UNIVERSITY OF CALIFORNIA, SAN DIEGO

# Registration Intent in the Domain Name Market

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Computer Science

by

Tristan Halvorson

Committee in charge:

Stefan Savage, Co-Chair
Geoffrey M. Voelker, Co-Chair
Kirill Levchenko
George Papen
Lawrence K. Saul

2015

The Dissertation of Tristan Halvorson is approved and is acceptable in quality and form for publication on microfilm and electronically:

______________________________

______________________________

______________________________

______________________________
Co-Chair

______________________________
Co-Chair

University of California, San Diego

2015

TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

First I'd like to my advisors, Geoff Voelker and Stefan Savage. They taught me a lot about taking a step back and thinking about the motivation and impact of the work I do. Stefan gave me a lot of advice on making relatable and understandable talks, and Geoff really made me understand the importance of making sure each graph is making a point, instead of just representing some data. Both made me feel welcome whenever I needed help or just wanted to chat, and both were (and still are) supportive of our department's graduate student community. Their involvement in both research and random departmental projects make UCSD CSE a great place to be.

The rest of my defense committee members were also very helpful. In particular, Kirill Levchenko felt like an advisor in everything but name for my first few years of grad school. While Geoff and Stefan taught me how to cleanly relate my ideas to other people, Kirill showed me what good research looks like in the first place. He spent a lot of time patiently working with me early in my first two years of graduate school, and his guidance made me self-sufficient on my later projects in a way I would not have been otherwise.

I'd also like to thank my wife Laura for putting up with strange schedules and a fair amount of stress during my graduate student career. I spent a lot of time escaped into work during conference deadlines, but she was both understanding and helpful in making sure I still took care of normal life things. I would also like to thank my family for their support. The physical distance meant I didn't get to visit as often as I'd like, but they were still happy to chat and give advice whenever I needed.

Thank you as well to Brian Kantor and Cindy Moore, both of whom run our group's computing infrastructure. They have done a great job keeping their services running and put up with some complicated requests for my project, including deployment of a new Hadoop cluster.

Getting a Ph.D is a lot of work and took me 6 years, but I met a lot of people along the way that made my time in San Diego enjoyable. Thanks to Nathan Bales, Dimo Bounov, Neha Chachra, Mike Conley, Joe DeBlasio, Louis Devoken, Ian Foster, Danny Huang, Kenny Katayama, David Kohlbrenner, Wilson Lian, Keaton Mowery, Nima Nikzad, Alex Rasmussen, Danny Turner, Bhanu Vattikonda, Jesse Vavrek, and Michael Vrable for being such great people and good friends.

Finally, I'd also like to thank all of my coauthors during my time at San Diego, some from within the department and many from other institutions. Thank you to Neha Chachra, Matthew F. Der, Brandon Enright, Mark Felegyhazi, Ian Foster, Chris Grier, Chris Kanich, Christian Kreibich, Kirill Levchenko, He Liu, Gregor Maier, Damon McCoy, Vern Paxson, Andreas Pitsillidis, Lawrence K. Saul, Stefan Savage, Janos Szurdi, Geoffrey M. Voelker, and Nicholas Weaver.

Chapters 1, 2, and 3, in part, are a reprint of the material as it appears in *Proceedings of the Passive and Active Measurement Workshop*. Tristan Halvorson, Janos Szurdi, Gregor Maier, Mark Felegyhazi, Christian Kreibich, Nicholas Weaver, Kirill Levchenko, and Vern Paxson, Springer, March 2012. The dissertation author was the primary investigator and author of this paper.

Chapters 2 and 4, in part, are a reprint of the material as it appears in *Proceedings of the International World Wide Web Conference*. Tristan Halvorson, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker, International WWW Steering Committee, April 2014. The dissertation author was the primary investigator and author of this paper.

Chapters 1, 2, and 5, in part, have been submitted for publication of the material as it may appear in *Proceedings of the International Measurement Conference*. Tristan Halvorson, Matthew F. Der, Ian Foster, Stefan Savage, Lawrence K. Saul, and Geoffrey M. Voelker, ACM, October 2015. The dissertation author was the primary investigator and author of this paper.

## VITA

| | |
|---|---|
| 2009 | Bachelor of Science in Computer Science<br>University of Wisconsin–Madison |
| 2009–2015 | Research Assistant<br>University of California, San Diego |
| 2012 | Master of Science in Computer Science<br>University of California, San Diego |
| 2015 | Doctor of Philosophy in Computer Science<br>University of California, San Diego |

## PUBLICATIONS

Tristan Halvorson, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. XXXtortion? Inferring Registration Intent in the .XXX TLD. In *Proceedings of the International World Wide Web Conference (WWW)*. Seoul, Korea, April 2014. Tristan Halvorson,

Janos Szurdi, Gregor Maier, Mark Felegyhazi, Christian Kreibich, Nicholas Weaver, Kirill Levchenko, and Vern Paxson. The BIZ Top-Level Domain: Ten Years Later. In *Proceedings of the Passive and Active Measurement Workshop*. Vienna, Austria, March 2012. Chris Kanich, Nicholas Weaver, Damon McCoy, Tristan Halvorson, Christian

Kreibich, Kirill Levchenko, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. Show Me the Money: Characterizing Spam-advertised Revenue. In *Proceedings of the USENIX Security Symposium*. San Francisco, CA, August 2011. Kirill Levchenko, Neha Chachra,

Brandon Enright, Mark Felegyhazi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, Damon McCoy, Andreas Pitsillidis, Nicholas Weaver, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the IEEE Symposium and Security and Privacy*. Oakland, CA, May 2011.

# ABSTRACT OF THE DISSERTATION

Registration Intent in the Domain Name Market

by

Tristan Halvorson

Doctor of Philosophy in Computer Science

University of California, San Diego, 2015

Stefan Savage, Co-Chair
Geoffrey M. Voelker, Co-Chair

Businesses view a good domain name as a key part of their Internet presence. Short, memorable strings in `com` and `net` can sell for thousands or millions of dollars, or may simply be unavailable. To address this problem, ICANN has introduced alternative top-level domains (TLDs): first a few at a time, such as `biz` and `info`, but then hundreds per year starting in 2013. Alternative TLDs give registrants more opportunities to get a name they will like, but give the same chances to name speculators. Additionally, existing companies find themselves with new namespaces in which to defend their trademarks. One would hope most registrants want domain names to develop an Internet presence, and that defensive registrations and name speculation are the exception. Unfortunately, few studies quantify domain name registration types, and none of them at scale.

This dissertation identifies the intent of registrants in hundreds of TLDs. We combine registration intent with user visit data and pricing information to provide a comprehensive view of each TLD in our set. We open the dissertation with two case studies, first of `biz` and then of `xxx`, and quantify the types of registration behavior in each. Then we combine the lessons learned from each to scale our methodology to hundreds of TLDs in ICANN's New gTLD Program.

We find that defensive and speculative domain registrations are extremely common, and that some TLDs seem to encourage them. Domain registrants in `xxx` spent $24 million USD in the first year to defend their names and trademarks, and less than 7% of yearly fees come from primary registrants. By contrast, over 30% of new domain registrations in `com` host legitimate content.

ICANN's New gTLD Program generated new domain registrations, instead of simply shifting them from old TLDs to new ones, but most of them receive low quality registrations. Parked domains make up 32% of registrations in the new TLDs, and free promotional domains cover 12% of the space. Even our most permissive pricing models show that roughly 10% of registries will never become profitable at current registration rates.

# Chapter 1

# Introduction

As a technical system, the Domain Name System (DNS) provides a straightforward service. DNS servers provide a mechanism to look up human-readable names (e.g., www.ucsd.edu) and convert them into routable IP addresses (e.g., 132.239.180.101). The service treats domain names equally, and domain name sellers, known as *registrars*, provide most names on a first-come first-served basis for a nominal yearly registration fee.

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. For example, ToysRUs reportedly bought `toys.com` in 2009 for $5 million USD [59], and we've found unverified but first-hand accounts of sales up to $35 million. The technical operation of DNS may treat domain names equally, but companies pay a premium for short domains, brands, and generic words.

In part for historic reasons, DNS is arranged hierarchically: names associate with a suffix, usually a top-level domain (TLD). The Internet Corporation for Assigned Names and Numbers (ICANN) is the highest level organization that runs the DNS. They do not own the technical infrastructure or manage individual TLDs; instead, they provide contracts to third-party companies, known as *registries*. The original DNS design included generic TLDs (gTLDs, such as `com` and `edu`), although other kinds like country code TLDs have been added since. Of the original TLDs, `com` proved the most open and the best choice for most registrants, becoming the de facto TLD for Web addresses.

To the average user, `com` became synonymous with the Web, its iconic status earning it a place in the Oxford English Dictionary in 1994. As a consequence of its popularity, the `com` landscape quickly became crowded. To ease the pressure on `com`, ICANN moved to create more gTLDs, introducing `biz` and `info` in 2001. 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 are *defensive*, generated by brand or trademark owners trying to protect their names. The success of this endeavor hinged on how both businesses and users would perceive TLDs: whether users accepted the notion of TLDs as simply reflecting different potential homes for various Internet entities, or whether they viewed "dot com" as the sole TLD where they would expect to find prominent Web enterprises. Would users find it confusing to encounter the same subdomain in different TLDs? Would businesses feel compelled to defend their trademarks in new TLDs? Or would the expansion of the DNS namespace provide additional opportunities for new businesses?

ICANN continued to add new TLDs intermittently through 2011, such as `jobs` and `mobi` in 2005, and the controversial `xxx` in 2011. Then 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 program 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.

Even though a lot of time has passed between the introduction of the DNS and ICANN's new program of rapid expansion, few studies have examined how registrants use DNS and how new TLDs will impact current or prospective domain owners; we discuss some of them in Chapter 2. These studies tend to focus on specific types of registrations, and in one or a small handful of TLDs.

## 1.1 Contributions

While industry analysts debated between multiple ways registrants might use alternative TLDs, this dissertation shows that a data-driven approach can empirically measure primary, defensive, and speculative domain registrations, and that all play a significant role in the new TLDs' monetization structures. The revenue split between those registration types differs by TLD, which suggests that better policies from ICANN could significantly reduce the cost overhead to existing trademark and domain holders.

Our approach begins with a list of domains gathered from zone files for each TLD. We actively crawl DNS and Web for each new domain and use this data to classify the content of each domain. We combine this data with publicly available Whois data, user visit information provided by third parties, and registrar pricing. By merging these varied data sets together, we get a complete view of what type of content each domain hosts, as well as information about each owner. We use this content-based classification to estimate the intent of the domain's registrant. We begin our work with two case studies on very different TLDs, and then we combine the lessons we learned from each to generalize our intent inference methodology.

Our first case study focuses on the `biz` TLD, introduced as an alternative to `com` with the explicit goal of relieving `com`'s dense name space. We compare ownership information between domains common to both `biz` and `com` and also evaluate higher-level registration intent. We use this data to show that biz contains many primary registrations with real content, but also that many com domain owners simply duplicate their Web presence in biz. We describe our experiences with this TLD in Chapter 3.

In our second case study we take a look at xxx, intended for adult content. xxx had a contentious introduction process and was denied for inclusion multiple times before ICANN finally approved it. Fewer companies treat xxx as an alternative to com, so we focus less effort on comparing ownership information between TLDs in this case study. Instead, we focus on its first year from the point of view of a consumer. We examined the domain content and registration options available to the consumer and determined that xxx contained proportionally far more defensive registrations than biz, but mostly from companies that did not want others to purchase their names. We present detailed results about this case study in Chapter 4.

Shortly after the conclusion of our second case study, ICANN started delegating the first new TLDs from its new program of rapid expansion. Understanding the ways registrants use domain names can inform both future policy decisions by ICANN and better financial decisions by domain registries about which TLDs to create. The new rollout provides a fantastic opportunity to measure many TLDs at once, but also presents challenges in scalability. We used specialized knowledge in both case studies: for biz, the fact that ICANN introduced it as an alternative to com influenced our methodology, and for xxx we rely on specific page templates to classify some Web content, such as those for reserved domains. In Chapter 5, we show our solutions to these scaling challenges and also provide results for hundreds of the new TLDs. We find that the new gTLDs mostly include undesirable registrations, such as domain parking and defensive domains. Additionally, registrations in the new TLDs receive disproportionately fewer visits compared to new registrations in the older TLDs.

Finally, as part of our analysis of the new gTLD program, we also investigated registry revenue streams and expenses. We use this information to build a pricing model that describes which registries become profitable. This data is heavily related to our other work with the new gTLD program, so Chapter 5 also includes this data and discussion of the resulting pricing models. We find that 15% of new gTLDs will never be profitable at current registration rates, even with the least conservative models.

This thesis directly provides data on how people use domain names, and looks at many TLDs to explore the impact of delegating new ones. We find that new TLDs provide little benefit to primary domain registrants. Instead, these new TLDs generate name speculation and cause companies to defensively register their names and trademarks. By the end of the 2015 fiscal year, domain name sales for the new program reached only 18% of ICANN's estimates [41], so ICANN and its registries and registrars did not meet their expected revenues and in many cases could not cover expenses (Section 5.6). Assuming ICANN wants to consider and improve the needs of registrants when setting domain name policy, our data strongly suggests against the delegation of additional TLDs.

Chapter 1, in part, is a reprint of the material as it appears in *Proceedings of the Passive and Active Measurement Workshop*. Tristan Halvorson, Janos Szurdi, Gregor Maier, Mark Felegyhazi, Christian Kreibich, Nicholas Weaver, Kirill Levchenko, and Vern Paxson, Springer, March 2012. The dissertation author was the primary investigator and author of this paper.

Chapter 1, in part, has been submitted for publication of the material as it may appear in *Proceedings of the International Measurement Conference*. Tristan Halvorson, Matthew F. Der, Ian Foster, Stefan Savage, Lawrence K. Saul, and Geoffrey M. Voelker, ACM, October 2015. The dissertation author was the primary investigator and author of this paper.

# Chapter 2

# Background

DNS treats all names equally at the technical level, but different names accrue significantly different value. Since names are human readable, by design, certain words more naturally resonate with particular commercial categories (e.g., books.com) and others alias existing trademarks (e.g., mcdonalds.com). This aspect, combined with the "first come, first served" model of domain registration, has effectively turned domains into property rights with the attendant problems thereof. Thus, some popular domains can re-sell for millions of dollars [56], and a whole cadre of "domainers" purchase choice domains speculatively with no intent to use them. Still others, in a practice called "typo-squatting", register domains that are lexically similar to popular domains to poach their traffic from users who mistype [32]. In response to these behaviors, ICANN introduced the Uniform Domain-Name Dispute-Resolution Policy (UDRP) in 1999 to resolve domain disputes between trademark holders and domain name speculators through non-binding arbitration. However, the UDRP process can be both time consuming (six weeks or more) and costly (typically $1,000 plus any attorney's fees) and thus many brand holders register domains on a prophylactic basis simply to minimize risk.

In this chapter, we provide background on the DNS, new TLDs, and registration policy. Section 2.1 identifies the major players in the domain name business and their relationships between each other. Section 2.2 explains ICANN's role in introducing new

Top-Level Domains (TLDs). Our earliest study was not the first to examine the kinds of registrations in these alternative TLDs, and we describe some previous reports on domain name usage in Section 2.3. Finally, we more precisely define the goals of our studies and give a brief overview of registration intent in Section 2.4.

## 2.1 DNS as a Business Model

DNS was introduced in 1983, and for a short time its management was relatively ad-hoc. In 1988, the United States Department of Defense funded the Internet Assigned Numbers Authority, which officially took responsibility for managing DNS. This responsibility shifted once more in 1998 to the newly-created Internet Corporation for Assigned Names and Numbers (ICANN), this time funded through the U.S. Department of Commerce, and ICANN still manages the DNS today. As part of its duties, ICANN has "the authority to manage and perform a specific set of functions related to coordination of the domain name system, including the authority necessary to . . . oversee policy for determining the circumstances under which new TLDs are added to the root system" [36]. Since then the United States has given up control of ICANN altogether [62], and ICANN does its best to attend to the needs of the international community, not just those of the United States.

ICANN is responsible for managing the high-level direction of the DNS, but otherwise performs only a few technical functions (such as running one of the thirteen root DNS servers). ICANN provides contracts to third-party companies to manage nearly all of the DNS technical infrastructure.

The DNS is organized hierarchically, and its business architecture matches the technical in this respect. Companies known as domain name *registries* manage individual TLDs. For instance, Verisign runs `com`, which includes managing the policies for its subdomains within the bounds of its contract and operation of its technical infrastructure.

Companies that wish to become a registry and run a new TLD must provide ICANN with a proposal which details why introducing the TLD is a net positive for the Internet community and demonstrates their ability to operate the technical infrastructure. ICANN may choose to award the contract or not, but has a very specific set of rules it must follow when choosing to do so and must be transparent about the entire process.

Registries do not directly sell domain names. Instead, ICANN provides a separate set of contracts to companies that wish to do so, known as *registrars*. Historically ICANN would not accredit a single company as both a registry and a registrar, but in 2010 the ICANN board reconsidered, and now individual companies may perform both roles. Some of the most common registrars include GoDaddy, Name.com, eNom, and Network Solutions.

When a domain name registrant purchases a domain name, portions of that payment go to each of these actors. ICANN gets a flat amount per domain name sale, but only for registries whose registrations meet a predetermined volume. The registry operator gets a flat wholesale price. To increase competition, ICANN requires registries to provide the same wholesale price to all registrars. Registrations in some TLD types, such as sponsored TLDs, may include additional fees that go to the sponsoring organization. For instance, xxx registrations include a $10 fee for the International Foundation for Online Responsibility (IFFOR), which is responsible for ensuring xxx registrations go to members of the adult entertainment industry. The remaining balance goes to the registrar.

Registrars set the registration costs of domain names, not registries, which means domain names within a TLD do not have a uniform price. In most cases domain names within a TLD will not differ by a large amount, since the underlying fees match for all registries. However, domain name registrations require recurring payments (usually every year), and registrars typically sell domains for many different TLDs, so the instantaneous pricing for domains in a particular TLD is not always rational in the short term. For instance, a registrar may provide one-year registrations at half the wholesale price in some TLDs, losing money on the intial registration, in the hopes that the customer later renews the domain at the full rate.

## 2.2 Introducing New TLDs

For most of its existence, the DNS has had a small number of generic TLDs (gTLDs), and a larger set of country code TLDs (ccTLDs). Most TLDs had a specific purpose. For instance, `com` was open for all registrations, but ccTLDs allowed countries to operate their own local DNS trees, providing independent sovereign control on part of the name space.

In 1999 ICANN formed the Domain Name Supporting Organization (DNSO), an advisory body within ICANN, to handle matters concerning the Domain Name System [24], chartering, within the DNSO, Working Group C to study the issues surrounding the formation of new generic top-level domains (gTLDs). The Working Group's task was to gauge the need for new gTLDs, and, if deemed required, to determine what should be their nature and deployment policy [10]. In March 2000 Working Group C released its final report, addressing these questions [64]. The ICANN board adopted its recommendations in July 2000 [23].

