering algorithm with $k = 400$ to organize these Web pages into groups of high similarity (based on the Euclidean distance between their feature vectors). We set k to be intentionally large because we wished to discover especially cohesive clusters of replicated Web pages.

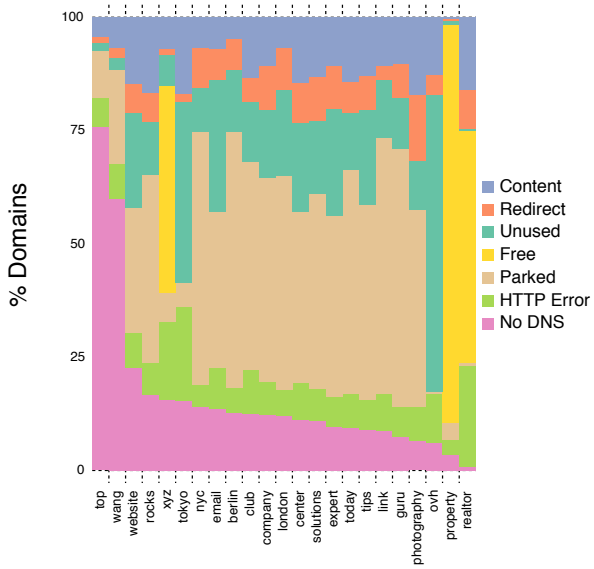

Figure 3: Domain classifications in individual TLDs for the 20 most common. We have sorted TLDs by fraction of "No DNS" to better highlight the category breakdowns of successful content.

Next we manually inspected the resulting clusters using a custom visualization tool. The tool displays screenshots of how the Web pages rendered in our crawler and provides a link to the HTML source next to each screenshot. To facilitate efficient manual review, the tool presents a condensed view of the clusters by showing only a sample of Web pages in each one. Specifically, it sorts the Web pages in each cluster by their distance to the cluster centroid, then displays the top and bottom-ranked pages as well as a random sample of pages in between. If all Web pages in this sample are visually nearly identical, we can conclude with confidence that the entirety of Web pages in the cluster have been appropriately grouped. Furthermore, we can classify Web pages in these perfectly homogenous clusters all together.

By examining the clusters, we placed domains into three broad categories according to their content: parked, content-free, and meaningful content. Our clustering approach was particularly effective at identifying large numbers of parked domains and content-free (or unused) domains that host a default registration page. The class of Web pages with meaningful content exhibits the most variety: Web content is highly diverse and unlikely to have the same degree of replication as the other two classes. Thus at this stage, we focused only on bulk labeling of clusters that clearly contained parked or content-free Web pages. If it was not visually obvious how to label a cluster in bulk, then its pages remained unclassified at this point. (In practice, though, we found that Web pages with content often were grouped together in clusters with wide diameters.)

After this phase of clustering, manual inspection, and labeling, we then aimed to classify domains that were not included in the initial subset. Now equipped with a large number of labeled examples, we used nearest neighbor classification to discover many more candidate Web pages which are likely parked or content-free. First, we extracted HTML features from the remaining Web pages, then mapped the pages into the same feature space as the original subset.

Then for each unlabeled Web page, we found its nearest neighbor by Euclidean distance in the labeled set and, if the distance was less than a strict threshold, we marked the page as a candidate for its neighbor's class. This thresholding minimizes false positives. This step continues to focus only on parked and content-free pages; no content pages were classified in this way. We modified our visualization tool to display candidates next to their nearest neighbor; if the Web pages were visually nearly the same, then we were confident in assigning the appropriate label to the candidates.

In one round of this nearest neighbor method, we were able to label many of the remaining (non-content) Web pages in our data set with high confidence. However, since we only clustered about one tenth of the Web pages at the outset, we likely missed different templates that did not appear in the initial subset. Thus, we iterated this approach to achieve greater coverage. That is, we clustered the remaining unlabeled Web pages, manually inspected and labeled homogenous clusters, and performed thresholded nearest neighbor classification—now with a larger set of labeled examples. We iterated this process until there were no more obviously cohesive clusters. Finally, after identifying all parked and content-free domains, we manually inspected a random sample of the remaining unlabeled Web pages. The results gave us confidence to conclude that the remaining Web pages contain legitimate content.

5.3 Content Categories

We use this content clustering methodology to create a cluster label for each domain. Then, we took any page metadata (e.g., DNS errors, HTTP status code, the redirect chain, etc.) and combined these features together to make a final classification.

The rest of this subsection describes how we combined those features to determine a final content category. For domains that might fall into multiple categories, we prioritize categories in the order listed in Table 3. For example, for parked domains that redirect to a different domain, usually as part of the parking program, we only classify as "Parked" and not "Defensive Redirect".

5.3.1 No DNS

Registrants purchase domain names from a registrar and pay a yearly fee to keep them, yet a large fraction of domains in the new gTLDs do not even resolve. Some of these registrants associate name server information with their domains, but these servers do not respond to DNS queries, or only respond with the DNS `REFUSED` error code. For instance, `adsense.xyz` has an NS record for `ns1.google.com`, but its name server returns `REFUSED` for all queries (which recursive resolvers usually report as `SERVFAIL` to the end user). Out of 3,638,209 domains in the new TLDs, we had 567,390 DNS failures with an associated NS record, or 15.6%.

Other registrants buy domains and then do not associate name server information with them. Since the zone files only contain associations between domain names and name servers, they contain no entries for this set of domains. We do not have a list of these domain names and do not have a clear mechanism to find them.