The Working Group C report [64] addressed two important questions: whether to create new gTLDs, and, if so, how to introduce them. On the need for new gTLDs, the report argued: "Expanding the number of TLDs will increase consumer choice, and create opportunities for entities that have been shut out under the current name structure." Moreover, the report observes:

> Existing second-level domain names under the .com TLD routinely change hands for enormously inflated prices. These are legitimate trades of ordinary, untrademarked words; their high prices reflect the artificial scarcity of common names in existing gTLDs, and the premium on `.com` names

> in particular.
>
> . . .
>
> If the name space is expanded, companies will be able to get easy-to-remember domain names more easily, and the entry barriers to successful participation in electronic commerce will be lowered. *Addition of new TLDs will allow different companies to have the same second-level domain name in different TLDs* (emphasis added).

Arguing against the consensus position of the report, some members of the Working Group suggested that "an increase in the number of top-level domains could confuse consumers," and that expanding the domain space "will likely increase trademark owners' policing costs and the costs of defensive registrations." Members expressed concerns about "*trademark holders simply duplicating their existing domains*" (emphasis added).

On the nature of the initial rollout, the report weighed two opposing approaches: authorizing "hundreds of new TLDs over the course of the next few years" and introducing new TLDs "slowly and in a controlled manner, and only after effective trademark protection mechanisms had been implemented and shown to be effective." In arguing for the introduction of many new TLDs, some members argued that "a small number of new gTLDs with no commitment to add more . . . would encourage pre-emptive and speculative registrations based on the possibility of continued artificial scarcity". The report ultimately proposed "deploying six to ten new TLDs."

On May 15, 2001, ICANN introduced the new `biz` and `info` gTLDs. The `biz` TLD in particular was intended as an alternative to `com` to increase choice in the face of second-level string scarcity. We describe the introduction of `biz` in more detail in Chapter 3, and follow it up with our study of long-term `biz` usage. Our goal is to measure how well ICANN's earliest new TLDs dealt with these issues, or if the defensive and speculative behavior became common as feared.

Through 2011, ICANN introduced TLDs in a controlled process that followed from the result of the Working Group C report. We describe the introduction of the last of these domains, xxx, in Section 4.1. Then in 2013, ICANN's New gTLD Program took effect. New TLDs are no longer the subject of intense individual scrutiny in the same way as biz, info, and xxx, but instead must go through a simplified and standard process. TLD applications can still take years in the new program, but the hurdles are not nearly as high. Since then ICANN has delegated hundreds of new TLDs to the root zone, more than doubling its size, and hundreds more applications await in the pipeline.

The giant shift in the delegation process allows us to compare the effects of both scenarios laid out by the Working Group C report; few TLDs at a time in the earlier ones, and many TLDs at once in the new program. By analyzing the types of registrations in all of these TLDs, we can learn how the TLD deployment models compare and if either or both provides real benefits to the Internet community.

Our work also investigates whether some TLDs extort brandholders for defensive registrations through predatory pricing schemes. For example, Vox Populi Registry Inc. proposed the sucks TLD as part of the new program, a string to which ICANN's Government Advisory Committee (GAC) filed an early warning [14].[1] They state that "The string (.sucks) has an overtly negative or critical connotation. As a result, many individuals, businesses and organisations may seek to protect their brands or reputations in this TLD." Additionally, they ask for the registry to "detail appropriate mechanisms to limit the need for defensive registrations."

ICANN delegated the TLD, and then in March 2015, Vox Populi announced their pricing plans: a suggested sunrise price of $2,499 USD per domain [55], and general availability at $249 [46]. ICANN sent a letter to the Federal Trade Commission (FTC), forwarding concerns about Vox Populi's exploitive pricing and asking if the FTC felt they

[1] An early warning does not mean anything official, but indicates that later in the application process when ICANN asks for the GAC's advice, they plan to express some concerns. The early warning gives the registry ample time to address those problems.

should step in. In their response [27], the FTC urged ICANN to "consider whether the current rights protection mechanisms adequately perform their intended function." They also asked ICANN to address consumer protection concerns at a wider scale, because they felt that "other registries have also engaged in troubling tactics."

Our research most comprehensively looks at xxx in terms of its extortive pricing. Unfortunately, sucks did not enter general availability until after the conclusion of our study. However, the methodology proposed in this thesis could lend itself well to identifying these sorts of behaviors and quantifying their monetary impact on brandholders.

## 2.3 New TLD Usage

Not surprisingly, custom and convention have made some TLDs more popular than others. For example, com had 108 million domains registered as of October 2012, including 57% of the Alexa Top 100,000 Web sites. In contrast, biz had only 2.3 million domains at the same time, and covers only 0.3% of the Alexa ranks. Since different TLDs serve different purposes, the sheer number of registrations is not an ideal metric of success. For example, since edu only serves educational institutions it is artificially constrained in size. Yet it serves its function effectively and is widely supported by its stated community. Instead, another way to evaluate the "health" of a TLD is the extent to which its registrations host unique content (under the assumption that the names are therefore being used to provide direct access to this information). A TLD that instead has a high fraction of speculative or defensive registrations offers much less value to the Internet community. However, crisply determining intent can be difficult in practice.

ICANN has, however, commissioned several reports to address this concern. In 2004, Summit Strategies International produced a report on the issues encountered during introduction of the new gTLDs [22]. Three questions in particular are of interest to us in the context of this work: How effective have start-up mechanisms been in protecting trademark owners against cybersquatting and other abusive registrations? How often and how successfully have advance filtering and other mechanisms for enforcement of registration restrictions been used, both in sponsored gTLDs and in restricted unsponsored gTLDs? What effect have the new gTLDs had on the scope and competitiveness of the domain name market, in terms of opening new markets, and in their effect on existing TLDs and registrants?

With respect to these questions, the report found that "the new gTLD start-up periods proved generally effective at protecting the interests of trademark holders." Regarding registration restrictions, "random sampling indicated fewer problems than expected in `.biz`, with 1.8% of the registrations appearing to fail to satisfy the criteria and another 9.6% being unclear." Finally, on the question of competition, the report is more equivocal:

> The new gTLDs have introduced some competition, but how much is debatable. Examining market share, extent of actual choice and price elasticity suggests that impact has been minimal. Other evidence, however, indicates that TLD expansion has attracted about 20% new registrants and led to new uses among 40–60% of registrants.

In its recent move to open TLD registration to the general public, ICANN commissioned several reports analyzing the economic consequences of their new initiative. Among other concerns, the reports address the danger that new gTLDs will compel trademark holders to defensively register their marks in each new TLD. The 2009 Carlton report [7] dismisses this as a problem, arguing that "many registrations that 'redirect traffic' to other sites serve productive purposes of attracting and retaining Internet traffic, not merely to prevent cybersquatting." Furthermore, the report argues: "While some of the registrations for domain names under the new gTLDs may have been made for defensive purposes, the limited number of registrations for new gTLDs indicates that the vast majority of `.com` registrants did not find a compelling reason to undertake defensive registrations in the new gTLDs."

The 2010 Katz, Rosston, Sullivan report [26] looked back at past TLD introductions and found that "the competitive benefits of additional TLDs may not be large." They argue that `com` has remained dominant and that no alternative TLD shows any sign of impacting its high registration volume, and that "registrants with well-established domain names are probably unlikely to switch to a new gTLD." The authors also express concerns about defensive registrations, but state that "mechanisms for protection may mitigate this problem."

These studies raise many valid concerns about introducing additional TLDs, including the types of registrations they will contain and the impact of these registrations on existing TLDs. However, they also have limitations, such as speculation, conflicting data, and data sets as small as 100 domains. Our studies attempt to answer these same questions, but in a more comprehensive manner. Why do people register domain names? Do they want to defend a trademark or show real content? Do rights protection mechanisms discourage defensive registrations and save brandholders money? Can we measure this at scale for all domain names in as many TLDs as possible?

## 2.4 Registration Intent

One of the most significant contributions of this dissertation is our investigation of why people register domain names. To this end, we need some way to classify the intent of domain registrants. We classify registration intents into one of the following three categories.

**Primary** *Primary* domains identify a company, product, service, or organization, either publicly or internally. In other words, the registrant *actually uses* primary domains. For example, NeuStar, Inc. and UC San Diego use `neustar.biz` and `ucsd.edu`, respectively, as the primary domain names by which they identify themselves on the Internet.

**Defensive** A registrant uses a domain registration only to *defend* a name while not actively employing the domain to identify itself, its service, or network resources. Examples of defensive registrations are `google.biz` and `gooogle.com`, both of which redirect to `google.com`.

**Parked** Registrants *park* domains with the purpose of reselling them or generating advertising revenue from accidental user visits to the site.

When introducing new TLDs, ICANN needs to balance the needs of the potential registry, new registrants, the existing DNS community, and everyday Internet users. Registries, however, are private companies and are naturally interested in making a profit. Classifying domains by registrant intent in this way reveals which kinds of registrations provide the most revenue for the registry. In Chapters 4 and 5, we pay particular attention to the non-standard registration options, such as trademark protection and name reservation, to better assign revenue information to domains in each intent category.

Finally, not all domain names fit cleanly into one of these registrant intent categories, so in some cases we have an additional category or remove some domains from classification. For example, domains given to registrants for free as part of a promotion do not provide real information on why people buy domain names, as the owner may not have been willing to spend money on the name if the promotion did not exist. We provide explanations for these specialized categories when they appear in future chapters.

Chapter 2, in part, is a reprint of the material as it appears in *Proceedings of the Passive and Active Measurement Workshop*. Tristan Halvorson, Janos Szurdi, Gregor Maier, Mark Felegyhazi, Christian Kreibich, Nicholas Weaver, Kirill Levchenko, and

Vern Paxson, Springer, March 2012. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is a reprint of the material as it appears in *Proceedings of the International World Wide Web Conference 2014*. Tristan Halvorson, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, has been submitted for publication of the material as it may appear in *Proceedings of the International Measurement Conference*. Tristan Halvorson, Matthew F. Der, Ian Foster, Stefan Savage, Lawrence K. Saul, and Geoffrey M. Voelker, ACM, October 2015. The dissertation author was the primary investigator and author of this paper.

# Chapter 3

# biz Ten Years Later

On May 15, 2001 ICANN announced the introduction of the biz and info generic top-level domains (gTLDs)—the first new gTLDs since the inception of the Domain Name System—aiming to "increase consumer choice and create opportunities for entities that have been shut out under the current name structure." The biz gTLD, in particular, was to become an alternative to the popular com top-level domain.

In this chapter we begin our domain name usage analysis with a case study of the biz gTLD. We define a simple registration intent methodology and use it to determine whether biz has evolved into the role intended by ICANN, and whether early concerns about defensive or speculative behavior were justified. Using DNS zone files, DNS probing, and Web crawler data, we investigate whether biz has become a viable alternative to com, giving trademark holders who find themselves unable to register a com name an attractive alternative, or if it has merely induced defensive registrations by existing trademark holders who already had equivalent com domains. We find that a quarter of all biz domains are parked, and another 24% show evidence of defensive behavior. Additionally, our findings show that biz domains do not share the same popularity as other TLDs like com, based on public sources of Web visit rates.

## 3.1 Introduction

Following the adoption of the Working Group C report in July 2000, ICANN posted a public call for applications for new gTLDs. Seven proposals were ultimately chosen in November 2000. On May 15, 2001 ICANN announced agreements for the operation of the `biz` and `info` gTLDs—the first new gTLDs since the inception of DNS—with NeuLevel and Afilias, respectively [25]. NeuLevel received applications for roughly 280,000 domains during the initial preliminary registration "Land Rush" period from May 25 to September 25; on October 7, 2001, the `biz` gTLD became operational.

### 3.1.1 NeuLevel Proof of Concept Report

In 2004 NeuLevel produced a report for ICANN detailing the `biz` startup [38]. As part of the `biz` launch process, trademark owners were given the option of filing a "Trademark Claim Form," which "was designed to help companies protect their trademarks and service marks during the launch of the `.biz` TLD by enabling these companies to stake claims to domain names prior to the commencement of service." The $90 filing fee allowed the claimant to challenge an applicant for a domain containing the trademark under the "Start-up Trademark Opposition Policy," a dispute resolution policy under which the claimant would need to prove rights to the trademark. NeuLevel received 80,008 Trademark Claim Forms, collecting $6.3 million in fees.

The report also examined a sample of 100 `biz` domains for evidence of registration of the same names under different TLDs. Their survey found that for 85% of `biz` registrations, the same name existed in `com` (half of which were by the same registrant).

### 3.1.2 Previous Analysis of `biz`

In 2002 Zittrain and Edelman did a survey of `biz` registrations [73]. They found that 90.5% of names registered in `biz` also occur in `com` (consistent with our finding nine years later). To assess whether `biz` and `com` registrants overlapped, the survey authors examined Whois records, comparing records based on registrant postal code, email address, and name server second-level domains. They found that 35.4% of `biz-com` pairs matched in at least one attribute, 25.8% in two, and 12.9% in all three.

2011 marked ten years from the introduction of `biz` and nine years from the Zittrain and Edelman report. ICANN was in the middle of approving the new `xxx` TLD, and was also working on its rollout of the New gTLD Program. The time was right to perform a new study of `biz`, to evaluate the name ecosystem well after the TLD's introduction. Were the large number of defensive registrations an artifact of the rollout mechanisms used by NeuLevel, or did the defensive behavior they saw have a lasting effect? What did this mean for newer TLDs?

### 3.1.3 The `biz` Case Study

This chapter takes stock of the `biz` gTLD, which NeuLevel largely promoted as an alternative to `com`. We find that over 20% of domains in both `biz` and `com` are *parked*, which makes the frequency of parking in `biz` more than in `com`, contrary to the original intention to avoid domain speculation in `biz`. Furthermore, between 10% and 25% of `biz` registrations appear only to exist to defend against name infringement.

The rest of this chapter proceeds as follows. Section 3.2 describes our data collection methodology, followed by our analysis in Section 3.3, and a brief discussion of our findings in Section 3.4. Section 3.5 concludes the chapter.

## 3.2 Data & Methodology

This chapter answers two basic questions: how do owners of `biz` domains use their domains, and, if in active use, whether the `biz` domain forms the *primary* domain of the registrant or whether it merely *defends* one registered under another gTLD. To start, we obtained the biz and com zone files, dated June 27, 2011, to coincide with the 10-year anniversary of the biz TLD's addition to the authoritative root server. We use three sets of domain names: all 2.1 million biz domains, their 2 million (94%) com namesakes, and a random sample of 2 million com domains. We rely on four sources of data to classify domains: zone files, active DNS queries, Whois registration records, and Web content.

### 3.2.1 Zone Files

A zone file contains the DNS records used by a name server, typically in BIND format. We obtained the zone files for the biz and com gTLDs from their respective registries. We used the zone files to get the list of biz and com domains and determined which domains were registered in both gTLDs. We also gathered their name server information for the DNS crawler.

### 3.2.2 Whois

We retrieved the whois registration information for each biz–com pair in our data set. Since whois records consist of free-form text, we use a customized version of phpwhois [44] to parse the whois records and extract the domain registrant (owner) information. Many registrars have limits on the number of whois queries they will answer. Additionally, phpwhois could not parse all whois entries. Due to these limitations, we could only extract registrant information from 65% of our biz–com pairs.

To assess whether the biz domain and its com namesake share the same owner, we compare the registrant information returned by phpwhois. We first exclude domains that use whois privacy mechanisms (e.g., the Domains by Proxy service). We compute the Levenshtein distance between both domains for each of: the registrant's name, e-mail address, phone, and fax number. We mark each of these fields as missing if they are absent or less than 5 characters long in either domain. We mark a field as a match if present and the Levenshtein distance between biz and com does not exceed 2 (requiring an exact match does not significantly alter our results). We also mark registrant names as a match if the name from one domain forms a substring of the other domain.

We consider two whois records a strong match if at least two of the four categories match and at most one category is missing. We consider them a weak match if any of the four categories match. We use both types of matches, but differentiate between them in our analysis.

We note that for a significant number of com domains, the whois record we retrieved only contained the registrant name but no further fields. We therefore cannot have a strong match for any of these domains.

### 3.2.3 DNS Probing

We queried the DNS records for a list of all biz domains and their com counterparts, as well as the randomly selected set of com domains. For each name, we queried (starting at the root) to find the authoritative name servers for the biz and com versions of the domain. We performed the crawling with a custom Python library on September 12th and 13th, 2011.

### 3.2.4 Web Crawl

We collected the content of the Web pages belonging to the registered domain in our data sets. First, we downloaded the pages for the domains in the biz zone file, e.g., foo.biz. Then, we crawled the corresponding com domain foo.com to check the registration purpose for the biz domain. When downloading the Web pages, we recorded the HTTP status codes for success, redirection, errors and other standardized events. We also recorded unknown errors. Note that at times we could not retrieve the Web pages, either because the domains' owners want to serve no Web content, or due to the time interval (several months) between the zone file creation and our active Web crawling.

### 3.2.5 Content Classification

To identify parked domains, we built a simple classifier that searches for a set of regular expressions in the downloaded content. We created highly specific patterns to match templates for the largest known parking sites. We relied on unique features of the page, such as JavaScript libraries or image servers used by the parked pages.

## 3.3 Analysis

In Section 2.3, we described our goal to understand the effects of introducing a new gTLD: whether it would lead to "trademark holders simply duplicating their existing domains" or "will allow different companies to have the same second-level domain name in different TLDs"; whether it "would encourage pre-emptive and speculative registrations based on the possibility of continued artificial scarcity" [64]. To answer these questions, we first group biz domains into three functional categories:

Recall from Section 2.4 that defensive registrations prevent another party from misrepresenting itself as the registrant or from simply capturing traffic (intended for the registrant) for advertising purposes. A defensively registered domain is one not used by the registrant to name and identify products, services, or network infrastructure.[1]

Although it is nearly impossible to divine the registrant's intention with absolute certainty, certain network-visible characteristics of a domain serve as indicators of primary or defensive use. In particular, we consider domain ownership, Web content, and hosting infrastructure sharing as indicators of primary or secondary use. Our results are summarized in Figure 3.1.

[1] The difference between a defensive registration and either cybersquatting or typosquatting (registering misspellings of popular brands) lies in the identity of the registrant: when the registrant also owns the intended brand name or trademark, the registration is defensive; when the registrant is a third party with no

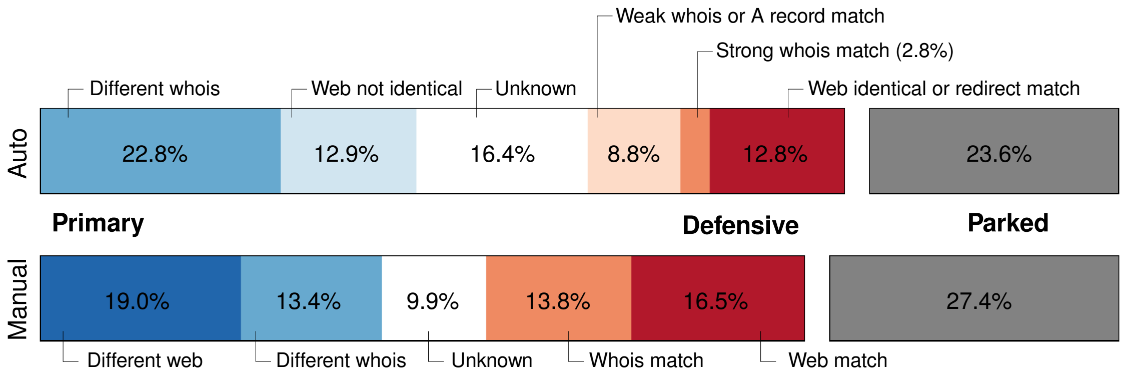


**Figure 3.1.** Disposition of `biz` domain names with respect to their `com` namesakes based on our automatic classification (top) and manual classification of a sample of 485 domains (bottom). Excludes 6.2% of `biz` domains without a `com` namesake.

In the first distribution of Figure 3.1, we show each characteristic we use, whether it indicates a primary or a defensive registration, and the strength of the classification. We first include a domain in the "parked" category if its Web content matches a known parking template (see Section 3.3.1). Otherwise, we use its strongest characteristic to classify the domain. We group the classifiers into "very confident", "somewhat confident", or "not confident", represented in the figure by the darkness of the bands.

In addition to the automatic classification above, we hand-classified 485 domains chosen at random from the 2 million names in the `biz` zone which have a namesake in `com`. For each of these domains, we compared their whois records. Independently, we also classified the `biz` Web content as parked, legitimate, or unavailable, and then compared legitimate content to `com` to determine its intent. The bottom distribution in Figure 3.1 shows the results of this manual classification.

### 3.3.1 Parked Domains

A parked domain is not actively used by the registrant, and does not represent a name or brand used by the registrant. Registrants typically hold parked domains with the

legitimate claim to the name, the registration constitutes cybersquatting or typosquatting.

**Table 3.1.** The Web behavior of domains in the biz and com gTLDs. The biz column shows statistics for the 2.1 million domains in the biz TLD, the com column for a random sample of 2 million com domains.

| *Category* | *biz* | *com* |
|---|---|---|
| No server | 23.5% | 17.4% |
| HTTP Error | 3.4% | 3.3% |
| Parked | 22.8% | 19.4% |
| Redirect | 18.5% | 17.3% |
| On-site | 5.1% | 8.5% |
| Namesake | 4.1% | 0.4% |
| Other site | 9.1% | 8.7% |
| Content served | 31.7% | 39.9% |
| Same as com | 3.0% | — |
| Distinct | 27.7% | — |

intention of selling them at a profit or monetizing accidental Web traffic through advertising. Parked domains are easily identified by prominent advertising on the domain's site that one may purchase the domain and usually includes additional advertising.

We rely on the Web content hosted at the domain as our primary indicator of a parked domain. Table 3.1 shows the proportion of parked sites in the biz and com gTLDs. Figure 3.1 shows the number of parked biz domains (23.6%) having a com namesake.

### 3.3.2 Identical Web Content

Owners of defensively registered domains frequently reroute all Web traffic to the intended (primary) domain, usually via HTTP redirection [12]. Because the browser actively follows HTTP redirects, this method has the advantage of changing the user-visible address bar to reflect the new address (the target of the redirect). The user thus sees the *correct* address, consistent with the branding of the site.

Table 3.1 includes statistics about this mechanism: 18.5% of biz and 17.3% of com domains host a Web server that redirects the user. However, 4.1% of biz sites redirect to a site hosted at the same domain name in a different TLD, compared to 0.4%

**Table 3.2.** Comparison of registrants of biz domains and their com namesakes using Whois records, showing absolute and relative number of biz–com name pairs in each category. Rightmost column shows value relative to total number of name pairs (93.8% of biz names).

| *Category* | *Abs* | *Rel* |
|---|---|---|
| Unknown | 693,393 | 35.1% |
| Privacy guard | 281,417 | 14.2% |
| biz only | 97,802 | 5.0% |
| com only | 82,161 | 4.2% |
| both | 101,454 | 5.1% |
| Match | 424,683 | 21.5% |
| Weak | 308,337 | 15.6% |
| Strong | 116,346 | 5.9% |
| No match | 573,388 | 29.1% |

for com. We also see defensive registrations where both the biz and com Web servers redirect a user to the same third domain. In our set of domains in both biz and com, 6.1% have identical redirects.

Despite the effectiveness of HTTP redirects and their advantage of "correcting" the user, site operators also may simply maintain identical Web sites under both domains. To detect this condition, we compared the content of each biz site to that of its com namesake. Upon examination, we found that 3.0% of non-parked sites did indeed serve the same content. We only classify pages as identical if the source matches exactly, meaning a match almost certainly indicates a defensive registration, but a mismatch only weakly indicates a primary one.

While identical content serves as a good indicator of a defensive registration, differing content does not provide our classifier with a good indicator of a primary registration. Different Web content does not necessarily mean the page really differs; some pages include the domain name itself in content or URLs, and many storefront pages will shuffle the order of the items on their front page, and our conservative automatic classification does not account for this possibility. Manually and visually classifying the Web pages, however, is resilient to these issues: we can more confidently classify primary registrations in the manual data set.

As shown in the dark red portion of Figure 3.1, 12.8% of `biz–com` domains have identical Web content or redirects, a strong indicator of a defensive registration. 12.9% of `biz–com` domains have different Web content and no stronger classifiers, and so are weakly classified as primary registrations.

### 3.3.3 Common Registrant

The identity of the registrant provides another classification feature. One registrant owning both a `biz` domain as well as its `com` namesake likely suggests defensive registration. To identify such cases, we extracted registrant information from publicly available Whois records, as described in Section 3.2.2.

We could retrieve and successfully parse both Whois records for 65% of all `biz–com` pairs.[2] Of these 65%, 10.1% of `biz` domains and 9.3% of `com` domains showed some manner of "privacy protection" mechanism, blocking the registrant information from appearing in the Whois record and leaving 50.6% of all `biz–com` pairs that could potentially match.

We grouped these pairs into three categories based on the degree to which we believed we identified the same registrants: weak matches, strong matches, and no match. Using the methodology described in Section 3.2.2, we determine 5.9% of pairs a strong match, another 15.6% a weak match, and 29.1% unlikely to be the same registrant. (Put another way, we found at least some degree of a match for around 40% of the pairs we could assess.)

[2] The delegated nature of the `com` Whois system means that these 65% necessarily constitute a biased sample, because being able to retrieve and parse a given Whois record depends on the registrar, specifically on their query rate limitations and record formatting.

As shown in Figure 3.1, we consider whois data to be more reliable than common hosting or different Web content, since either of those may be incidental. We consider it to be less reliable than HTTP redirects and identical Web content, since those are strong indicators of a defensive registration. After using stronger indicators, we classify 11.6% of biz domains as defensive based on whois and A record data (see Section 3.3.4) and 22.8% as likely primary.

### 3.3.4 Shared Infrastructure

We also used DNS crawling to observe infrastructure sharing between a biz domain and its com counterpart. In particular, we used CNAMEs and common A records as evidence of defensive registrations.

We identify a CNAME match in two different cases: first, when domains in both biz and com have CNAMEs pointing to the same domain; second, when the domain in biz has a CNAME pointing to its com namesake, or vice versa. Of the 2 million domains in both biz and com, 32,431 (1.6%) show common CNAMEs, which demonstrates a clear relationship between the two domains. We include these in the “redirect match” category in Figure 3.1.

We see many more domains with common A records. Our crawler observed common A records in 439,890 domains (22%) between biz and their com counterparts. We see three plausible explanations for such sharing: first, defensive registrations; second, coincidental common hosting, with unrelated owners of the biz and com employing common hosting infrastructure; third and finally, parking the biz and com domains in the same domain parking infrastructure. Since we classify parked domains first and only distinguish between primary and defensive registrations after considering all parked domains, we can ignore the third case.

While common A records do suggest a defensive registration, we cannot reliably

**Table 3.3.** TLD frequency in the Alexa listings and the Open Directory Project. In the Alexa 1,000,000, biz ranks (in frequency of occurrence) between com.cn and ir, while in the ODP, it falls between cat and za. Only one biz domain resides in the Alexa 500.

| TLD | Alexa 1M | Alexa 500 | ODP |
|---|---:|---:|---:|
| com | 55.3% | 64.6% | 41.7% |
| net | 6.26% | 4.60% | 3.74% |
| org | 4.01% | 2.80% | 9.00% |
| ru | 3.75% | 2.40% | 1.46% |
| de | 3.70% | 1.40% | 9.33% |
| info | 1.82% | 0% | 0.480% |
| biz | 0.396% | 0.200% | 0.188% |

distinguish true sharing and common hosting. Because of this, we consider common A records weak evidence of a defensive registration.

## 3.4 Discussion

Our analysis finds 22.8% of biz domains *parked* with a known parking service. We can with certainty classify another 12.8% of biz domains as *defensive* registrations, leaving two thirds undetermined. At least 27.7% of these served some kind of content (excluding cases where this content proved identical to the com namesake). In addition, of the pairs for which we could assess non-private registrant information, we found at least a degree of match between the biz and com registration in 40% of the instances (Chapter 3.3.3), indicating a substantial level of registrations likely made defensively.

To get at the fundamental value (to registrants) of the biz TLD, we can approach the question from the other direction: how many biz domains do registrants use actively? We assessed the popularity of biz domains in the Alexa [1] Web site rankings, as well as the popularity of biz domains in the Open Directory Project [43]. We show the results in Table 3.3, along with other common TLDs.

Our results show that biz domains appear in these sources disproportionately less often than com domains. The com zone is about 46 times larger than biz, but biz domains appear 140 times less frequently in the Alexa 1 million, 323 times less frequently in the Alexa 500[3], and 218 times less frequently in the Open Directory Project. This data suggests a disproportionally lower popularity of biz domains compared to com.

## 3.5 Conclusion

In this chapter we examined the current state of the biz TLD on its ten-year anniversary. We found that in many respects, most notably in the prevalence of domain speculation (parking), biz resembles com. And while one could conclude that it has failed to rival com, biz did extract defensive registrations from existing domain owners. Although registering these domains costs no more than $10 each, this cost is dwarfed by the additional costs of defending trademarks (via resolution procedures and litigation) in a new TLD.

Many people and organizations hold primary registrations in biz, and its introduction added value to the domain name system. However, roughly as many registrants defended a trademark instead of providing content. Such a significant number of defensive registrations in one of the earliest alternative TLDs should have served as a warning of the dangers of introducing a new TLD, and one could expect ICANN to consider better ways to address this type of registration in future TLDs. In the next chapter, we show how defensive registrations impacted the introduction of xxx and how effective the new rollout process was.

Chapter 3 was supported in part by the National Science Foundation under grants CNS-0905631 and CNS-1111672; by the Office of Naval Research under MURI Grant No. N000140911081; and a fellowship within the postdoctoral program of the German Academic Exchange Service (DAAD). It was also connected to the scientific program of

[3] Only one biz domain appears in the Alexa 500, for the blogging site livedoor.biz

the “Development of quality-oriented and harmonized R+D+I strategy and functional model at BME” project through the New Széchenyi Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

Chapter 3, in part, is a reprint of the material as it appears in *Proceedings of the Passive and Active Measurement Workshop*. Tristan Halvorson, Janos Szurdi, Gregor Maier, Mark Felegyhazi, Christian Kreibich, Nicholas Weaver, Kirill Levchenko, and Vern Paxson, Springer, March 2012. The dissertation author was the primary investigator and author of this paper.

# Chapter 4

# The Introduction of xxx

Now we shift our focus towards a second case study, this time of the xxx TLD. While biz is a truly generic TLD like com, xxx is intended for those in the adult entertainment industry and therefore shows a different style of user behavior. Many independent groups, including the adult entertainment industry itself, were concerned that it would primarily generate value through defensive and speculative registrations, and did not demonstrate the intent to register primary domains. Fewer registrants share content in xxx with com, an activity that was common in biz. Additionally, many trademark holders feared association with adult content, and went on to register xxx domains that do not even resolve.

In this chapter, we measure the validity of these concerns using an improved methodology, combining the lessons we learned while investigating biz and the kinds of behavior we saw in xxx. We characterize each xxx domain and infer the registrant's most likely intent. We find that at most 3.8% of xxx domains host or redirect to potentially legitimate Web content, with the rest generally serving either defensive or speculative purposes. Indeed, registrants spent roughly $13M up front to defend existing brands and trademarks within the xxx TLD, and an additional $11M over the course of the first year. Our analysis shows that defensive registrations are not limited to generic TLDs. Additionally, our results suggest ICANN should consider how trademark holders will perceive prospective TLD strings and disallow ones that will include defensive registrations as a significant part of their revenue model.

## 4.1 Introduction

The xxx TLD was created to serve the needs of the adult entertainment industry and its approval by ICANN was a matter of great controversy (both from within the adult industry and without). Critics argued that the TLD did not serve any real need and that its creation would largely be monetized via speculators and defensive registrations. The xxx TLD provides an excellent case study in the current benefits of a new TLD, as the rollout periods make determining registration intent significantly easier than for other TLDs. The remainder of this section describes the controversies surrounding the TLD and its rollout processes.

### 4.1.1 The xxx TLD

ICM Registry created the first proposal for a xxx TLD in 2000, which ICANN rejected later that year. In their report, they cited the complex political climate and the lack of need for such a TLD, especially in their initial rollout [48]:

> ICM Registry's application for an xxx TLD does not appear to meet unmet needs. Adult content is readily available on the Internet. To the extent that some believe that an xxx TLD would segregate adult content, no mechanism (technical or non-technical) exists to require adult content to migrate from existing TLDs to an xxx TLD.

Despite these points, they also indicated their willingness to reconsider the proposal in the future.

After ICANN rejected their third and newest proposal in 2007, ICM Registry appealed for an independent review of their application [52]. In their appeal, ICM Registry complained that ICANN's rejection was "arbitrary, lacking in transparency, and discriminatory," and suggested that "ICANN materially violated its Articles of Incorporation and Bylaws." The review found in favor of ICM Registry but left the ultimate decision up to ICANN, urging it to "evaluate the continued uncertainty and risk associated with its decision, including risks to ICANN resulting from potential legal actions" [19]. In 2010, against the recommendation of its Governmental Advisory Committee (GAC), ICANN accepted ICM Registry's newest sponsored proposal, to be rolled out in 2011 for the express use of the "adult entertainment industry".

One of the largest concerns around an adult-oriented TLD was the trademark resolution process, an issue ICM Registry acknowledged even in their earliest proposals. In particular, due to the social stigma associated with the adult entertainment industry, brand and trademark holders wanted assurance that their marks would not be unfairly associated with the xxx TLD. For this reason, among others, ICM registry introduced a multi-phase rollout for xxx registrations, described in Section 4.1.3.

### 4.1.2 Criticisms and Challenges

Criticisms of the xxx TLD have come from both the adult entertainment community and general brand holders.

The former class has been very suspicious that xxx would both incur additional costs to their businesses and that it might be a harbinger of future regulations forcing them to register exclusively under xxx where they could be easily filtered.

In November 2011, shortly before the opening of General Availability, Manwin Licensing International, the owner of many popular adult entertainment sites in other TLDs, filed a complaint against ICM Registry [29, 30]. In their complaint, they claim ICM Registry uses "monopolistic conduct, price gouging, and anti-competitive and unfair practices." Manwin also describes the necessity of defensively registering one's names, and further claimed issue with the entirety of the defensive registration process, described in the next section. This includes ICM Registry's naming restrictions, which require registered trademark status on *exact strings* to participate in the Sunrise periods, and its "monopolistic pricing."

In a settlement announced in May 2013, the companies settled under terms that temporarily reduce the cost of xxx registrations ($8.99, comparable to .com, for up to 10 years) for the remainder of the month, with similar periodic discounts in the future.[1]

### 4.1.3 XXX Deployment Phases

In Chapter 2, we provided a timeline that included the creation of ICANN, their decision to introduce new gTLDs, and some early studies on the types of registrations in biz. Some of this information is relevant for xxx as well, but even more informative is its launch period. Partly due to the controversy surrounding xxx and partly to reduce the number of speculative registrations, the xxx TLD initially had many options for users to choose from when registering or defending their names. This section describes the rollout period in detail, while Table 4.1 summarizes the registration options. We use GoDaddy as our source for all pricing information [15] because it is the largest registrar for xxx domains with a quarter of all registrations [31], as well as the one most prominently advertised by ICM Registry [13].[2]

ICM Registry made the very first xxx domains available through its *Founder's Program*. Users could register these names by bidding on them, beginning in late July 2011. ICM Registry's goal in auctioning off the initial batch of xxx domain names was to ensure that the most highly desirable adult-relevant domains would be purchased

[1] While not indicative of all registrations, the xxx zone shows a 14x increase in the number of new domains per day during this discounted registration period. Within a few weeks of its end, the churn rate returned to its pre-discount levels, giving some indication of price sensitivity on the registration process.

[2] We use retail prices in our calculations since this reflects the gross revenue paid into the xxx TLD. However, it is important to remember that this money is shared between the registrar, the registry, the registry's sponsor and ICANN. Our understanding is that the *wholesale* price for xxx is effectively $62, with $10 per domain going to the International Foundation For Online Responsibility (IFFOR), who is the sponsor for xxx, and $2 going to ICANN.

**Table 4.1.** Domain registrations on or before December 6, 2012 (the opening of general availability) and associated first-year fees.

| *Registration Period* | *Domains* | *Fee per Domain (USD)* | *Total Revenue (USD)* |
|---|---|---|---|
| Registry Reserved Names | 15,000 | 0 | 0 |
| Approved Performer Program | 3500 | 0 | 0 |
| Founder's Program | 1524 | Variable | 4.2 M |
| Sunrise A | 12,496 | 210 | 2.6 M |
| Sunrise B | 66,442 | 200 | 13.3 M |
| Landrush | 5022 | 200 | 1 M |
| General Availability | 55,367 | 100 | 5.5 M |
| Totals | 159,351 | | 26.6 M |

by someone who wanted to grow a business and actually use the name, rather than for speculation purposes. These proposals required an accompanying description of the intended monetization model. ICM Registry auctioned 1,524 domain names in this way, including names like `casting.xxx`, for a total of roughly $4.2 million USD [20].

In addition, ICM Registry reserved some `xxx` domains from registration. These domains fell into two distinct categories. The first category, domains in the *Approved Performer Program (APP)*, contains the names of specific celebrities, as well as actors and actresses in the adult entertainment industry [66, 70]. The registry held on to these domains until February 2012, a few months after the beginning of general availability. These then became available only to their respective persons, with the first year of registration included. ICM Registry claims this program included 3,500 domains at its inception [70].

The other category, *Registry Reserved Names (RRNs)*, includes domains made unavailable by ICM Registry on behalf of others, including "words of cultural and/or religious significance" [71] and words requested by the government [57]. These domains will never be registrable and they resolve to a Web page with the text "This domain has been reserved from registration." While some of these names can be guessed (e.g., the first and last name combination of every U.S. Senator or Representative as of December 2011), ICM Registry does not supply an exhaustive list. Consequently, we are unable to empirically measure the correct number of these (see Sunrise B below), but ICM Registry's chief executive officer is reported to have claimed 15,000 domains were so reserved [45].