While we cannot enumerate these domains, we can infer their presence through the ICANN monthly reports. The monthly reports provide a summary of domain activity and transactions for all registered domains (i.e., domains with a yearly fee). We can use the difference between the number of domains in the ICANN reports and the number of domains in the zone file as an estimate for the number of domains with no name server information.

Our analysis shows that out of 3,754,141 total domains in the reports, 207,184 domains (5.5%) do not appear in their respective zone files. Registrants pay for these domains like any other, but they do not resolve.

Error Type	*Result*	
Connection Error	110,144	30.4%
HTTP 4xx	82,298	22.7%
HTTP 5xx	138,471	38.2%
Other	31,814	8.8%
Total	362,727	100.0%

Table 4: Breakdown of HTTP errors encountered when visiting Web pages.

5.3.2 HTTP Error

We next classify domains that resolve to an IP address, but return no result or an HTTP error code when queried on port 80. We suspect some of these error conditions are temporary. Others are likely longer-term misconfigurations by owners who do not care about the content hosted on the domains, making them likely brand defenders. Alternatively, these domains might serve a legitimate purpose that is motivated by content other than Web. Because we use the status code from the final landing page, even HTTP 3xx status codes indicate errors, typically a redirect loop.

We received 362,727 responses to that we classified as HTTP errors. Table 4 provides a breakdown. Notably, most domains in this category exhibit connection issues such as timeouts or return HTTP 5xx return codes, meant for internal server issues. The variety of errors is multifarious: overall we received responses with 43 unique HTTP status codes.[2]

5.3.3 Parked

Many domain registrants do not have a plan to monetize the content of their domain names. Most of them are speculating on the name itself, intending to sell it later for a profit. Some may have a plan to develop the site later in its lifetime, but have not put up any content yet. Still other owners initially created Web properties that turned out to be unsuccessful, and later parked them while waiting for expiration. Whatever the reason, domain parking is common in all TLDs. We discovered 1,161,892 parked domains in our data set, or 31.9% of all domains in the zone files.

Potential domain speculators have the choice of a large number of parking services. Some parking services also act as domain registrars (e.g., GoDaddy and Sedo), while others focus solely on parking. Registrants use their services by setting their name server (NS) record to the parking service's DNS servers, redirecting their Web traffic to the parking service, or setting a CNAME. Parking services that also act as registrars may or may not use different name servers for parked domains compared to normal registrations.

Parked domains come in two main varieties [3]. Most domain parking monetization is through pay per click (PPC) advertising. These parked pages look much like search result pages with links pertaining to words in the domain name. Each link on this page is an advertisement. Other parked domains use pay per redirect (PPR). When the target domain's owner purchases "direct navigation traffic" from an ad network used by the parking program, the parking service will redirect the user to a page run by an ad purchaser. Decisions to serve PPC or PPR to any particular visitor happen in real time based on characteristics provided by the traffic purchaser, including domain keywords or traffic from limited geographic regions.

[2]Six domains responded with the HTTP response code 418, an error code added as part of the Hyper Text Coffee Pot Control Protocol in a satirical RFC [13]. The return code means "I'm a teapot".

Feature	*Domains*	*Coverage*	*Unique*
Content Cluster	1,080,283	92.3%	277,754
Parking Redirect	638,757	55.0%	81,468
Parking NS	279,903	24.1%	124
Total	1,161,892	—	

Table 5: Our capture methods for parking and how many domains each catches. We identify most parking domains with more than one classifier; column 2 shows how many domains each classifier identifies, while the last column shows how many are unique to that classifier.

As a starting point, two previous studies also needed to classify parked domains as part of their work. Alrwais et al. focus on how parking programs operate, and use domains from known parking name servers as their source [3]. Vissers et al. focus on classifying parked domains, but use parking pages from known parking name servers as their inputs [28]. However, our problem is slightly different since we want to identify random pages from the Internet as parked or not. Some parking programs host both legitimate and parked pages using the same name servers, including one of the largest parking services, GoDaddy. We need a different approach to identifying parking than either of these papers suggest.

We identify parked domains with three mechanisms. First, we use our k-means content classifier to identify PPC parking services. Often there are many of these pages for each parking service, with variations only in the displayed links; all layout and remote resources remain constant for any given parking service. As such, they tend to cluster well and are easy to identify with this method.

Second, we use the visit's full redirect chain, acquired with the methodology described in Section 5.3.6, to identify PPR parking. These domains usually redirect through an ad network before landing at their final destination for accounting purposes. We manually inspected redirect chains for visits to known parking name servers to compile a set of URL features that indicate parking. For instance, if any URL contains "zeroredirect1.com" or both "domain" and "sale", we classify the domain as parked.

Finally, we use known parking name servers, such as those for `sedoparking.com`. We use this method only for servers we are confident host solely parked domains. We start by taking the intersection of the different sets used by Alrwais et al. [3] and Vissers et al. [28]; the intersection includes all but one of the name servers from the latter set. For each name server in the set intersection, we use our k-means classifier to determine if domains using that name server are parked or not. For those we did not identify as parking (a very small set), we manually inspect a random selection of screenshots and their redirect chains. If we believe them all to be parking traffic missed by our classifier, then we assume all domains using the name server are parked. With this additional verification step, we concluded with high confidence that all 14 name servers in our set are used strictly for domain parking. Finally, we added one additional name server (`parklogic.com`) to our set, which we found to be dedicated to parking services through our classification experiments.