After this auction, ICM Registry opened the *Sunrise A* registration period on August 28, 2011. During this phase, users could get first pick of domains relating to their existing trademarks, or domains they already owned under a different TLD. Simultaneously, ICM Registry also opened the *Sunrise B* phases, which provided the first (and only) chance for brand holders to *permanently* reserve their documented trademarks by paying a one-time fee. No potential registrant can later register a Sunrise B domain, not even the original applicant.

Domains reserved from registration through Sunrise B follow the same process as Registry Reserved Names; both resolve to the same Web page and have the same whois record, using a specific contact email at ICM Registry. We are able to empirically measure the aggregate of these two categories, 81,442 domains, but cannot distinguish between them. Thus, we derive the number of Sunrise B registrations, 66,442, by subtracting the claimed 15,000 RRNs from this total.

At the close of both Sunrise periods, ICM Registry announced the total number of sunrise domains as 78,938 [21]. The 66,442 Sunrise B registrations thus implies 12,496 Sunrise A registrations.

A one-week *landrush* phase followed the sunrise period, where companies with competing applications could bid for a name. Finally, on December 6, 2011, xxx opened up to *general availability*. Like other TLDs, names are now available on a first-come, first-served basis. ICM announced the number of first-day general availability registrations as

55,367 and a total TLD size of 159,351 domains [60]. Based on that total and the sizes of the categories, we calculate the size of the only remaining category, landrush, as 5,022 domains.

### 4.1.4 The xxx Case Study

By combining public ICANN records, analysis of the xxx zone files, active whois requests and crawling the content of live sites, we have constructed what we believe is the most comprehensive picture to date of how xxx is used in practice. We find that almost half of all xxx domains do not appear in the zone file at all, and of the approximately 107k domains that do resolve, three-fourths are objectively defensive registrations. A significant fraction of the remainder point to parked sites, suggesting speculative use. Finally, our calculations suggest that the clear majority of registration revenue, both during the pre-registration period and for ongoing operation, is driven by defensive concerns rather than entrepreneurship. Because the xxx TLD situation is clearly unique, we believe better understanding the differences between registrations in this TLD and biz could provide additional insights about ICANN's far broader expansions in the name space.

## 4.2 Data Collection

We use data gathered from the xxx zone file and ICANN reports, as well as actively gathered Web, whois, and name server responses. Additionally, we use Alexa and SIE passive DNS to examine the xxx TLD visit patterns of Internet users. This section describes our data collection methodology, as well as the data collection methodology of each external service we rely on.

### 4.2.1 Zone File

Just as with com and as per the ICANN .XXX Registry Agreement [67, 72], ICM Registry allows users to download complete copies of the xxx zone file, updated once every 12 hours, except for certain purposes (such as bulk advertising). The xxx zone file we download contains only NS and A records for second-level domains. We largely use the zone file as a list of resolving xxx domains for any particular day. The zone file contained 106,789 unique domains on April 12, 2013, the date of our latest Web crawl.

### 4.2.2 ICANN Monthly Reports

An addendum to the ICANN .XXX Registry Agreement also specifies that the registry must compile domain registration reports, once per month [68]. ICANN publishes these reports for all gTLDs on their Web site [31]. These reports provide unique domain registrations, adds, transfers, and deletes per registrar.

These reports only include *registered* domains, i.e. those sponsored by a registrar. For each registered domain, some registrant must pay an annual fee or the domain becomes available again. Registered domains do not necessarily resolve. Additionally, ICM Registry includes some domains in the xxx TLD themselves; since no registrar sponsors these domains, ICM Registry does not include them in the ICANN reports.

ICANN publishes these reports once per month with a three-month delay. Partly based on our measurements (Section 4.3.1), we believe the information in the reports reflects data as of the last day of the month. While the size of these reports varies, the April 2013 report contains summary information for 88 unique registrars and 108,337 domains.

Together, the zone file and ICANN reports provide at least summary information about every registered or resolving domain in the xxx TLD. We expect most non-registered and non-resolving second-level domain strings to be those that are still available for registration, although other categories do exist. The only other domains we believe to fall in this set are those names reserved from registration by ICANN in all new gTLDs (Appendix 6 of the Registry Agreement [69]), such as all names two letters or fewer and ICANN- or IANA-related names.

### 4.2.3 Web Crawl

For each registered domain appearing in the xxx zone file, we collect the Web content served with the top-level URL (e.g., http://icm.xxx). We save the DOM, a screenshot, the HTTP status code, and any associated headers per domain.

Though we also discuss our use of Web data in Section 3.2.2, we improved our Web crawler significantly for this study in terms of features, accuracy, and robustness. Our Web crawler is based on a custom extension for Mozilla Firefox [28] with supporting code written in Python. By using a browser with significant market share, a real JavaScript engine, and widespread plugin support, we are able to follow even convoluted redirects and parse a wide variety of content. We can also conveniently record iframe contents, which are popular on both parked and redirected Web pages.

For this study we use two crawls of the xxx zone file, one on January 10, 2013 and the other on April 12, 2013. While most of our analysis focuses on the most recent, having two Web crawls allow us to investigate category shifts over time.

During the January 2013 Web crawl, the connection between the network file server and two of our crawler instances failed. For 7.8% of our Web visits with HTTP 200 (OK) status codes, we recorded headers and associated data, but no DOMs or screenshots. We crawled the zone in order with each crawler instance processing 100 domains at a time, causing these failures to cluster into bursts alphabetically. This outage had no effect on our April data set.

Table 4.2 gives a breakdown of the HTTP and DNS resolutions for both Web crawls. This table represents every domain in the TLD; data for domains that do not appear in the zone are calculated with the ICANN reports using the methodology described in Section 4.3.1. The zone contains significantly fewer domains in April than it did in January, which reflects domains that were not renewed at the end of the year. The registration period corresponds to a little over a year and a month due to ICM Registry's grace period for renewing registrations, a minimum of 35 days [17].

### 4.2.4 Active Whois

We actively crawled whois records for our xxx case study similar to our work in biz. However, we were less interested in identifying shared ownership and instead used them to identify Registered Non-Resolving (RNX) domains (Section 4.3.1). RNX domains return valid whois records, while non-registered domains return the string "NOT FOUND", thereby providing an easy method to differentiate between them. We also used whois to identify reserved domains (Section 4.3.1), which all return the same contact email address.

We used a modified version of pywhois for whois crawling and parsing. We crawled the xxx versions of edu domains and domains in the Alexa top 750,000 at a rate of 10,000 per day between January 24, 2013 and April 11, 2013. In addition, we crawled all domains in the xxx zone file on April 17, 2013, as well as each xxx domain in our SIE data set (Section 4.2.6) that was never in a xxx zone file.

### 4.2.5 Alexa

Alexa [1] makes domain name rankings and traffic information publicly available. Alexa bases their rankings on data collected by the Alexa toolbar, an opt-in browser addon. Besides performing Alexa's data collection, the toolbar provides the user with traffic rankings and reviews of the sites they visit, along with an analysis of how traffic is

**Table 4.2.** The Web and DNS resolution results of xxx domains.

| *Category* | *2013-01-10* | *2013-04-12* |
|---|---:|---:|
| In Zone | 116,861 | 106,790 |
| DNS Errors | 3,961 | 3,176 |
| Other Errors | 730 | 741 |
| HTTP 200 | 112,170 | 102,873 |
| DOM Stored | 103,436 | 102,873 |
| No DOM (bug) | 8,734 | 0 |
| No Zone | 90,156 | 86,710 |

driven to the site. This toolbar is available for Chrome, Firefox, and Internet Explorer. We use the Alexa rankings as one gauge of domain popularity (Section 4.5) and when searching for Registered Non-Resolving domains (Section 4.3.1).

### 4.2.6 SIE Passive DNS

We use passive DNS data collected by SIE [54] to identify those domains that users resolve in practice. This data contains DNS queries performed by recursive DNS resolvers. Organizations provide data to SIE on an opt-in basis, and it includes data from a variety of different providers, including multiple residential ISPs and universities. SIE groups queries with responses whenever possible.

Due to caching effects, it is difficult to use passive DNS data to accurately estimate the visit rate of any particular domain. Every registered domain in the xxx zone has a TTL of one day. Barring early cache evictions, a resolver will not look up a registered domain more than once per day even if many users configured with that DNS resolver visit the site during the day. As a result, we limit our use of SIE data to very coarse-grained measurements. To alleviate concerns over two resolvers in our data set that account for a significant portion of all xxx queries, and to remove highly infrequent lookups, we include only those domains queried by at least three resolvers. Apart from this requirement, we use SIE as a binary data source; either the domain is included or it is not.

For each A or AAAA query of a xxx domain, we keep the timestamp at date granularity and the registered domain. We distinguish between *live* and *dead* domains. Live domains have a valid response, no error code, and a non-empty answer section. These domains appear in the zone file, and the existence of such a record suggests that some user visited the domain in search of content. Any domain for which SIE contains a query but the domain is not live is a dead domain. These domains may be registered and not resolving, or they may not be registered at all. Dead domain lookups suggest that some user would have visited the corresponding domain, if it were registered and populated with some sort of content. We use dead domains when searching for domains that are Registered Non-Resolving (Section 4.3.1).

## 4.3 Content Categories

As a first step towards categorizing the registration intent of each domain, we categorize their contents. In this section we describe our domain classification methodology, including the indicators we use to identify each type, which data sources we draw from, and the number of domains in each category. We also examine how domain classifications have changed over time.

### 4.3.1 Classification

After crawling all domains, we cluster them based on resolution status and Web content using text shingles [6] to find the most prevalent page types. For each large cluster, we then create regular expressions to match and classify the content. For example, all Web pages corresponding to a particular domain parking service may use a single CSS file served off a different domain. Any pages with this particular CSS file likely belong to the same parking service, regardless of whether or not they are all in one cluster. We apply a *content tag* to any Web page matching a number of regular expressions for the tag, with the threshold varying by tag.

We created tags to match the largest 40 clusters, including every cluster with 50 or more domains. We also tagged some smaller clusters, typically because they were popular earlier in our study.

One consequence of our content classification methodology is that we are unlikely to find primary registrations with the clustered Web content. Domains will only cluster together if they have the same or similar Web content, but few content creators will likely need many domains for the same content. Instead, this technique is much more useful for identifying defensive, reserved, or parked domain registrations.

We classify each domain into one of eight categories:

**Reserved** domains are permanently unregistrable.

**Approved Performer Program** domains correspond to celebrity names and have a separate registration process.

**Premium Domain Names** are potentially desirable names with larger up-front registration costs.

**Parked** domains are advertised as for sale, often with additional automatically-generated advertisements as well.

**Non-Resolving** domains have been registered but do not resolve, and are usually not even in the xxx zone.

**Content** domains give meaningful Web responses.

**Unused Web** domains host little or no Web content.

**Unknown** domains could not be automatically classified.

The rest of this section discusses each classification in detail.

### Reserved Domains

As described in Section 4.1.3, there are a few categories of reserved domains. This section deals with two of those, Registry Reserved Names and Sunrise B reservations. We group these together because we are unable to distinguish between them.

Reserved domains are straightforward to identify. Each returns an identical ICM-generated Web page over HTTP stating "This domain has been reserved from registration". These domains each have NS records that are subdomains of `ICMREGISTRY.NET`, which host all of ICM Registry's domains. The whois information for these domains also specifies ICM Registry as the contact point for the domain, consistent with their intent as stated in [71].

We match whois instead of Web content to classify a domain as reserved. The contact email address `registryescrow@icmregistry.com` appears in the whois record for each reserved domain, a different address than for other kinds of ICM Registry domains. Additionally, we found 172 instances of registrants copying ICM Registry's domain reservation Web page and hosting it on their own Web server with no changes, making the Web content ambiguous.

Using this methodology, we find 81,442 reserved domains. Even after checking for distinguishing characteristics in their whois records, we are unable to empirically separate Sunrise B and RRN reservations. ICM Registry reported 15,000 RRNs [45], which we use as an estimate, and classify the 66,442 remaining domains as Sunrise B reservations. These domains resolve but are not registered, so they appear in the zone file but not the ICANN reports.

### Approved Performer Program

Domains in the Approved Performer Program, described in Section 4.1.3, resolve to a Web page indicating their participation in the program and specifying a process

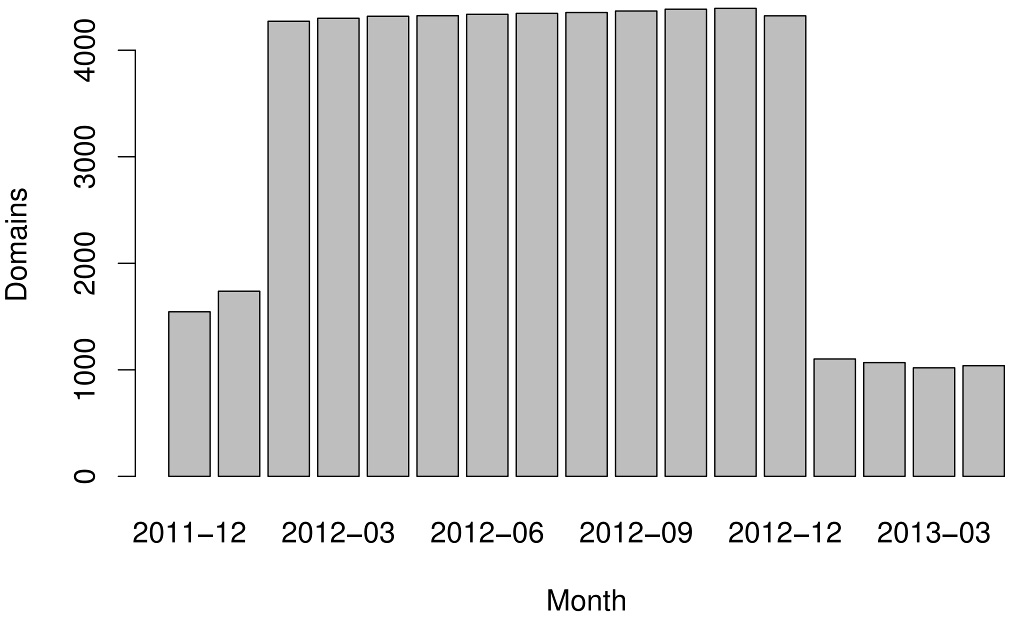


**Figure 4.1.** `xxx` registrations with `name.com` by month.

by which the person in question may take ownership of their domain name. Since they resolve, they also appear in the zone file. We identify them using our tagging methodology described in Section 4.3.1.

ICM Registry registered these domains through name.com to ease transferring ownership, and also to provide the rest of the first year free to the registrant. According to whois, this registration was valid from February 3, 2012 to January 24, 2013, at which time the sponsoring registrar switched back to ICM Registry. Thus they appear in the ICANN monthly reports during this time period, but not otherwise. We use this set of domains to discover whether the ICANN reports provide the unique number of domains per registrar across the entire month, or from a single snapshot.

Figure 4.1 shows the total domains registered through name.com each month, as reported to ICANN. As expected, we see an uptick of `name.com` registrations between January and February 2012, and a drop-off between December 2012 and January 2013. Since these domains switched registrars on January 24, we conclude that the ICANN reports must only use data from the last seven days of the month. We make the simplest assumption, and assume ICM Registry reports totals for the last day of the month.

We identify 2,415 Approved Performer Program pages as of April 12, 2013. The increase of 2,534 `name.com` registrations between January and February 2012, which include both normal domain registrations and those via the Approved Performer Program, suggests that ICM Registry only sold a very small number of these domains in the intervening 15 months. When ICM Registry started the program in October 2011 [70] they claimed withholding 3,500 of these names [66], making the first few months of the program significantly more successful.

### Premium Domain Names

ICM Registry also withholds potentially desirable domain names from registration as part of their Premium Domain Names (PDN) program. These domains require potential registrants to purchase them for larger sums prior to registering the name. Premium Domain Names successfully resolve, returning ICM Registry's search portal for `xxx` domains when visited via HTTP. They appear in the zone file, but do not appear in the ICANN monthly reports since they are reserved directly by ICM Registry. We find 991 of these domain names.

### Parked Domains

Many domains are parked with domain parking services, often but not always established through a registrar. Domain registrants speculatively monetize parked domains through ad revenue and domain resale. While there are many templates for parked pages, most contain these features:

- Text indicating the domain's availability ("This domain for sale").
- Advertisements relating to words in the domain name, often laid out to look

like search engine results.

❖ The name and/or logo for the domain parking service.

Our results show domains parked through many different parking services. While each parking service can set up their infrastructure in any way they choose, we expect all parked domains to have Web content. This expectation clearly holds for parked domains monetized through ad revenue, as these registrants can only make money if they actually serve advertisements. Speculatively parked domains do not necessarily require Web content, but likely get more offers if they advertise the domain as being for sale. For these reasons, we identify parked domains through their Web content, not through their whois records or hosting infrastructure.

Though we note that parked domains can be motivated by either domain resale or ad revenue, we consider domain resale to be the most likely motivator behind parked domains. Remember that `xxx` domains are significantly more expensive than domains in other TLDs, costing roughly $100 per year. Ad revenue is unlikely to scale with domain registration costs, which intuitively suggests the other strong motivator, domain resale, as the most likely reason for these domains' registrations.

These domains are both registered and resolving, so they appear in both the zone file and the ICANN monthly reports. We find 8,262 domains parked through 21 different programs using our content tags, as described in Section 4.3.1, including 3,561 domains parked through GoDaddy and 1,923 through Sedo.

### Non-Resolving Domains

Some `xxx` domains are configured with name servers but do not successfully resolve when queried through the DNS. Upon further investigation, we discovered that many domains in this category are hosted on otherwise working name servers, but return `REFUSED` or `SERVFAIL` responses when queried with the domain in question. For example, Google has registered picasa.xxx with the name server ns1.google.com (and sequentially numbered variants), but returns REFUSED for lookups of this domain (which recursive servers often report as SERVFAIL). Other name servers return no error, but no answer, authoritative, or additional sections either. We see some less common failure modes, such as temporary failures.

Further, some xxx domains are *Registered Non-Resolving (RNX)*. These domains do not appear in the zone file because they have no associated DNS records, but some registrant is still paying the normal domain registration fee. The consequence of this kind of registration is that no other party will be able to successfully register that domain name, making this a clear case of a defensive domain registration. ICM Registry and some registrars cite this method as the correct way to defend your trademark after the beginning of general availability [16, 49].

Given that this category is the logical equivalent of domains in the ICANN reports but not the zone file, we can calculate how many of these exist at any given time by using the categorizations from the zone file. Some domains from the ICANN reports also exist in the zone file; let *Overlap* represent the total number of domains in this set. Then we can find the number of Registered Non-Resolving domains, *RNX*, using:

$$RNX = Report - Overlap$$

While this method will compute the number of RNX domains, the size of the overlap set changes over time and is difficult to measure directly. We can find more accurate historical values for *RNX* by rewriting this formula in terms of the size of the zone file, *Zone*, and the size of the more stable and measurable categories. This includes Sunrise B applicants, Registry Reserved Names, Premium Domain Names (PDNs), and sometimes the Approved Performer Program (Section 4.3.1).

Using this set of categories, we can calculate *RNX* as follows:

$$\begin{aligned}
Report &= Overlap + RNX \\
Zone &= Overlap + Reserved + APP + PDN \\
Report - Zone &= (Overlap + RNX) - \\
&\quad (Overlap + Reserved + APP + PDN) \\
Report - Zone &= RNX - Reserved - APP - PDN \\
RNX &= Report - Zone + \\
&\quad Reserved + APP + PDN
\end{aligned}$$

Figure 4.2 shows the number of RNX domains for each month from general availability of the `xxx` TLD through April 2013. Our methodology shows the existence of 86,710 RNX domains in April 2013. Added to the 3,176 non-resolving domains in the zone file, we find a total of 89,886 non-resolving `xxx` domains.

We were able to identify some of these RNX domains by actively crawling their whois records. Domains that have not been registered or reserved return “NOT FOUND”, while those that have been registered return valid whois information, regardless of whether or not they have any associated NS entries.

To preliminarily identify registered non-resolving domains, we performed whois queries for the `xxx` version of each domain in the `edu` zone file (such as `ucsd.xxx`), as well as for each domain in the Alexa top million Web sites. We also crawled whois information for each SIE lookup attempt for a dead domain. These sources of likely defenders identified 19,873 (23%) of the RNX domain names, but many other defenders remain unidentified.

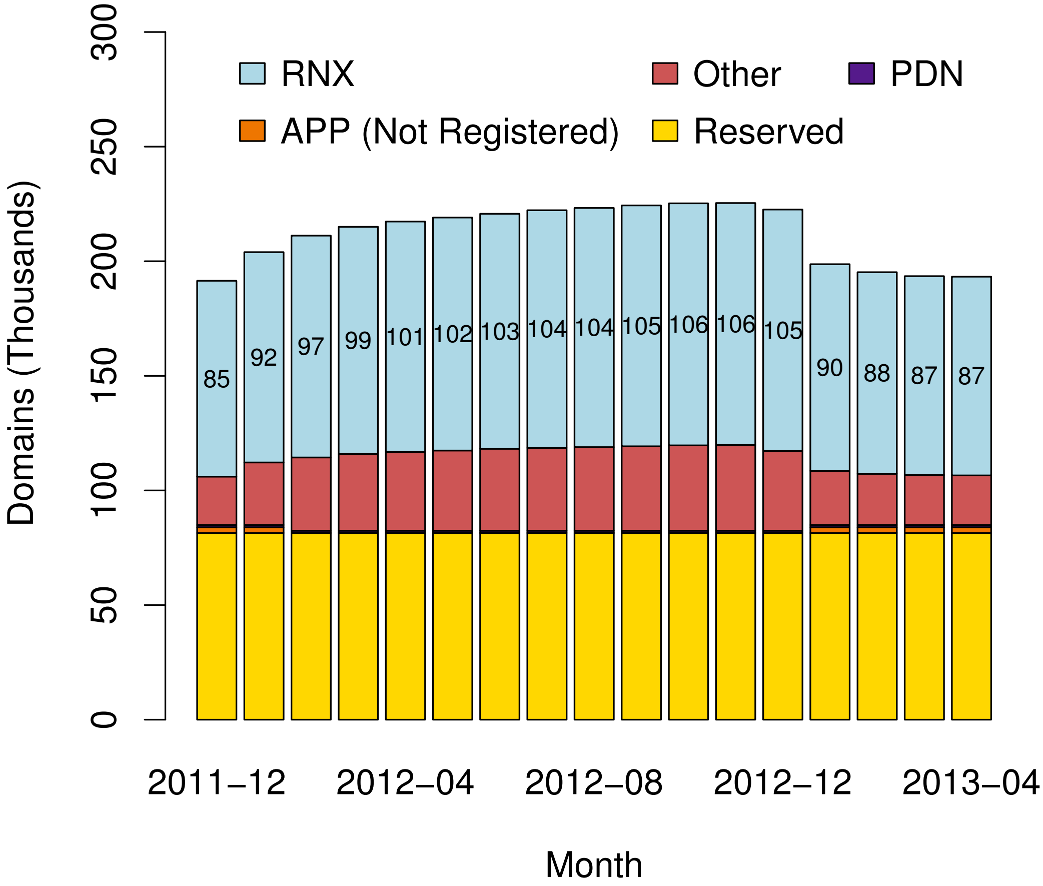


**Figure 4.2.** Breakdown of all domains in the zone and ICANN reports, which we use to estimate the number of RNX domains. The top two bars show domains in the ICANN reports, while the bottom four bars show the zone file. The red bar in the middle overlaps them.

### Unused Web

We categorize domains as “Unused Web” if the domains resolve but do not return any useful Web content. Some xxx domains successfully resolve through the DNS but do not respond to HTTP requests on port 80. The remainder run a public Web server on the default port and return either a blank Web page, the default Apache Web page, a PHP error page, or a similar content-free response.

There are several possible explanations for these responses. First, this state could be transient; that is, other kinds of domains could fall into this state until their operators notice and rectify the problem. Second, these domains could signal defensive registrations; perhaps some brand holders are not familiar with the suggested defensive methods and misconfigure their servers or NS entries. Alternatively, these domains could be primary registrations, but intended to support services other than the Web.

Regardless of intent, we identify 570 domains as returning HTTP errors and 609 as returning successful but content-free responses, for a total of 1,179 unused Web domains.

### Content

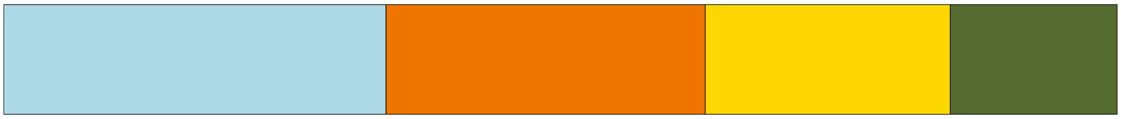
A handful of our large Web content clusters contain legitimate content that end users may find interesting, and thus we categorize them as "content" domains. Additionally, most adult pages have an 18 USC 2257 notice, which states their compliance with record-keeping requirements for each adult performer on the site. We classify 231 xxx domains as content due to our content tags on large clusters and 1,072 due to their 18 USC 2257 warning, for a total of 1,303 content domains.

### Unknown

Finally, our automatic classification techniques do not correctly classify some xxx domains. Our classification criteria err on the strict side to avoid false positives, and often rely on manually-generated Web content tags as described in Section 4.3.1. The 7,885 domains we are unable to classify with the methodology above fall into this category.

We suspect much of this category contains legitimate content. Our content classification depends heavily on manually-generated tags, which we only create for our largest clusters. Many types of legitimate content will not cluster with any other xxx domains, while our other categories will tend to contain pages that cluster together better.

To verify our suspicions about these unclassified domains, we performed a manual



| | *Category* | *Domains* | *%* | *Total (Est)* |
|---|---|---|---|---|
| 🟦 | Content (Redirect) | 103 | 34% | 2707 |
| 🟧 | Content (Local) | 86 | 29% | 2260 |
| 🟨 | Unused | 66 | 22% | 1735 |
| 🟩 | Parked | 45 | 15% | 1183 |

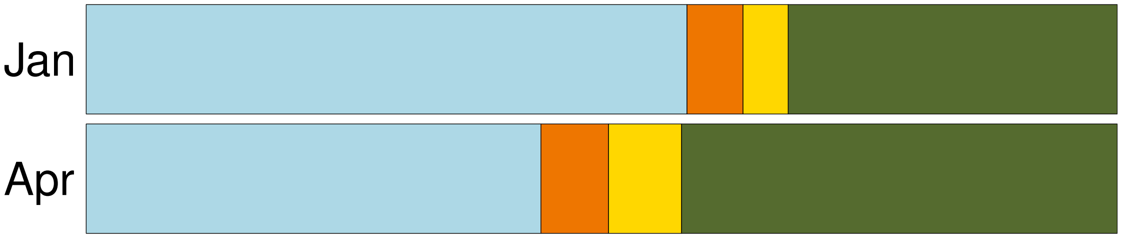
**Figure 4.3.** Categories of 300 random Unknown domains.

sample of the Web content for these pages. We examined the Web content for 300 domains in this category, chosen uniformly at random. We use the Web contents and screenshots from our April 12 Web crawl while classifying these domains, so the content matches exactly with the content we were unable to automatically classify. Figure 4.3 summarizes our results. In each case we were able to classify the domain into one of our existing categories. As expected, most domains (63%) that we could not automatically classify fell into the content category, with the rest dividing into unused and parked.

For any domain that showed primary content, we also judged whether or not it was adult-oriented. We used a very liberal policy for assigning content to this category. For instance, `pureheaven.xxx` had broken image and product links for adult-oriented items during our crawl, which we considered adult. We found 76% of originally unknown content or redirect pages to be adult in nature.

### 4.3.2 Overall Classifications

Figure 4.4 shows the content class breakdown for each domain in the `xxx` zone file that is both registered and resolving (except Approved Performer Program domains in January). Both classifications ignore all `xxx` domains which had HTTP 200 response codes and we were unable to store the DOM in January (Section 4.2.3). While we believe these domains to be distributed across the content categories like the other `xxx` domains,



| | *Category* | *January* | % | *April* | % |
|---|---|---|---|---|---|
| 🟦 | Parked | 15,090 | 58.3% | 7663 | 44.4% |
| 🟧 | Unused | 1407 | 5.4% | 1139 | 6.6% |
| 🟨 | Content | 1135 | 4.4% | 1231 | 7.1% |
| 🟩 | Unknown | 8265 | 31.9% | 7342 | 42.3% |
| | Totals | 25,897 | | 17,375 | |

**Figure 4.4.** Content classifications for all *registered* and *resolving* xxx domains during both crawls.

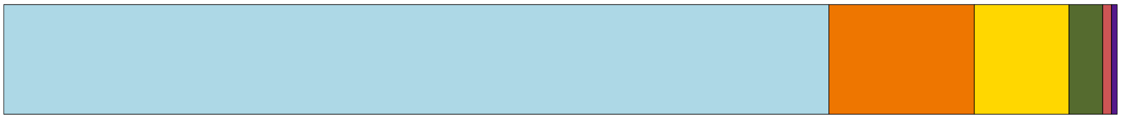
removing them from both sets will make category size comparisons between the data sets more straightforward.

Even among only domains that both resolve and are registered, parked domains are the most prevalent. After breaking up the unknown category using the data from Figure 4.3, parked domains still make up the largest category of both classifications. Note, though, that the sum of all of these categories is still much smaller than either of the two largest categories of xxx domains in general, RNX and reserved domains.

When comparing the classifications across time, two features stand out. First, there are many fewer registered resolving domains in the April data set than in January. Second, there are far fewer parked domains in April than there are in January, even after accounting for the general dropoff. In the following sections, we investigate both of these features.

### 4.3.3 Elapsed Domain Names

Figure 4.4 shows that the xxx zone file contained fewer registered resolving domains in April 2013 than it had three months before, similar to the reduction in zone



| | *Category* | *Domains (Jan)* | *Elapsed* | *%* |
|---|---|---|---|---|
| 🟦 | Parked | 15090 | 7979 | 52.9% |
| 🟧 | Unknown | 8265 | 1406 | 17.0% |
| 🟨 | Non-Resolving | 3866 | 916 | 23.7% |
| 🟩 | Unused | 1407 | 326 | 23.2% |
| 🟥 | Content | 1135 | 85 | 7.5% |
| 🟪 | XXX Support | 3302 | 56 | 1.7% |

**Figure 4.5.** Elapsed domain names across categories. The last column shows the probability that a registrant chose not to renew a random domain from January.

file size in Section 4.2.3. On January 10, 2013 the xxx zone file contained a total of 116,833 entries, only 3,048 fewer than its peak of 119,881 a month earlier; this drops off to 111,554 only ten days later, with 5,279 fewer domains. We see 10,768 registered and resolving domains on January 10 that do not appear in the April 12 zone. Since this time period falls roughly one year after the beginning of general availability, we can use this gap to approximate those domains the registrant chose not to renew after one year.

Figure 4.5 shows the content classification for each of the 10,768 xxx domains in the January 10, 2013 zone, but not the April 12 one. We actively crawled whois for these domains on May 6 and found that 9,584 of them were no longer registered, leaving 1,184 as registered but not resolving.

From the breakdown, most elapsed domains were parked, especially when compared to registered and resolving content in general. Using our content tags, we classified 15,090 domains as parked at the 21 roughly largest parking programs during our January classifications; over half (7,979, or 53%) of them disappeared over the next 92 days. This data suggests that domain parking in the xxx TLD has not been particularly profitable. Regardless of whether these domains were registered for ad revenue or resale potential, it

**Table 4.3.** Domains that changed content classifications between January and April.

| *Category* | *Total Domains* | *Total Changed* | *% Changed* |
|---|---:|---:|---:|
| Parked | 15090 | 573 | 3.8% |
| Non-Resolving | 3866 | 547 | 14.1% |
| Unused Web | 1407 | 490 | 34.8% |
| Content | 1135 | 99 | 8.7% |
| Unknown | 8265 | 738 | 8.9% |
| Support | 3302 | 18 | 0.5% |

appears as if their registrants no longer believed these domains were likely to pay off.

### 4.3.4 Classification Changes

When classifying domains with data collected over a small time interval, our classifications gain the appearance of being very stable. An alternate hypothesis is that some classifications, such as unused or parked domains, are transient domain states; that is, domains are only unused or parked for short periods of time while their owners set up real content.

We use our January 10 Web crawl and a methodology similar to Section 4.3.3 to look for transient domain states. We classify domains from the January 10 set using the same methodology and content tags as we use for the April 12 data set. For each potential categorization, Table 4.3 shows the number of domains originally in that category, the subset of those that switched categories in the next 92 days, and the fraction of domains that switched categorizations.

For most categories, we see very little change in classification. Less than 4% of parked domains switched classifications. This result and that of Section 4.3.3 strongly suggest that parked domains in xxx tend to be parked for long periods of time.

The largest outlier is the category of domains with either no Web server, or that serve only empty Web content. Nearly 35% of domains in this category switch to a

**Table 4.4.** Registration motivations of all *registered* xxx domains (55% of all domains) for which we could infer intent. The remaining domains are reserved domains, which are defensive by definition.

| *Intent* | *Domains* | % |
|---|---:|---:|
| Primary | 6,270 | 5.9% |
| Speculative | 9,445 | 8.9% |
| Defensive | 89,886 | 85.1% |
| Total | 105,601 | |

different one within 92 days. When we introduced these domains in Section 4.3.1, we suggested three different potential purposes for them: primary, defensive, or transient. While the motivations behind the other 65% are still unclear, this evidence supports transience as the explanation for a significant portion of these domains.

## 4.4 Registration Intent

Upon categorizing each xxx domain, we would like to determine the registration intent, whenever possible. As described in Section 2.4, in this section we group xxx domains into one of three functional groupings: primary, defensive, or speculative. Table 4.4 summarizes our results. In particular, we find that an overwhelming number of registered domains (85%) are defensive in nature, with only 5.9% of domains registered for the purpose of serving content.

Before categorizing domain registrations by motivation, we must identify and separate all varieties of reserved domains. The primary purpose of further categorizing domains by motivation is to compare both the number and the amount spent on domains registered for different purposes. Because of this goal, we first need to identify and separate those domains with no recurring registration costs. This includes the reserved and unregistrable domains, Premium Domain Names, and those in the Adult Performer

Program.[3] We find a total of 84,848 domains with no recurring registration costs using these criteria, all of which are defensive in nature.

In Section 4.3.1, we proposed three different motivations behind xxx domains with unused Web content. While the classification change results in Section 4.3.4 were not conclusive in this regard, it did suggest the transient domain hypothesis is the most common. As a result, we do not include these domains in the rest of this section. While it is possible that some of these domains should have definitive classifications, the size of the entire category is relatively small and so the misclassification of a subset of these domains will have little impact on the results.

The rest of this section discusses the registrant motivations assigned to the other 105,601 domains in the remaining content categories: "Parked", "Non-Resolving", and "Content". We also describe any known issues with our decisions.

### 4.4.1 Defensive Registrations

RNX domains are the most obvious category of defensive domain registrations. ICM Registry and GoDaddy, the largest registrar in xxx with a quarter of all domains [31], both specify RNX as the intended mechanism for defensive registrations after the conclusion of the Sunrise B phase [16, 49]. RNX domains also do not resolve, making it difficult for a devious user to use them even for obscure purposes.[4]

[3] While domains in the Adult Performer Program were registered through name.com for a year, we are unsure if ICM Registry obtained these through the normal domain registration process. Regardless, the majority of the domain registration cost would have returned to ICM Registry, so they should still not be included in this total.

[4] These registrations could also be speculative, but we assume speculative registrants want to broadly advertise their domains as being for sale, at a minimum by resolving to a server hosting recognizable Web content.

### 4.4.2 Primary Registrations

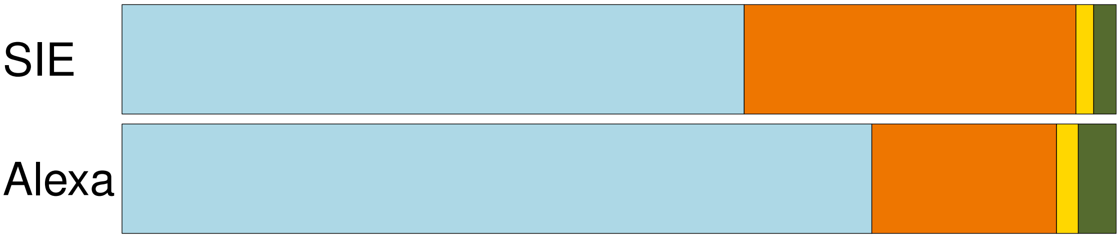
We assume all domains that point to content are primary registrations. We do not differentiate between direct and redirected content. We also ignore whether or not the content is adult in nature and therefore even belongs in the xxx TLD. At first this might seem strange, as an off-domain redirect might be a good indicator of a defensive registration.

To understand our policy, consider the definition of a primary registration. When one domain redirects to another, the primary domain is the one that the content owner actively advertises. If we could categorically say that xxx domains are undesirable to all registrants, redirects would strongly indicate defensive registrations. While valid for trademark defenders, it is not valid for content producers, who may want to use the connotations of the xxx TLD for marketing purposes. The registrant may serve content via a domain in a different TLD for historical reasons, but prefer and advertise the xxx version, thereby making it the primary registration.

We noticed a fair number of redirects in our manual sample (see Section 4.3.1); by extrapolating our manual sample results, we estimate 2,500 xxx domains whose Web contents redirect off-domain. This quantity makes up a large portion of domains with content but is dwarfed by the number of reserved (81,442) and RNX (86,710) domains. Given that there is no clear and simple policy for redirects, we assume they are all primary registrations and treat this classification as an overestimate.