Table 5 shows how many parked domains we identify with each method. We identify most parking domains with more than one of our three methods. In particular, we identify all but 124 of nearly 280,000 domains on our set of parking name servers with another approach. This high detection accuracy provides validation of our other parking classifiers, and further increases our confidence that we have identified the prevalent parking behaviors.

Mechanism	*Domains*	*Coverage*	*Unique*
CNAME	2,020	0.9%	729
Browser	211,065	89.3%	203,941
Frame	30,437	12.9%	24,571
Total	236,380	—	

Table 6: The mechanisms domain owners use to redirect to a different domain. Most domain owners use only browser-level redirects, but frames are still very common. Very few content domains use multiple redirect methods.

5.3.4 Unused

In our analysis, we find many Web pages that fit in none of the above categories, but also do not provide meaningful content. Most of these are placeholder pages served by a large registry with instructions for the owner on how to develop their domain. Others are empty Web pages, or the default template provided by a software package. Whatever the reason, these pages do not provide meaningful content to end users and we refer to them as "Unused".

Unused pages often appear in bulk, so we identify them using our k-means classifier. With this technique, we find 504,928 content-free domains in our data set, or 13.9% of domains in the new TLDs.

5.3.5 Free

Domains we identify as part of a promotion, such as those described in Section 2.3, get their own content classification. Most of these domains fall into the "Unused" category through a strict categorization, but the registrant plays a different role for these (which will be relevant when determining intent in Section 6).

Though not part of a promotion, the `property` TLD largely contains domains owned by Uniregistry, its registry. The TLD showed slow growth in all other time periods, but on February 1, 2015 it grew from 2,472 to 38,464 domains in a single day. Uniregistry owns all of these domains and hosts a standard sale page with the text "Make this name yours." We place these registry-owned content placeholders into the "Free" category as well.[3] In total, we find 432,323 free domains in the new TLD program (11.9%).

5.3.6 Defensive Redirects

Many domains in the new gTLDs have at least one redirect, and most of these point to a different domain. The role of the redirect depends on the type of content. Some parking programs redirect from the initial domain to a standard parking page, using the URL parameters to pass a domain identifier for revenue sharing purposes. Defensive registrations often redirect to the owner's other domain names, typically in an older TLD. We check for three kinds of redirects: CNAMEs, browser-level redirects, and single large frames. Table 6 shows how many domains redirect with each mechanism.

A CNAME is a DNS record type that acts like a symbolic link between two domains. Any DNS query that results in a CNAME causes the resolver to perform the same query on the target. Sometimes the result is another CNAME, which our DNS crawler must follow before finally resulting in an answer to the original query. Most domains with a CNAME only have a single CNAME, but chains of up to four are not uncommon in CDNs. For example, in our February 3 data set, the domain `tangyao.xyz` has a CNAME to `scwcty.gotoip2.com`. This domain has its own CNAME to `hkvhost660.800cdn.com`.

[3]We do not classify them as "Parking" because they do not show ads and they are owned by the registry.

Redirect To	*Number*
Defensive	236,380
Same TLD	7,135
Different New TLD	5,843
Different Old TLD	98,923
com	124,479
Structural	75,073
Same Domain	74,379
To IP	694
Total	311,453

Table 7: Which locations our visits were ultimately redirected towards.

Browser-level redirects happen when DNS resolves to a host running an HTTP server, but a query to that server returns a redirect which our browser will follow automatically. For example, an HTTP request to `tucsonphotobooth.com` returns an HTTP 302 redirect to `bumblebeephotobooth.com`, which modern browsers obey without user interaction. A domain owner can do this in a very large number of ways, such as with a 300–399 status code, an HTTP header, an HTML `meta` tag, or using JavaScript to set window.location. We find and store these redirects at crawl time, so we are robust to these and less common methods.

In practice, we find many pages that return valid HTML, do not redirect, and present only a single large frame to the end user, such that all visual content comes through the frame. Although it does not use an explicit redirection mechanism, this technique provides the same effect: a user visits one domain on their browser, and sees content from another. Since these frames serve the same purpose as a CNAME or an HTTP redirect, we consider these to be redirects as well.

To determine if a page contains only a single large frame, we first check how many frames the page contains. We do this in JavaScript in the browser, so we do not need to use textual analysis to find them. The remaining challenge is to differentiate between pages with a single large frame, and pages with real content that have a smaller frame, such as for page navigation or tracking purposes.

We differentiate between these classes using the DOM. First, we remove non-visible components from the page, as well as anything having to do with the frame itself: the `head` tag, `frameset` and `iframe` tags, and long URLs. These modifications are safe because we operate on the DOM, not the original HTML, so non-visible components that transform visual components (such as JavaScript) have already run. By examining the string length of the resulting DOM, the pages we crawl fall cleanly into two classes. Altogether, 49% of the filtered DOMs have a string length of less than 55 characters, but show variable behavior based on the few remaining tags. The remaining pages distribute mostly evenly with a few spikes corresponding to common page templates. A visual examination of the pages in these clusters shows that the short pages do show only a single large frame, while most of the large pages have other visual content.

The most important two pieces of the overall redirect chain are the starting domain and the final page that serves content. To determine the last, we check for a single large frame first, then a browser-level redirect, and finally a CNAME. A domain with all three behaviors serves its real content through the frame; the CNAME and browser-level redirects only point to the next resource. We classify redirects by the domain they point to: same-domain, same-TLD, "com", new-TLD, old-TLD, or IP.