### 4.4.3 Speculative Registrations

We consider parked domains to be speculative in nature. Parked domains incentivize the registrant to buy likely desirable domains without adding any useful content for their visitors, regardless of whether their monetization model favors ad revenue or resale.



| | *Category* | *SIE* | *Alexa* |
|---|---|---|---|
| 🟦 | Unknown | 632 | 138 |
| 🟧 | Content | 337 | 34 |
| 🟨 | Unused | 18 | 4 |
| 🟩 | Other | 23 | 7 |

**Figure 4.6.** Classifications for domain names appearing in passive DNS (top) and Alexa (bottom) within three days of our Web crawl. The SIE data only includes domains resolved by at least three resolvers.

## 4.5 Visit Patterns

So far, we have focused mostly on the costs of the xxx TLD by showing that there are many defensive and speculative registrations. In this section we examine the benefits. We would like to examine the TLD from the perspective of its early rejections by ICANN and determine the degree to which the TLD serves any unmet needs. One way to examine this question is to look at how many users visit xxx domains. We answer this question using our two sources of visit information, SIE and Alexa.

Figure 4.6 shows the breakdown of xxx domains in our SIE passive DNS data set within three days of April 12, 2013, the date of our latest Web crawl. This date range is contemporaneous with our Web crawl, so we expect our classifications to remain accurate.

We classified nearly every xxx domain (96%) appearing in the passive SIE data as either “unknown” or “content”. This result matches our intuition, as we expect users to be looking up domains with real content on them. A small fraction of these domains fall into the “unused” category. These domains may have been registered for purposes

**Table 4.5.** The five most popular xxx domains in Alexa.

| *Ranking* | *Domain* |
|---:|:---|
| 6,452 | perfectgirls.xxx |
| 12,450 | hornytube.xxx |
| 14,220 | hdsextube.xxx |
| 15,178 | rule34.xxx |
| 24,660 | entire.xxx |

other than Web content, such as IRC or email. Alternatively, these domains may offer Web content on a different port or content path than the default, and therefore not be represented in our data set. While xxx domains are significantly more expensive than domains in other TLDs, our manual sample showed us evidence of domains meant only for domain registrants and their close friends, such as blogs, a personal resume, and a handful of vanity domains. Some domains in our SIE data set may reflect this use case.

We performed a similar analysis on the top one million Alexa-ranked domains on April 12, 2013, of which only 184 xxx domains appeared. Figure 4.6 shows our categorizations for 183 of these domains; the last dropped out of the zone file on April 11. Table 4.5 shows the top-ranked xxx domains and their rankings. Only one domain appears in the top 10,000; for comparison, the name TLD also only has one domain in the top 10,000, but biz has 17 and info has 50.

The top Alexa domains predominantly fall into the unknown and content categories and closely resemble the SIE data, as we would expect. Together, unknown and content account for over 94% of xxx domains in the Alexa top million. We classified two Alexa domains as non-resolving; one of these appears to no longer be resolvable, while the other was likely a transient DNS failure during the time of our Web crawl. ICM Registry runs 5 domains in Alexa (classified above as "other"), such as search.xxx.

**Table 4.6.** Upfront and yearly renewal costs for xxx domains by registrant motivation.

| *Registration Intent* | *Peak Registrations* | *Initial Costs (USD)* | *Number of Domains* | *Recurring Cost (USD)* |
|---|---|---|---|---|
| Reserved | 84,848 | 13.3 Million | 84,848 | 0 |
| Defensive | 109,559 | 11.0 Million | 89,886 | 8.99 Million |
| Primary | 8,666 | 5.7 Million | 7,433 | 0.74 Million |
| Speculative | 22,313 | 3.3 Million | 11,196 | 1.12 Million |
| Totals | 225,386 | 33.3 Million | 193,363 | 10.9 Million |

## 4.6 Registration Costs

The domain classifications form a basis for measuring the impact of the xxx TLD in terms of its costs and benefits. The benefits are clear: primary domain registrants have a new domain name, likely one that is easier for users to remember or stumble upon, which might make their business more appealing. The cost lies in the money spent by domain registrants for primary, defensive, and speculative purposes.

Domain registration costs vary by registrar, but performing cost estimates requires us to choose a price. ICM Registry's domain registration page features GoDaddy most prominently [13], so as in Section 4.1.3 we base our price estimates using GoDaddy's registration costs of $100 USD [15]. We use the registration cost instead of the $62 wholesale price to capture money spent instead of any one particular party's revenue.

Table 4.6 shows the estimated registration costs for each domain in xxx. Recurring cost estimates use the registration intent numbers from Section 4.4. The only exception is transient domain names; these names are still registered (i.e., someone is paying the registration cost for them), but we have not previously categorized them. For purposes of our cost estimates, we divide these between primary and speculative registrations, based on the ratios those domains otherwise exhibit. Since all measurements are from the period shortly after one year of general availability, the recurring costs section of this table reflects domains for which users chose to renew initial registrations after one year.

We estimate that in terms of initial costs, only 17% were for primary domains and the rest were speculative or defensive in nature. In terms of recurring cost, the situation is even more stark: only 6.8% of recurring costs are for primary domains. These estimates underscore the conclusion that the vast majority of registration revenue, both during the pre-registration period and for ongoing operation, is driven by defensive concerns rather than entrepreneurship.

Initial cost estimates use data from a variety of sources. The peak number of domain reservations uses the categories described in Section 4.4. The only category of domain reservations for which registrants paid money is Sunrise B, and that cost comes directly from Table 4.1. We were able to calculate the peak number of RNX domains, our only category of defenders, in Section 4.3.1, as well as the peak number of domains in xxx. Primary and speculative registrations make up the remaining 30,979 domains.

We use the January content classifications to estimate the ratio of primary to speculative registrations. We use this data instead of the April data because it is temporally closer to the peak. Additionally, the exodus of parked registrations seen in Section 4.3.3 makes the April classifications less reliable for this purpose. We expect our estimate to be slightly biased towards content, since the zone size had already fallen by roughly 3,000 domains by January 10, and because more speculative registrants choose not to renew than primary registrants.

Finally, for our initial cost estimates, we assume all of the most expensive domain names (those in Sunrise A and the Founder's Program) all went to primary registrants. While some general availability registrations were certainly primary instead of speculative in nature, this assumption again lets us bias towards primary registrations. Additionally, while some domains in the Founder's Program may serve parked content, the uneven and unknown cost of individual domain names in this category creates the potential for fine-grained estimates to be incorrect in unpredictable ways. Instead, we assume the most expensive registrations were for primary content, and treat the resulting number as a high estimate.

## 4.7 Conclusion

The introduction of new TLDs is typically meant to increase consumer choice with respect to second-level domain strings by opening up second-level domains that have already been claimed in other TLDs. As we have previously shown with the `biz` TLD (Chapter 3), some people take advantage of such opportunities to register new content, while others feel compelled to defend their names, and still others seek to resell desirable domains without using them themselves. In general-purpose TLDs like `biz`, these registration types are all fairly common.

Aspects of the new `xxx` TLD amplify concerns about defensive and speculative registrations. Brand and trademark holders are particularly concerned about any potential association with the connotations of such a TLD, while previous members of the adult entertainment industry do not consider it to serve any real need. This chapter empirically studies the validity of these concerns by measuring the costs and benefits of the `xxx` TLD.

By gathering public data from ICANN, zone file records, active whois and Web content, we are able to build a complete view of `xxx` domains. We show that concerns over defensive registrations are particularly valid: nearly 92% of all domains in the `xxx` TLD exist for solely defensive purposes, including 83% of all domains with recurring registration costs. Speculative registrations make up 60% of the remainder, and only 7,433 domains, or 3.8% of all `xxx` domains, serve real Web content. Additionally, we find that registrants spent $24M in the first year of registration solely for defensive purposes.

Chapter 4 was supported by National Science Foundation grants NSF-1237264, by the Office of Naval Research MURI grant N00014-09-1-1081, and by generous research, operational and/or in-kind support from Google, Microsoft, Yahoo, and the UCSD Center for Networked Systems (CNS).

Chapter 4, in part, is a reprint of the material as it appears in *Proceedings of the International World Wide Web Conference 2014*. Tristan Halvorson, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. The dissertation author was the primary investigator and author of this paper.

# Chapter 5

# The New gTLDs

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. As we showed in Chapters 3 and 4, 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 chapter 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.

## 5.1 Introduction

Starting in 2013, delegation began of a whole new wave of TLDs. 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.

In this chapter, we identify the intent of registrants in the new gTLDs. While brandholders saw fit to defend their trademarks in the old TLDs and domain speculators thrived, the current much faster expansion could change both. With 579 new TLDs in the last 18 months, we expect many smaller companies to find it infeasible to defend their name in each. Additionally, with so many TLD options for any second-level name, speculators may find it difficult to resell even desirable names. Has such a rapid expansion of the TLD namespace merely extorted companies for registrations, like in `biz` (Chapter 3) and `xxx` (Chapter 4), or has ICANN's new application method made it easier for primary registrants to get the domains they want?

### 5.1.1 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) [39]. Each applicant prepared an extensive submission covering business, technical and operational issues and paid a $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 who would amortize these expenses across many such applications (e.g., Donuts, Google, Amazon).

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.

### 5.1.2 TLD Rollout

After delegation, the TLD life cycle becomes very registry-dependent. 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 that up with a “land rush” phase where registrants can get an earlier chance at any domain names 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. In principal, the mechanism for such a restriction is to apply for an exemption to ICANN's Registry Operator Code of Conduct – a legal framework that protects registrants in public TLDs and includes provisions about what registries may do with registrants' personal information and the types of agreements the registry may have with domain registrars. However, we find that many private TLDs with few registered domains do not use this mechanism in practice.

### 5.1.3 Promotions

For a registry operator for a new gTLD, the influx of new gTLDs creates some difficult marketing challenges. Even registrants looking for a domain in a new gTLD have hundreds of options. Registries can and do compete based on TLD string and price, but successful registries have learned that they need to aggressively target their chosen demographics. The tight competition has lead the new gTLD space to become inundated with promotions, usually including free domains in specialized circumstances. We must understand how registries sell domains to learn what the registrant intended, so we highlight some of the bigger promotions in the new gTLD space.

#### xyz

The xyz TLD is the largest in the new program and targets generic registrations as an alternative to com. In the middle of 2014, Network Solutions, a large registrar, began offering xyz domains for free to some of their customers. Owners of com domains found free xyz domains automatically added to their accounts 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 [34, 58].

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.

In our data set, 351,457 xyz domains (46%) 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, 82.0% 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.

### **science**

The science TLD allows generic registrations, but targets the scientific community. Starting with general availability on February 24, 2015, the AlpNames registrar began offering 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.

**realtor**

The National Association of Realtors owns the realtor TLD and targets accredited realtors, but also requires all registrants to prove they are members of their association [47]. 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.

### 5.1.4 Our Study of the New gTLDs

In the rest of this chapter, we describe the data and methodology with which we classified registration intent in the new TLDs. We describe our data sources and infrastructure in Section 5.2. Section 5.3 provides our methodology for categorizing domains based on their content, and in Section 5.4 we combine them into registration intent categories. Section 5.5 shows the impact of the new TLDs on registrations in the older TLDs. We hypothesize about registry revenue in Section 5.6 and provide profitability models based on several TLD and registry features. Section 5.7 describes our analysis of new domain registrations in Alexa and a well-known blacklist, and Section 5.8 concludes our work.

## 5.2 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.

### 5.2.1 Zone Files

Just as in biz and xxx, ICANN requires most registries to provide zone file access to others 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 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 seems 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. We simplify the zones and store all NS, A, and AAAA records on our HDFS cluster. We store the raw zones on our archive server for future use.

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

### 5.2.2 ICANN Public Data

ICANN requires each registry to provide a handful of summary reports. We have used most of them at some point while developing our methodology. The monthly transaction reports feature most heavily in our methodology and are comparable to those we used in xxx. 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.

In the registry agreements for the old TLDs, ICANN did not specify on which day of the month the registry should compile their monthly report data, an important detail when synchronizing report and zone file data in quickly growing TLDs. In xxx, we were able to empirically measure that ICM Registry used a day within the last week of the month. We assumed the registry used the simplest method, or the last day of the month. In the registry agreements for the new TLDs, ICANN explicitly required registries to make their reports accurate as of the last day of the month, so we do not need to do a similar estimate.

We also heavily used ICANN's New gTLD Current Application Status listing [40]. We used the data they provided to determine TLD status and registry information as the new TLDs worked through the application system.

### 5.2.3 Our TLD Set

We have focused our analysis on why registrants spend money on domains in the new TLD program. We only include results for TLDs that started general availability by the date of publication of ICANN's latest monthly registry reports, or January 31, 2015.

Our interest in private TLDs extends only to the defense of the TLD string itself (i.e., we want to know when a company felt it necessary to spend 185,000 USD in application fees, likely with additional legal fees, to defend an entire TLD). Since the registry, registrar, and registrant match for all domains in those TLDs, our interest in private TLDs stops there, so we exclude them from our analysis.

As mentioned in Section 5.1.1, many private TLDs do not file Code of Conduct Exemption Requests, and we found it difficult to classify TLDs as public or private. We solve this problem by checking public information about the start of general availability, as provided by several large domain registrars and nTLDStats [42], 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, and we use this as our indicator for the remainder of the project.

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 their own rules for sunrise and general availability that we found unclear even with the help of a native speaker.

After removing private and internationalized TLDs from those that began general availability before January 31, 2015, we end with a set of 290 new TLDs. Table 5.1 gives an overview of the largest TLDs in our set. The total set of TLDs includes English words like `bike` and `academy`, geographical regions like `berlin` and `london`, and organizations like `airforce` and `gop`. 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). When analyzing

**Table 5.1.** The ten largest TLDs in our public set with their general availability dates.

| *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 |

content, we restrict our activity to these 290 TLDs.

### 5.2.4 Active Web

For each domain in the zone file of a new gTLD, we make an HTTP request on port 80 with a crawler based on Firefox [33], an improved version of the crawler discussed in Section 4.2.3. Our browser-based 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 all page headers, the response code, and the redirect chain. We have improved and maintained our Web crawler primarily for this project, but other projects in our research group use it as well. We have worked hard to make our crawler easy to use by multiple projects simultaneously; it schedules jobs from multiple parties and writes the output to a separate location for each project.

Our primary data set for this chapter 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.

### 5.2.5 Active DNS

Every time we Web crawl a domain, we also perform a DNS query using a crawler developed for [28]. 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.

### 5.2.6 Pricing Data

Many aspects of our analysis focus on the economic impact of the new TLD program, a task that would be impossible without domain pricing information. As described in Section 2.1, 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 make claims about pricing per TLD, so we'd 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.

### 5.2.7 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.

### 5.2.8 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 [61]. 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.

## 5.3 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.

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 a huge collection of unlabeled Web pages crawled from the wild, we must learn from scratch to classify the domains.

The first step in our approach is to cluster Web pages with highly similar content. This procedure groups together duplicate and near-duplicate Web pages, which commonly arise when HTML source code 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; see [9] for more details of this approach.

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.

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 identified three broad categories for classifying domains according to their content: parked, content-free, and meaningful content. Our clustering approach was particularly effective for 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 for now. (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 that we clustered. Now equipped with a large number of labeled examples, we utilized 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 (according to 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 is intended to minimize false positives. Note that 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 of Web pages. 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. Ultimately, using technical data from both crawlers in addition to textual analysis of HTML content, we assign each domain to one of the following six categories:

**Invalid DNS** domains do not successfully resolve DNS queries.

**HTTP Error** domains have valid DNS, but return something other than HTTP 200 when queried.

**Parked** domains are owned by an ad network or for sale by their owners and typically

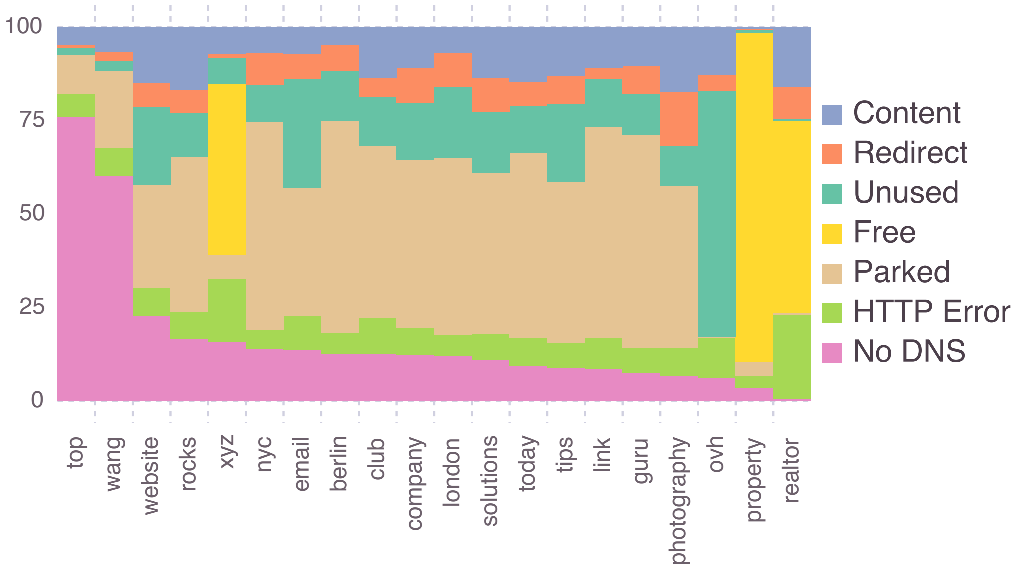


**Figure 5.1.** Our classifications separated by TLD for the 20 most common. We have sorted TLDs by fraction of invalid DNS to better highlight the category breakdowns of successful content.

return Web pages with mostly ads.

**Content-Free Web** domains host Web pages that are 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 legitimate Web content and are ready for users to visit.

Figure 5.1 shows our content classification for the 20 largest TLDs that allow public registrations in our data set. Table 5.2 shows the overall classification totals for all domains in the new TLD zones. We describe our classification methodology in more detail in the following sections. We performed our categorization in the order listed, so

**Table 5.2.** Overall content classifications for all domains in the zone file for the new public TLDs.

| *Content Category* | *Results* | |
|---:|---:|---:|
| No DNS | 567,390 | 15.6% |
| HTTP Error | 362,727 | 10.0% |
| Parked | 1,161,892 | 31.9% |
| Unused | 504,941 | 13.9% |
| Free | 432,323 | 11.9% |
| Redirect | 230,535 | 6.3% |
| Content | 378,401 | 10.4% |
| **Total** | 3,638,209 | 100.0% |

for instance 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 Invalid DNS

A large fraction of domains in the new gTLDs do not even resolve. Registrants purchase these domain names from a registrar and pay the yearly fee to keep them. Some 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 queries to its name server return `REFUSED` (which recursive resolvers usually report as `SERVFAIL` to the end user). Out of 3,638,209 domains in the new TLDs, we detect 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. These domains similarly do not resolve, but also do not appear in the zone files. As such, 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 (every domain a registrant pays a yearly fee to keep). 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.