Table 7 shows which of these six location types domains in the new TLDs tend to point towards. Though each of these domains has some form of redirect when fetching Web content, redirects to a page under the same domain name are less interesting because they reflect aspects of the structure of the Web page itself. Similarly, we cannot make any strong claims about redirects to a hard-coded IP address.

Instead, we only consider redirects to a different domain to fall into our redirect category. We do include redirects to other domains within the same TLD because in this case, the registrant is only using the destination domain for primary purposes. We find 236,380 off-domain redirects in our data set, or 6.5% of all domains in the new TLD zone files. 94.5% of defensive redirects point to domains in the old TLDs, with over half of those to `com`. In short, defensive redirects are only a small fraction of the overall registration behavior in the new TLDs.

5.3.7 Content

We classify domains under "Content" when they do not fit into another of our content classifications. The other aspects of our categorization pull out common errors, interesting features like redirects, and Web responses that appear frequently. Domains that do not fit into any of those categories resolve in the DNS, return HTTP 200 status codes, and provide vaguely unique responses to Web queries. Only 372,569 domains (10.2%) fall into this category. By comparing this category with the previous, we find that 38.8% of the 608,949 domains with real content redirect to a different domain to serve it.

6. REGISTRATION INTENT

In the previous section, we focused on understanding the types of content that domains in the new gTLDs host. In this section we explore the high-level intent of the domain's registrant. For each domain, we infer what motivated its registrant to spend money on the name. We classify registration intent into one of three broad categories:

Defensive registrants purchased a new domain to defend an existing Web presence.

Primary registrants own domains with the intent to establish a Web presence.

Speculative registrants intend to profit off of the name itself and never plan to develop a meaningful Web presence.

Before classifying domains by registration intent into one of the above categories, we must remove some types of domains. We ignore domains in the "Unused" and "HTTP Error" categories. We could guess that these domains tend to include more defensive than primary motivations since they are not user-ready and therefore the use of the name is the only relevant effect on the Internet. However, registrants likely buy domains they intend to develop all the time, and these domain names may transition to other categorizations given time or result in expirations.

We also ignore domains in the "Free" content category before deciding registration intent. In a typical domain registration scenario, we know registrants have expressed genuine interest in the domains they own because they paid money for them. Without ignoring the "Free" content category, we could not use the results of our registrant intent classifications to make any claims about why registrants purchase domain names.

Table 8 summarizes our results. In the following sections, we describe each registration intent category in more detail. We discuss what types of registrants we expect each category to cover and how we map content categories to registration intents for each domain.

Intent	*Results*	
Primary	372,569	14.6%
Defensive	1,010,954	39.7%
Speculative	1,161,892	45.6%
Total	2,545,415	100.0%

Table 8: Registration intent categorizations for the new public TLDs.

6.1 Defensive

Our defensive registration intent set begins with domains that redirect to a different domain name. Some off-domain redirects could reflect primary registrations: registrants could use their old name for technical or historical reasons but primarily use and market the new domain name. However, we find in practice that most are defensive, and many lead to sites whose branding and headers clearly advertise the landing domain.[4]

Additionally, we include domains that return invalid DNS results in this set. Owners of non-resolving domains could only use their names for private purposes, since traffic routed through the public Internet cannot correctly address a remote server. A more likely explanation is that the registrant only cares about the name. We include domains with invalid NS records as well as those that do not appear in the zone file (both described in Section 5.3.1), for a total of 774,574 non-resolving domains. Combined with the 236,380 defensive redirects, we find defensive registrations of 1,010,954 domains in the new TLDs.

6.2 Primary

Primary domains include all those purchased by a registrant with the intent to use that specific domain. Most primary registrants purchased their domain to establish a Web presence, but there are other kinds of primary registrants as well. We only classify domains in our "Content" category as primary registrations. Each of these domains resolves and could conceivably host content intended for end users. Our clustering technique did not find similar Web content for these domains, so registrants of those domains at a minimum host sufficiently unique content.

6.3 Speculative

Many registrants purchase domains to speculate on the domain itself with no intent to develop content. Most make use of the first-come first-served nature of domain registrations to grab domains they believe others will find desirable in the hope of selling them later for a profit. Others host parking-based advertising and pay-per-redirect services with the goal of monetizing through ad revenue, but still with no intent to develop unique content. In practice, most speculators in the first case also host parked content because it is essentially free (and often bundled with domain registration fees), and also serves as a signal to prospective buyers that the name is available.

From a content standpoint, the difference between a defensive and speculative registration is relatively narrow. Defensive registrants purchase domains to defend the string but with no intent to develop content, while speculative registrants purchase domains to resell later with no intent to develop content. However, speculative registrants are monetarily motivated on a per-domain margin, while defensive registrants have revenues outside the domain business. A speculator must monetize the name, but a defender does not. Therefore, we classify parked domains as speculative and non-resolving domains as defensive based on this distinction.

[4]Trademark holders make defensive registrations on their own brands. The same registration made by a different actor with malicious intent would instead qualify as cybersquatting.

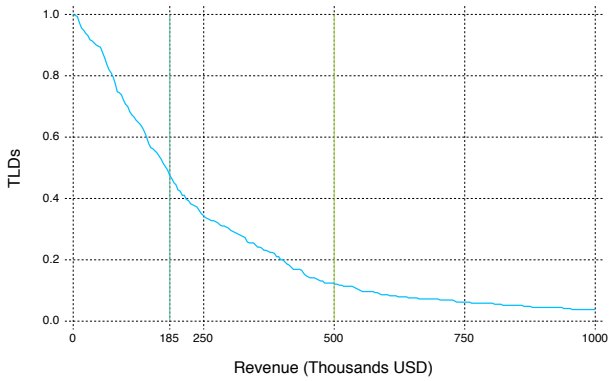

Figure 4: New gTLD program revenue as a CCDF across all TLDs. The vertical line at $185,000 USD corresponds to the minimum ICANN application fee, and the line at $500,000 USD corresponds to a more realistic estimate of the cost of establishing a new TLD.

7. REGISTRATION COSTS

Previously, we focused on the new TLD system from a registrant-centric perspective. In this section we look at the new TLD rollout from the point of view of the registries. We examine how registries make money and how they interact with registrars in practice.

7.1 Registry Financials

Using the methodology described in Section 3.7, we obtained pricing information for 2,006 (TLD, registrar) pairs, which account for 73.8% of all domain registrations. In only four TLDs do we record prices from fewer than three registrars; in each case, however, the one or two registrars we do record account for at least 97.5% of all domains. In the remaining 26.2% of domain registrations for which we do not have matching data, we use the median price for the TLD.

Figure 4 shows a complementary cumulative distribution function of the cost to registrants per TLD. A point on the line shows the ratio of new TLDs that have made at least the corresponding amount in registration costs. We included a vertical line at $185,000 USD, the standard application fee for a new TLD [19]. At this cost, roughly half of all TLDs made this money back. We estimate the total cost to registrants for domains in the new TLDs at $89 million USD through March 2015.

The application fee, however, only represents a lower bound on the amount each registry spent on their TLD. Additional costs to ICANN include a quarterly $6,250 fee [5], a per-domain transaction fee for registries with more than 50,000 transactions per year (a threshold only 18 TLDs have met), and additional application fees for TLDs that must enter the contention process. While registries do not have many other explicit costs, the TLD application process ran for years before the first delegation; presumably registries built up legal or personnel costs in the meantime. Registries also need to connect with registrars, market and brand their TLDs, build a Web presence, and run or outsource technical operations.

As a result, we also include 500,000 USD as a more realistic estimate for the cost of establishing a new TLD. While it is conjecture, some TLDs have already gone up for auction, like `reise` [23] and

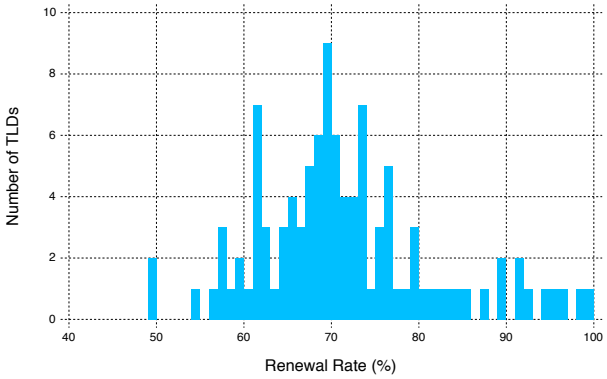

Figure 5: A histogram of renewal rates per TLD.

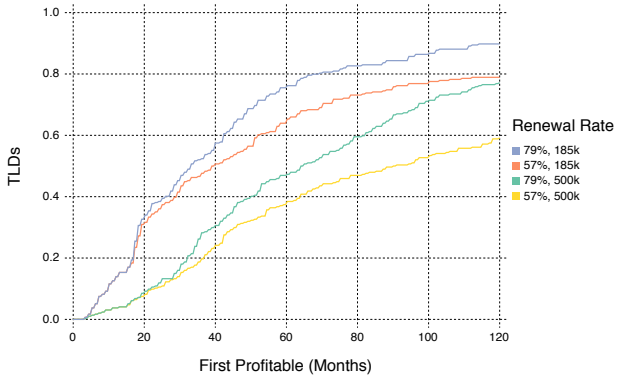

Figure 6: Registry profitability over time under different revenue models. A point on a line indicates the fraction of TLDs that were profitable within the given time since general availability.

`versicherung` [4], which set reserve prices of 400,000 USD and 750,000 USD, respectively. Given the small number of registrations each had at the time, these TLDs were valuable because they had completed the delegation process, suggesting the sale price roughly reflects the cost of delegation. With these reserve prices, we chose 500,000 USD as a rounded estimate in this range. At this estimate, only about 10% of TLDs are profitable.

Revenue from domain registrations does not all go to the registry. Instead, registries and registrars split revenue based on a previously agreed upon model. For instance, Verisign makes $7.85 USD per `com` registration [6] and $6.79 per `net` registration [18]. During our pricing data collection, we found registration prices for both `com` and `net` names ranging from $8 to $13 USD, or markups ranging from $0.15 to $6.[5]

Unfortunately, the new registry agreements do not specify maximum wholesale prices, only fees the registry must pay to ICANN. For calibration, we can get a handful of prices through registry-reported earning data. Rightside, one of the largest back-end registry providers, is funded through private investors and has released some revenue statistics online in a presentation meant for investors and analysts [25]. They provide end-of-November wholesale and total revenue numbers for five TLDs, two of them aggregated. Our estimate is too low for `reviews`,[6] but our other estimates overestimate the wholesale price by close to a factor of 1.4. Our model does not factor in premium domain name sales, a non-trivial revenue source that does not correlate well with wholesale price. As a result, Figure 4 represents a low estimate of domain name costs, and we discuss the limitations of our model further in Section 7.4.