During our previous study on the xxx TLD (Chapter 4), we found a large set of domains reserved by the registry that also appeared in the zone file. Since nobody registered them through a registrar, these domains did not appear in the ICANN reports, and we needed to account for them when calculating the number of non-resolving domains. The new gTLD program has hundreds of registries, and identifying these through common Web templates will not scale. However, registered non-resolving domains made up 85% of registered domains in xxx, so we cannot ignore their potential existence in the new TLDs.

The template for the registry agreement that new TLD operators sign declares that reserved names must either be inactive or registered through a registrar. Thus, they must appear in both the zone file and monthly reports, or in neither of them. However, each registry signs a separate agreement for each new TLD, and ICANN could include special provisions on reserved names for TLDs as contentious as xxx. We do not expect many registry-owned names, and not at quantities that will impact the remainder of our analysis, but cannot rule them out entirely.

We could verify that no registry-owned domains exist in the zone file but not in the ICANN reports, but doing so would require consideration of every domain name and is not scalable. Instead, we verify that these domains do not exist in quantities that would significantly impact our analysis. We do so by examining ownership information of the most common name servers in each TLD to verify that none of them are owned by a registry.

We convert each name server to its registered domain (e.g., `ns1djs.name.com` to `name.com`). We find the 50 most common name server domains across the new TLDs. Then we search each TLD for name servers that host at least 5% of the TLD's domains but do not appear in our global name server list, revealing an additional 99. We manually inspect their ownership information and found 7 name servers owned by a new registry. In every case, we find registration information about these domains in the ICANN monthly reports. Both our data and our interpretation of the base registry agreement suggest that at most a handful of domains appear in the zone files and not the ICANN monthly reports. Thus, we can find the number of registered domains with no entries in the zone with only summary information from these two sources.

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.

### 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.

We recieved 362,727 responses to our Web requests that we classified as HTTP errors. Table 5.3 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. We received responses with 43 unique HTTP status codes.[2]

[2] Six responses use the HTTP response code 418, an error code added as part of the Hyper Text Coffee Pot Control Protocol in a satirical RFC [18]. The return code means "I'm a teapot".

**Table 5.3.** The kinds of HTTP errors we get when querying Web pages.

| *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% |

### 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 previously created unsuccessful Web properties and parked them at the end 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.

We are aware of two previous studies that focus largely on parked domains and need to classify them as part of their work [3, 63]. Alrwais et al. focus on how parking programs operate and use domains from known parking name servers as their source. Vissers et al. focus on classifying parked domains, but use parking pages from known parking name servers as their inputs. In contrast, 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. There tend to be lots 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, including `sedoparking.com`. We use this method only for servers we are confident host solely parked domains. We start

**Table 5.4.** Our capture methods for parking and how many domains each identifies. 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.

| *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 | — | |

by taking the intersection of the different sets used by Alrwais et al. [3] and Vissers et al. [63]; 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 5.4 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 affirms that we have found the most common parking behaviors.

### 5.3.4 Content-Free Web

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 run by a large registry with instructions for the owner on how to develop their domain. Some are instead 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.

Content-free pages often come in bulk, so we identify them using our k-means classifier. With this technique, we find 504,941 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 5.1.3, get their own content classification. Most of these domains fall into the "Unused" category through a strict interpretation of our content categories, but the registrant plays a different role for these (which will be relevant when determining intent in Section 5.4). We cannot classify free `berlin` domains into this category because those were normal opt-in registrations with no standard page template.

Though not part of a promotion, the `property` TLD contains mostly domains owned by its registry, Uniregistry. 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. In total, we find 432,323 free domains in the new TLD program, (11.9%).

**Table 5.5.** The mechanisms domain owners use to redirect to a different domain. Most domain owners use only browser-level redirects, but frames are still common.

| *Mechanism* | *Domains* | *Coverage* | *Unique* |
|---:|---:|---:|---:|
| CNAME | 1,997 | 0.9% | 725 |
| Browser | 206,118 | 89.4% | 199,046 |
| Frame | 29,509 | 12.9% | 23,677 |
| **Total** | 230,535 | — | |

### 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 5.5 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`.

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. While most people do not think of these as redirects, they provide the same function: a user visits one domain on their browser, and sees content from another. Since these 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. 49% of the filtered DOMs have a string length of less than 55 characters, but show spiky 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, new-TLD, or old-TLD.

Though each of these domains has some form of redirect when fetching Web content, redirects to a page under the same domain name are not of interest to our registrant-focused analysis. These redirects instead just reveal something about the structure of the Web page itself. Similarly, we cannot make any strong claims about redirects to a hard-coded IP address. 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 230,535 off-domain redirects in our data set, or 6.3% of all domains in the new TLD zone files.

### 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 378,401 domains (10.4%) fall into this category. By comparing this category with the previous, we find that 37.9% of the 608,936 domains with real content redirect to a different domain to serve it.

## 5.4 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 purchase domain names to make money 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.

For domains we felt comfortable classifying registrant intent, Table 5.6 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.

**Table 5.6.** Registration intent categorizations for the new public TLDs.

| *Intent* | *Results* | |
|---:|---:|---:|
| Primary | 378,401 | 14.9% |
| Defensive | 1,005,109 | 39.5% |
| Speculative | 1,161,892 | 45.6% |
| **Total** | 2,545,402 | 100.0% |

### 5.4.1 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.

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 230,535 defensive redirects, we find defensive registrations of 1,005,109 domains in the new TLDs.

### 5.4.2 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 mostly unique content.

### 5.4.3 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 strictly 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. We classify parked domains as speculative and non-resolving domains as defensive based on this distinction.

## 5.5 Impact on Old TLDs

We next look at the impact of the new TLDs on registrations in the old TLDs. 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

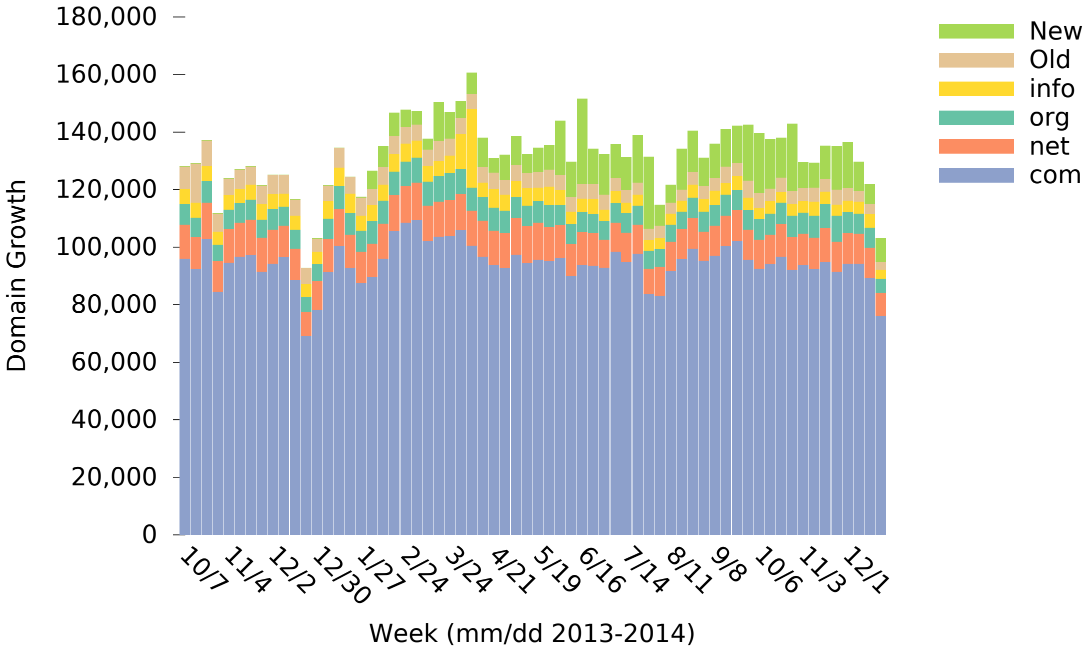


**Figure 5.2.** Number of new domains per day. Bars indicate the average rate for each week.

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.

We compare registrations in the old and new TLDs in a variety of ways. First, we look solely at registration volume to answer the question above at a high level. Then we compare our content categorization of the new TLDs to a random sample of the old TLDs to understand what kind of content exists in each. Finally, we compare the content in the new TLDs with new registrations in the old to learn how new primary, defensive, and speculative registrations in all TLDs differ.

### 5.5.1 Registration Volume

Figure 5.2 shows the number of new domain registrations per week broken down into various categories. Days which we did not have access to the zone files resulted in

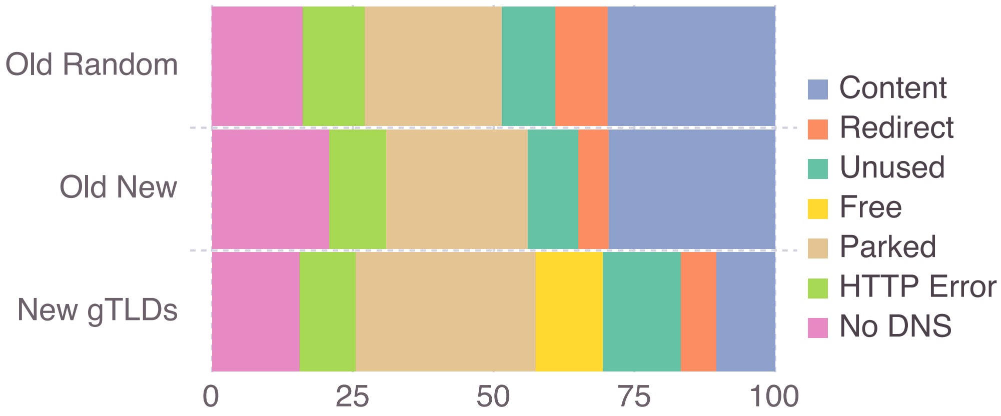


**Figure 5.3.** Our overall classifications for domains in the new TLDs, a random sample of the old TLDs, and a month of new domain registrations in the old TLDs.

slight drops in the graph. We show the most active old TLDs individually, the remaining old TLDs grouped into "Old", and all registrations in new TLDs in "New" (Table 5.1 breaks down registrations in the new TLDs in more detail).

Overall, the results show that the new TLDs are still a small portion of all registrations in the gTLDs (e.g., registrations in com continue to dominate), and that the new TLDs generally increase the total number of registrations rather than shift focus from old to new TLDs. Such registration behavior is consistent with substantial domain speculation in the new TLDs (Section 5.4.3).

### 5.5.2 Domain Classifications

We compare domain classifications using three distinct data sets. The first includes all domain registrations in the new TLDs as of February 3, 2015. The second includes 3 million domains from the old TLDs defined in Section 5.2.1 chosen uniformly at random. The third includes all domains in the same set of old TLDs that were added to their respective zone files during December 2014. (Delays in our com processing pipeline prevented us from using a more recent data set.)

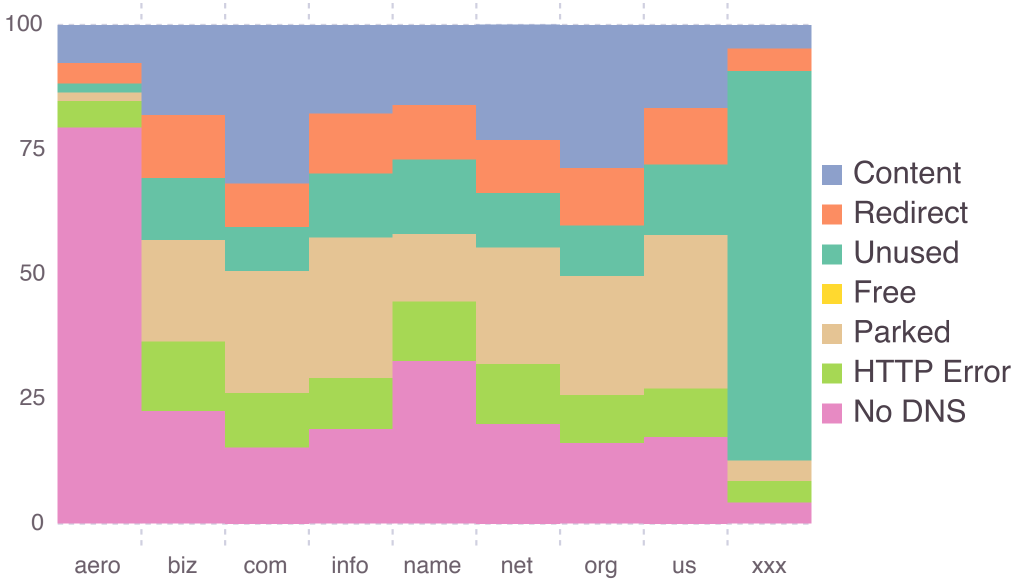


**Figure 5.4.** Our classifications for a random sample of 3 million active domains in the old TLDs.

Our overall content classifications for all three sets are visible in Figure 5.3. 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 15–20% 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.

### 5.5.3 Old TLD Categorizations

Figure 5.4 shows the content categories for our random sample of current registrations in old TLDs, and Figure 5.5 shows the same for new domain registrations in the old TLDs. We see some large changes in registration types between the two sets. Perhaps the most telling is xxx, which contains mostly ICM Registry's reserved domains in the random set but significantly more content in the new registrations. While at the

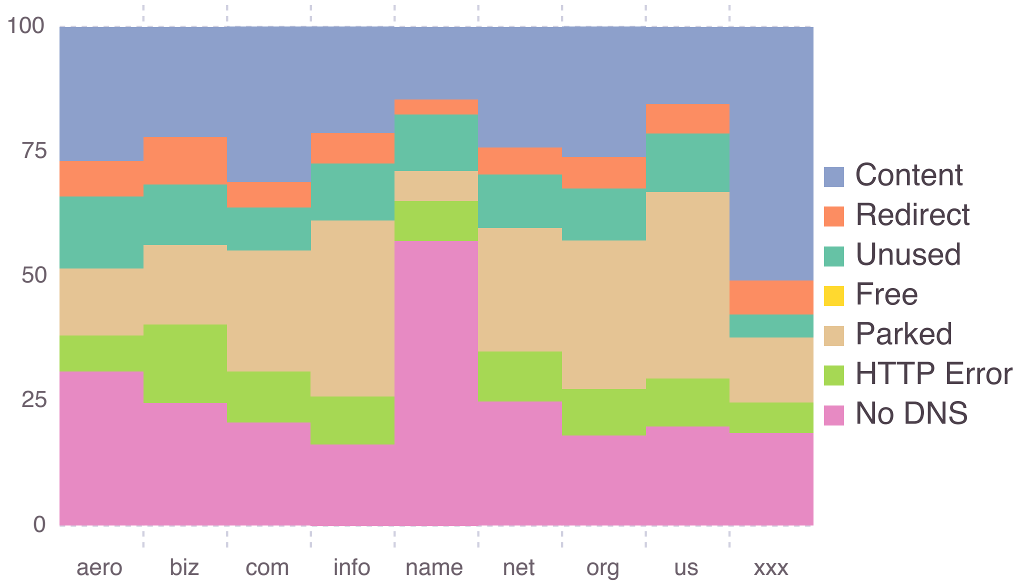


**Figure 5.5.** Our classifications for all new domains in the old TLDs between December 14, 2014 and January 14, 2015, well into the existence of the TLD program.

outset this looks promising, xxx has already had over three years to grow, and the lack of content-oriented domains in the random sample suggests that the low volume of new registrations do not significantly influence the overall classification.

## 5.6 Registration Costs

In previous sections, 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.

### 5.6.1 Registry Financials

Using the methodology described in Section 5.2.6, we obtained pricing information for 2006 (TLD, registrar) pairs, which matches 73.8% of all domain registrations. In

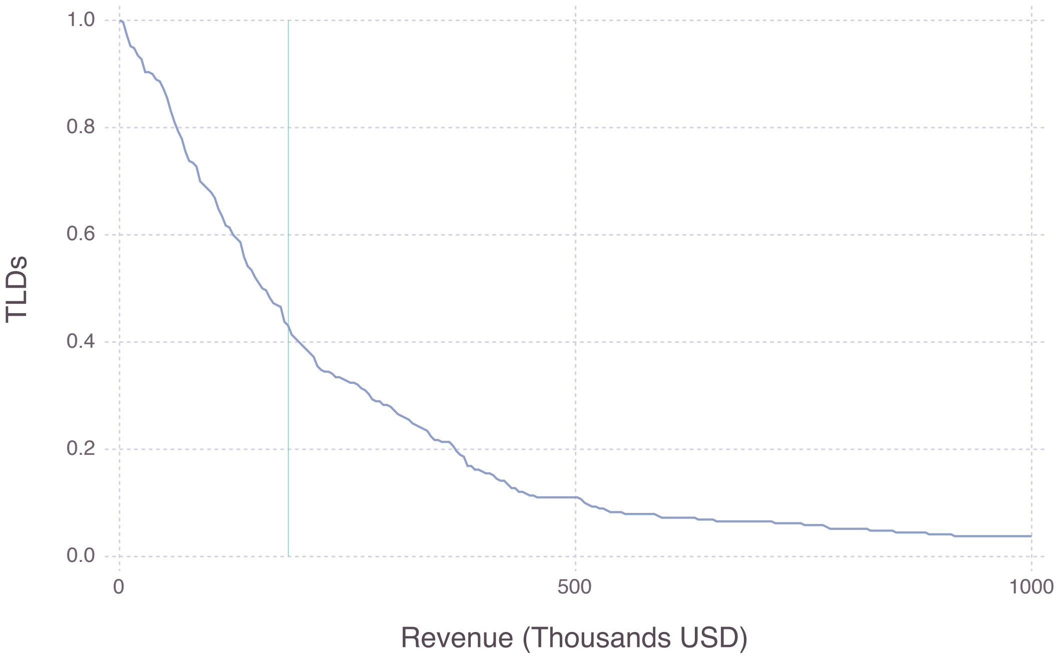


**Figure 5.6.** New gTLD program revenue as a flipped CDF.

only four TLDs do we record prices from fewer than three registrars; in each case the one or two registrars we do record include at least 97.5% of all domains. In the 26.2% of domain registrations for which we do not have matching data, we use the median price for the TLD.

Figure 5.6 shows a flipped 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 [39]; 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 represents the minimum 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 any of ICANN's contention processes. 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. We do not know how much it costs to run a registry, but if 500,000 USD is a good estimate[3] then just over 10% of TLDs are profitable under our model.

Revenue from domain registrations does not all go to the registry. Instead, registries and registrars split revenue based on a previously agreed upon model. Verisign makes \$7.85 USD per `com` registration [8] and \$6.79 per `net` registration [37]. 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.[4]

Unfortunately, the new registry agreements do not specify maximum wholesale prices, only the fees the registry must pay to ICANN. For calibration, we can get a handful of prices through registry-reported earning data. One of the largest backend registry providers, Rightside, is funded through private investors, and has released some revenue statistics online in a presentation meant for investors and analysts [53]. They claim end-of-November wholesale and total revenue numbers for five TLDs, two of them aggregated. Our estimate is too low for `reviews`,[5] but our other estimates over-

[3] While still conjecture, we do have one other piece of evidence that suggests 500,000 USD is a reasonable estimate for the cost of introducing a TLD. Some TLDs have already gone up for auction, like `reise` [50] and `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.