7.2 Renewal Rates

All registries in the new gTLD program anticipated the one year and 45 day mark since the introduction of the earliest TLDs [2].[7] This milestone provides the first chance for registrants in the new TLDs to renew their domain names, and hence reflects ongoing demand for domains in the new TLDs after one year's actual, rather than anticipated, experience with the domains. Donuts, the largest registry with over a hundred new TLDs, published statistics on renewal rates for their earliest TLDs [10, 24], likely in an attempt to attract registrars and investors [30]. However, Donuts limited their analysis to their own TLDs, and also did not provide numbers past 26 days.

Figure 5 shows a histogram of renewal rates by TLD. We only performed our analysis on TLDs where at least a hundred domains completed a full year of registrations plus the 45-day Auto-Renew Grace Period. The Donuts TLDs in our data set show renewal rates within a few percentage points of the numbers Donuts reported in April. We calculate an overall renewal rate of 71%.

7.3 Future Profit Modeling

In this section, we take a look at registry profitability using a variety of parameters. In face of the limitations of our profit modeling discussed in Section 7.4, we acknowledge that drawing higher-order conclusions from such limited data could lead to models that are incorrect in unpredictable ways. However, we would still like to attempt to classify "successful" TLDs, and profitability is a strong indicator of the success of any company.

We start by graphing TLD profitability over time under four different models in Figure 6. A point on a line indicates the fraction of TLDs that were profitable within the given time since general availability. We show four curves that reflect different values for two parameters. Two of the models assume an initial cost to the registry of only 185,000 USD, or the amount of the ICANN application fee. This is the minimum amount we know all registries must pay. The other models assume an initial cost of 500,000 USD, which better reflects our understanding of the cost of creating a registry. The second parameter is renewal rates. We show models with renewal rates of 57% and 79%, which reflect lower and higher than average rates and show the sensitivity of the model to renewals.

For each TLD, we collect registration volume data from the reports provided via ICANN. We consider TLDs for which we have three monthly reports after general availability. The first month typically contains a burst of registrations, and then the second and third provide two data points at a more typical registration rate. We model future months based on new registrations at this rate, and renewals of domains registered or renewed 12 months prior at the indicated renewal rate. We estimate the wholesale price as 70% of the total price at the cheapest registrar.

[5]Registries can offer "bulk discounts and marketing support and incentive programs" but must offer similar terms to all registrars [6].

[6]The price we found for `reviews` domains through two registrars owned by the same company as Rightside is less than its wholesale price. We found pricing for November through `archive.org` [17] and found that the price to registrants of a `review` domain has halved. We do not know if this reflects a reduction in its wholesale price or a promotion.

[7]The extra 45 days is for the Auto-Renew Grace Period, which allows registrars to keep the registrations for free. Usually the registrar uses this time to offer the registrant one last chance for renewal, in case they let it expire accidentally.

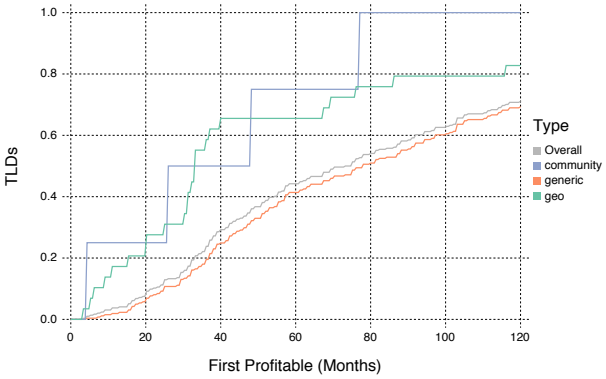

Figure 7: Modeling profitability by type of TLD. The gray line represents the aggregate, and the colored lines represent the set of TLDs of the indicated type.

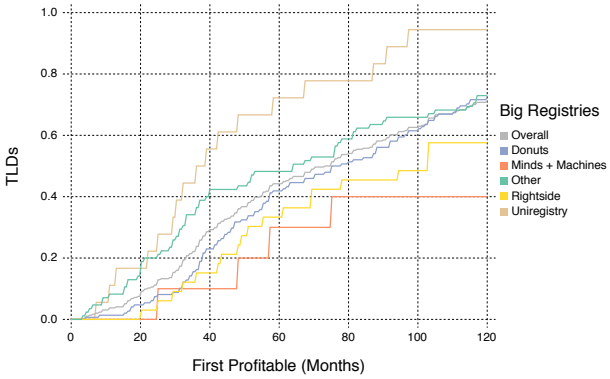

Figure 8: Modeling profitability by registry for the registries with the most TLDs. The gray line represents the aggregate, with colored lines representing individual registries.

Figure 6 shows that the initial cost plays a much larger role than the renewal rate in the short term, but that both parameters are important in the long term. We find that even under the most permissive model, with high renewal rates and no fees beyond those imposed by ICANN, 10% of TLDs still do not become profitable within the first 10 years.

Since there are a wide variety of registries operating new TLDs, and there is a wide variety of domain registration activity across the TLDs, we were interested to see if there were features that might separate profitable and unprofitable TLDs. To that end, we compared profitability based on four metrics:

- lexical string length;
- the registry for TLDs belonging to the top four registries, otherwise "Other";
- the type of registry ("generic", "community", or "geographic"); and
- whether or not the most common registrars all sell domains in the TLD.

In practice, we only found minor variations in profitability based on these metrics. We present results for the most significant differentiators, type and registry, below.

Figure 7 shows variations in profitability by type of TLD. The gray line represents the overall profitability CDF. It is equivalent to the profitability CDFs in Figure 6 with an initial cost of 500,000 USD and an overall renewal rate of 71%. The remaining lines represent non-overlapping TLD subsets which combine to the same overall set. Though community and geographical TLDs become profitable much sooner than generic TLDs, there are so few of them in comparison that the profitability of generic TLDs still closely tracks the overall rate.

Similarly, Figure 8 shows variations in profitability by registry. Of the large registries, only Uniregistry TLDs become profitable sooner than the average. Instead, our data suggests owners of multiple TLDs mainly benefit by spreading the risk. Many registries only manage between one and three TLDs, and those strings tend to become profitable sooner than most of the large registries.

7.4 Limitations

We see profitability as an important metric with which to compare registries, but our methodology has some limitations. In this section, we describe the known limitations and their expected impact on the results.

First, our pricing model does not include premium domain name sales, as described in Section 3.7. For the few TLDs for which we have seen premium domain revenue reports, these sales vary considerably. For different TLDs, we have seen the total revenue from premium domain sales range from $0 to the same amount as the total revenue of wholesale domains. It is also plausible that some TLDs could get more total revenue from premium domain names than from standard registrations. As a result, this category represents the largest unknown in our model. Premium domain names renew for the normal registration cost, so this unknown only affects the initial upfront purchase of the name and not ongoing renewal revenue.

Second, for any given TLD/registrar pair, we only record a single price, when domain name prices could change over time. To date we find that, after the beginning of general availability, domain prices do not change very frequently. Future studies could address this assumption by periodically regathering pricing data. For practical reasons, doing so would require deploying a more automated method of gathering prices than we used in this paper.

Finally, we estimate wholesale prices as 70% of the lowest price for domains in the TLD. We leave a better estimation of this price to future work.

8. VISITS

As an alternative to our registrant-focused analysis, we also analyze the new TLD program from an end user perspective. In particular, we want to know whether actual users visit domains in the new TLDs, and how that compares to similar domains in the old TLDs. We use a domain's presence or absence in the Alexa top million domains as a metric for whether or not users visit it. We do not consider the ranking order as we only care whether or not the domain gets traffic at all.

We begin by splitting new domain registrations from December 2014 into two sets, one for domains in the new TLDs and one for domains in the old TLDs. We find 326,974 registrations in December 2014 in the new TLDs, and 3,461,322 in the old TLDs. We compare these sets with the Alexa top million from April 13, 2014. We use a newer Alexa list to allow the new domain registrations time to develop their Web presence. Due to the order of magnitude size difference between our new registration sets, we report results per hundred thousand new registrations.

	New Per 100,000	*Old* Per 100,000
Alexa 1M	88.1	243
Alexa 10K	0.3	1.1
URIBL	703	331

Table 9: The rate at which new domains in the old and new TLDs appear in blacklists and Alexa. This table only includes domains registered within the same one-month time window to compare old and young TLDs on equal terms.

Table 9 summarizes our results. New domain registrations in the old TLDs are nearly three times more likely to appear in the Alexa top million when compared to registrations in the new TLDs. This ratio is also consistent with appearances on the Alexa top ten thousand. While this is a notable difference, it is also consistent with the proportion of primary registrations described in Section 5.1.

We use a similar method with the URIBL blacklist as an indicator of abusive behavior. We use the same sets of newly registered domains. We use a blacklist contemporaneous with our registration data because blacklist operators add abusive domains as soon as possible. Table 9 summarizes our results.

We find that domains in new TLDs are twice as likely to appear on the URIBL blacklist within the first month. Our data does not reveal why spammers find the new TLDs attractive. However, we can guess based on the registrar pricing data we collected as described in Section 3.7. Domains in new TLDs tended to cost more on average, but individual registrars sometimes sold them for significantly reduced prices. In the extreme we found `xyz` domains for less than $1 USD per year at some registrars.

Table 10 shows the ten TLDs for which a new registration is most likely to appear on a blacklist. Domains registered in December 2014 in most TLDs had less than a 1% chance of appearing on a blacklist in the same month, but the `link`, `red`, and `rocks` TLDs showed significantly higher rates of blacklisting. We found `link` domains for as cheap as $1.50 USD, but `rocks` domains cost at least $7.99 USD. The characteristics of these domains that consistently contribute towards higher rates of abusive behavior remains an open question.

9. CONCLUSION

ICANN greatly expanded the TLD name space to increase consumer choice and to allow more domain registrants to get short and memorable domain names. As we have found in previous TLD expansions [11, 12], new TLDs can increase primary domain registrations but can also lead to speculation and defensive registrations. ICANN's new rapid expansion of the available TLDs gives primary registrants a lot more choice, but also increases the demands on defensive registrants seeking to protect their marks.

We take a comprehensive approach to understanding how registrants use domain names in ICANN's new TLD program. We used data from many sources, including zone file data available to researchers, extensive crawls of Web and DNS information, and public data from ICANN, registries and registrars. We determined that only 15% of domains purchased by a registrant show behavior consistent with primary registrations and that domain parking drives over 30% of registrations in the new gTLD zone files. We use domain pricing information to estimate that only half of all registries have recouped their application fee in wholesale revenue. Similarly, we conservatively estimate that registrants have spent roughly $89 million USD on domain registrations in the new TLDs. Finally,

TLD	New Domains	Blacklisted	Percent
link	4,087	917	22.4%
red	7,599	614	8.1%
rocks	7,191	360	5.0%
tokyo	3,252	40	1.2%
black	919	10	1.1%
club	16,490	173	1.0%
blue	4,971	41	0.8%
support	435	3	0.7%
website	7,876	49	0.6%
country	1,154	7	0.6%

Table 10: The ten most commonly blacklisted TLDs.

we validate the expectation that users visit fewer new domains in new gTLDs than those in old, and that new domains are more than twice as likely to appear on a commonly available blacklist within the first month of registration. Taken together, our findings suggest that new gTLDs, while accruing significant revenue for registrars, have yet to provide value to the Internet community in the same way as legacy TLDs.

Acknowledgments

We would like to thank He Liu for providing the use of his active DNS crawler, and Brian Kantor and Cindy Moore for managing our hardware. Thank you as well to our reviewers for their feedback. This work was supported by National Science Foundation grant NSF-1237264 and by generous research, operational and/or in-kind support from Google, Microsoft, Yahoo, and the UCSD Center for Networked Systems (CNS).

10. REFERENCES

[1] Alexa. `http://www.alexa.com`.

[2] All eyes on Donuts as first new gTLD renewal figures roll in. `http://domainincite.com/18209-all-eyes-on-donuts-as-first-new-gtld-renewal-figures-roll-in`.

[3] S. Alrwais, K. Yuan, E. Alowaisheq, Z. Li, and X. Wang. Understanding the Dark Side of Domain Parking. In *Proceedings of the USENIX Security Symposium*, San Diego, CA, Aug. 2014.

[4] Another new gTLD up for sale with $750,000 reserve. `http://domainincite.com/19021-another-new-gtld-up-for-sale-with-750000-reserve`.

[5] Base Registry Agreement. `https://www.icann.org/resources/pages/registries/registries-agreements-en`.

[6] .com Registry Agreement. `https://www.icann.org/resources/pages/agreement-2012-12-05-en`.

[7] M. Der, L. K. Saul, S. Savage, and G. M. Voelker. Knock It Off: Profiling the Online Storefronts of Counterfeit Merchandise. In *20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2014.

[8] Did XYZ.com pay NetSol $3m to bloat .xyz? `http://domainincite.com/19139-did-xyz-com-pay-netsol-3m-to-bloat-xyz`.

[9] Donuts Launches Domain Namespace Expansion with 307 gTLD Applications. `http://www.donuts.domains/donuts-media/press-`

releases/donuts-launches-domain-namespace-expansion-with-307-gtld-applications.
[10] Donuts Renewal Trends: First Definitive Report. http://www.donuts.domains/donuts-media/blog/donuts-renewal-trends-first-definitive-report.
[11] T. Halvorson, K. Levchenko, S. Savage, and G. M. Voelker. XXXtortion? Inferring Registration Intent in the .XXX TLD. In *Proceedings of the International World Wide Web Conference (WWW)*, Seoul, Korea, Apr. 2014.
[12] T. Halvorson, J. Szurdi, G. Maier, M. Felegyhazi, C. Kreibich, N. Weaver, K. Levchenko, and V. Paxson. The BIZ Top-Level Domain: Ten Years Later. In *Proceedings of the Passive and Active Measurement Conference*, Vienna, Austria, Mar. 2012.
[13] Hyper Text Coffee Pot Control Protocol (HTCPCP/1.0). http://tools.ietf.org/html/rfc2324.
[14] ICANN Whois: Purpose. https://whois.icann.org/en/purpose.
[15] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the IEEE Symposium and Security and Privacy*, pages 431–446, Oakland, CA, May 2011.
[16] My Interview with Daniel Negari Addressing Reported Inflated .xyz Registrations. http://www.ricksblog.com/2014/06/interview-daniel-negari-addressing-inflated-xyz-registrations/.
[17] name.com Pricing for Common TLDs. https://web.archive.org/web/20141128024531/http://www.name.com/pricing.
[18] .net Fees. https://www.icann.org/sites/default/files/tlds/net/net-fees-01feb15-en.pdf.
[19] New gTLD Applicant Guidebook. https://newgtlds.icann.org/en/APPLICANTS/AGB.
[20] New gTLD Current Application Status. https://gtldresult.icann.org/application-result/applicationstatus.
[21] new gTLDs Launches. https://ntldstats.com/launch.
[22] .realtor Fact Sheet. http://www.realtor.org/sites/default/files/handouts-and-brochures/2014/DotREALTOR-Launch-Factsheet.pdf.
[23] .reise to start at $400k in no-reserve auction. http://domainincite.com/17988-reise-to-start-at-400k-in-no-reserve-auction.
[24] Renewal Trends: Day 26. http://www.donuts.domains/donuts-media/blog/renewal-trends-day-26.
[25] Rightside Analyst and Investor Day 2014, slide 104. http://edge.media-server.com/m/p/f9o6abq7.
[26] The Rest of the Story, 2014 Edition. http://www.npr.org/blogs/money/2014/12/31/374225531/episode-595-the-rest-of-the-story-2014-edition.
[27] URIBL. http://uribl.com/about.shtml.
[28] T. Vissers, W. Joosen, and N. Nikiforakis. Parking Sensors: Analyzing and Detecing Parked Domains. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, Feb. 2015.
[29] Whois Policy Review Team Final Report. https://www.icann.org/en/system/files/files/final-report-11may12-en.pdf.
[30] Why Donuts is revealing domain name renewal rates. http://domainnamewire.com/2015/03/31/why-donuts-is-revealing-domain-name-renewal-rates/.