[4] Registries can offer "bulk discounts and marketing support and incentive programs" to registrars but must offer similar terms to all registrars [8].

[5] 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` [35] and found that the price to registrants of a `review` domain has halved. We do not know if this reflects a

estimate 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.

Figure 5.6 represents a low estimate of registry pricing information and is likely accurate within a factor of two or three. We likely underestimate registry expenses by a similar amount.

### 5.6.2 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] because it provides the first chance for registrants in the new TLDs to renew their domain names.[6] Donuts, the largest registry with over a hundred new TLDs, published statistics on renewal rates for their earliest TLDs [11, 51], likely in an attempt to attact registrars and investors [65]. However, Donuts limited their analysis to their own TLDs, and also did not provide numbers past 26 days.

Figure 5.7 shows a histogram of renewal rates by TLD. We only performed our analysis on TLDs where at least 100 domains completed a whole year of registrations and 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%.

### 5.6.3 Future Profit Modeling

In this section, we take a look at registry profitability using a variety of parameters. These models use our "best guesses" for each of the issues we have raised above. We

reduction in its wholesale price or a promotion.

[6] 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.

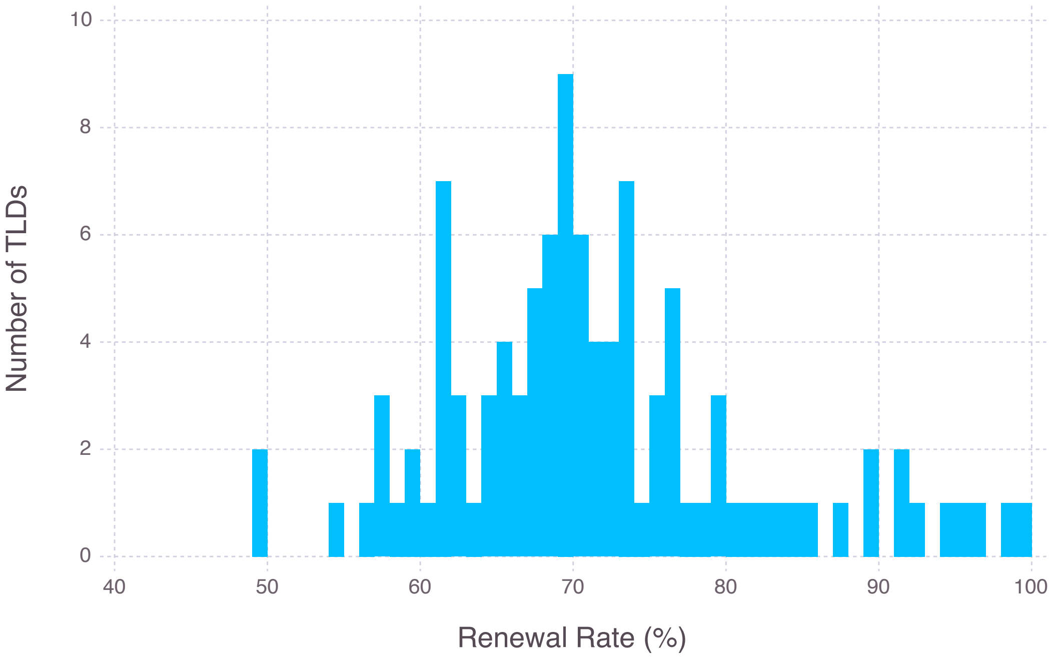


**Figure 5.7.** A histogram of renewal rates per TLD.

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 under four different models in Figure 5.8. Our models show 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 creating a registry. Our second parameter reflects varying renewal rates, and we base our choices off of our graph from Section 5.6.2. We show models with renewal rates of 57% and 79%. These reflect a lower and higher value surrounding the typical range and give an idea of how renewal rates impact the model.

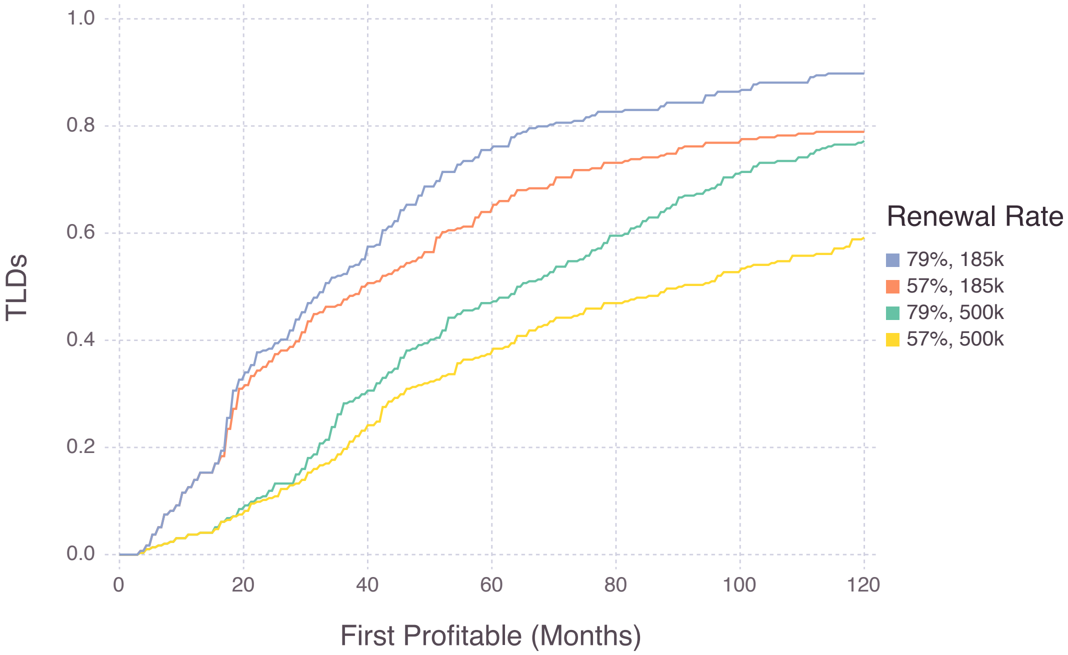


**Figure 5.8.** Registry profitability over time under different revenue models. A point on a line indicates the fraction of TLDs that were profitable within the given time since general availability.

For each TLD, we collect registration volume data with the ICANN monthly reports. We consider TLDs for which we have three 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.

Figure 5.8 shows that the initial cost plays a much larger role than the renewal rate in the short term, but that years out renewal rates still matter. 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.

Our expectation when modeling TLD profitability was that we might find some

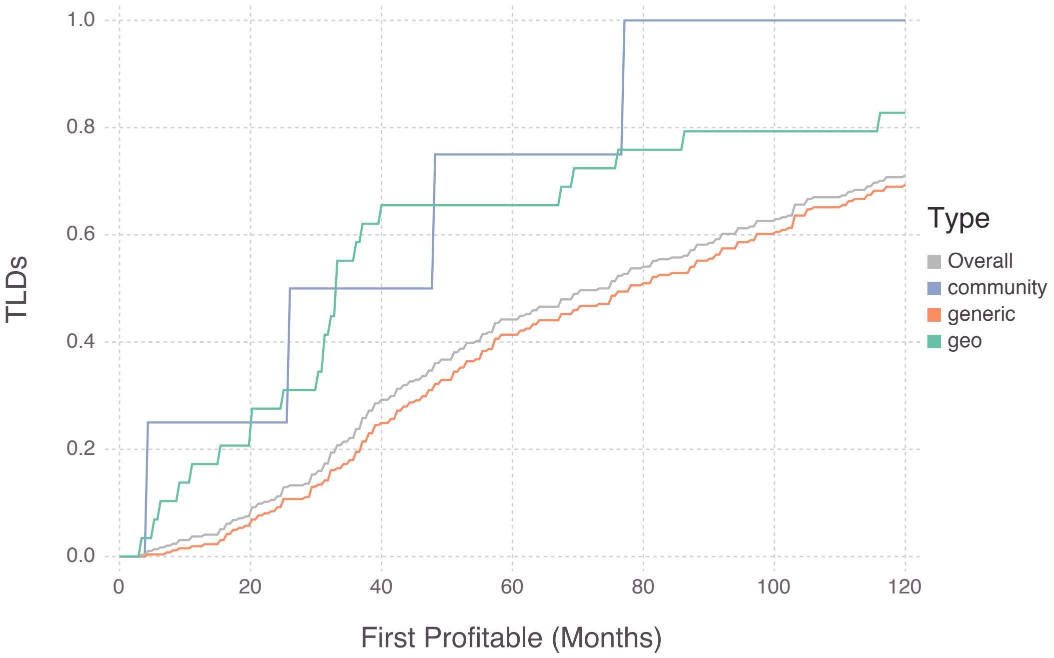


**Figure 5.9.** Modeling profitability by type of TLD. The gray line represents the aggregate, and the colored lines represent the set of TLDs of the indicated type.

metric, be it registry or lexical length, that would separate profitable and unprofitable TLDs. 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 “geo”); 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 biggest differentiators, type and registry, below.

Figure 5.9 shows variations in profitability by type.[7] The gray line represents the overall profitability CDF. It is equivalent to the profitability PDFs represented by

[7] This graph, like the data for the rest of our paper, does not include TLDs in the second largest category, private brand TLDs. As described in Section 5.2.3, we exclude brand TLDs because they are not monetized through public registrations.

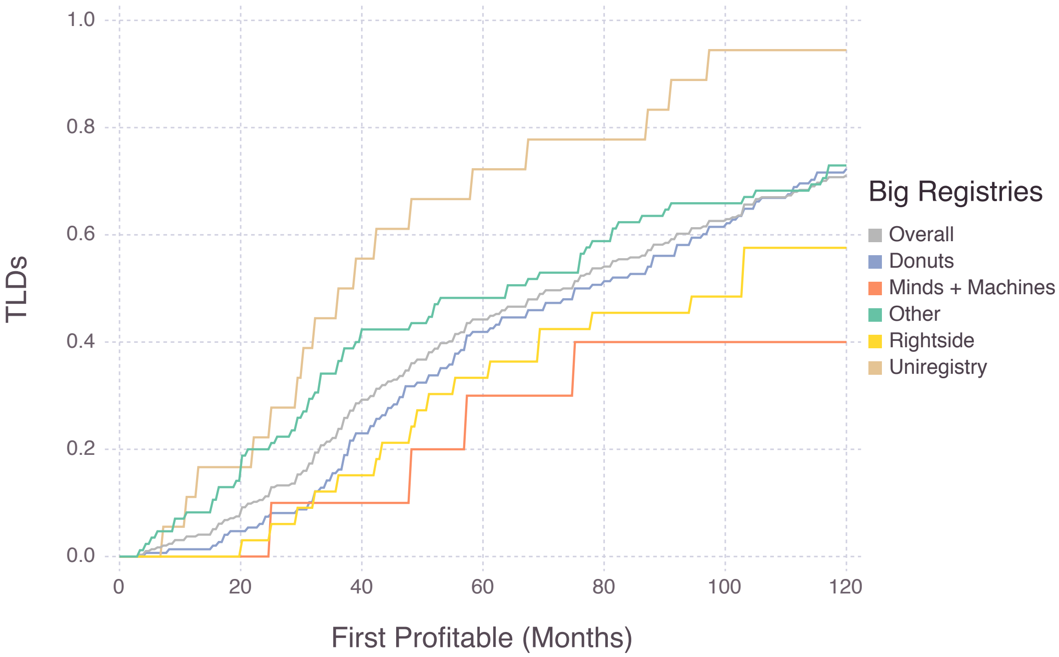


**Figure 5.10.** Modeling profitability by registry for the registries with the most TLDs. The gray line represents the aggregate, with colored lines representing individual registries.

Figure 5.8 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 5.10 shows variations in profitability by registry. Of the large registries, only Uniregistry TLDs become profitable sooner than the average. Instead, it appears as if 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.

While Figure 5.10 shows some unexpected results, remember that our model does not incorporate premium domain name sales as anything other than a normal domain

**Table 5.7.** The rate at which new domains in the old and new TLDs appear in blacklists and Alexa.

| | *New*<br>Per 1000 | *Old*<br>Per 1000 |
|---|---|---|
| URIBL | 7.03 | 3.31 |
| Alexa 1m | 0.88 | 2.43 |

sale. Any registries that receive significant revenue from premium domain names in comparison to normal sales may reach profitability significantly sooner than indicated by our model. Registries that do not focus on premium domain names will not receive the same artificial deflation.

## 5.7 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 how often real 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 thousand new registrations.

Table 5.7 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 old TLDs. While this is a notable difference, it is also consistent with the proportion of primary registrations described in Section 5.5.2.

We use a similar methodology with the URIBL blacklist as an inference method for abusive behavior. We use the same sets of newly registered domains. In contrast to Alexa, we select blacklist data that is contemporaneous with our registration data because we suspect blacklists will add abusive domains as soon as possible. Table 5.7 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 5.2.6. 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.

## 5.8 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 demonstrated in previous TLD expansions, 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.

This chapter showed 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, 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.

Chapter 5, in part, has been submitted for publication of the material as it may appear in *Proceedings of the International Measurement Conference*. Tristan Halvorson, Matthew F. Der, Ian Foster, Stefan Savage, Lawrence K. Saul, and Geoffrey M. Voelker, ACM, October 2015. The dissertation author was the primary investigator and author of this paper.

# Chapter 6

# Conclusion

The sale of domain names rests on a curious juncture of technology and business. On the surface, DNS simply provides a mechanism for converting human-memorable strings to host identifiers. The space of potential names seems vast at first glance, but much like the real estate market, each product is different and the better names go to those willing to pay. In an ideal world, registrants would acquire domain names for actual use, such as to promote unique content or for their private needs. Instead, many domain registrations go to resellers and companies that want to defend their trademarks. The addition of new TLDs to the Domain Name System exacerbates this problem, either because users perceive identical second-level domains in different TLDs as related, or because corporations believe they might.

In an effort to understand how commonly registrants defensively purchase domains, this dissertation provides a methodology to determine registration intent for domain names, and then applies it at scale to all domains in hundreds of TLDs. We provide data and lessons learned on some older TLDs and use them to scale our approach to all TLDs in ICANN's New gTLD Program. We find widespread instances of defensive registrations. Some TLDs contain more defensive registrations than others, as demonstrated by our analysis of xxx, and while rollout phases provide companies with ample opportunities to defend their names, their cost structures reflect the needs of the registry more than the needs of the brandholders.

## 6.1 Future Directions

ICANN's New gTLD Program is in full swing, and many TLDs from the first round of applications remain to be delegated. ICANN plans to open a second round of applications at that time. The lessons we as a community have learned about how people use domain names should absolutely feature in ICANN's discussions between now and then, but there is a lot of time remaining for the market to evolve, and for newer registries to provide alternative options on how to limit the need for defensive registrations. Additional studies between now and then can only help ICANN and potential new registries make better decisions about their future.

Besides just performing the same study in new TLDs as they roll out, the additional time provides some other new opportunities we were not able to take advantage of in our analysis of the topic. For instance, we would like to learn how many registrants renew their domains, and if renewals provide any insights on which types of names provide long-term value to the registry. Do primary registrants hang onto their names longer than speculative or defensive ones? Do organizations defensively register names for a few years, and then decide the cost is not worth the benefit and let their registrations lapse?

Finally, extra time would greatly help in refining our pricing models. We assume registration rates are linear and estimate future registrations based on the last two months in the ICANN reports. These reports come out three months after the data they represent, and we only have sufficient data to model the price at all for around two hundred of them. Additional time would allow us to apply our pricing models to more of the TLDs, but would also provide additional data for our existing models to refine our expectations. Perhaps after the initial availability phase new registrations speed up or slow down, possibly in relation to other factors like the TLD’s length. The work presented in this dissertation starts the journey to understanding these topics. We encourage ICANN to make policy decisions fueled by real data, and to this end urge anyone interested in the domain name market to help out.

## 6.2 Final Thoughts

This research focuses on increasing the quality of domain name registrations, and to this goal sees defensive registrations as less valuable than primary registrations. However, one should not blame registrants for defending their names in ways they see necessary, especially given the prevalence of cybersquatting and phishing on the Internet. Instead, ICANN should consider these behaviors and how to limit them in future TLDs, or cease the introduction of new TLDs altogether.

# Bibliography

[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] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic Clustering of the Web. *Computer Networks and ISDN Systems*, 29(8):1157–1166, 1997.

[7] D. Carlton. Report of Dennis Carlton regarding ICANN's proposed mechanism for introducing new gTLDs. http://www.icann.org/en/topics/new-gtlds/carlton-re-proposed-mechanism-05jun09-en.pdf, June 2009.

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

[9] 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.

[10] Domain Name Supporting Organization. Working Group C – creation of new gTLD. http://www.dnso.org/dnso/notes/19990625.NCwgc.html, June 1999.

[11] Donuts Renewal Trends: First Definitive Report. http://www.donuts.domains/donuts-media/blog/donuts-renewal-trends-first-definitive-report.

[12] R. T. Fielding, J. Gettys, J. C. Mogul, H. F. Nielsen, L. Masinter, P. J. Leach, and T. Berners-Lee. Hypertext Transfer Protocol — HTTP/1.1. RFC 2616, The Internet Society, June 1999.

[13] Find and Buy Your .XXX Domain Name. http://www.buy.xxx.

[14] GAC Early Warnings. https://gacweb.icann.org/display/gacweb/GAC+Early+Warnings.

[15] GoDaddy Registration Prices on 2011-09-28. http://web.archive.org/web/20110928063106/http://www.godaddy.com/tlds/xxx-domain.aspx?tld=xxx&ci=51450.

[16] How do future brands protect themselves? https://member-icmregistry.custhelp.com/app/answers/detail/a_id/26/.

[17] How do I acquire NON-Renewed names? https://member-icmregistry.custhelp.com/app/answers/detail/a_id/85/.

[18] Hyper Text Coffee Pot Control Protocol (HTCPCP/1.0). http://tools.ietf.org/html/rfc2324.

[19] ICANN Options Following the IRP Declaration on ICM's .XXX Application. http://www.icann.org/en/irp/icm-v-icann/draft-options-post-irp-declaration-26mar10-en.pdf.

[20] ICM Registry Announces the Successful Conclusion of the .XXX Founders Program. http://www.icmregistry.com/press/icm-registry-announces-the-successful-conclusion-of-the-xxx-founders-program/.

[21] ICM Registry closes sunrise period with 80000 applications for XXX domain names. http://www.icmregistry.com/press/icm-registry-closes-sunrise-period-with-80000-applications-for-xxx-domain-names/.

[22] S. S. International. Evaluation of the new gTLDs: Policy and legal issues. http://www.icann.org/en/tlds/new-gtld-eval-31aug04.pdf, July 2004.

[23] Internet Assigned Numbers Authority. Iana report on establishment of the .biz and .info top-level domains. http://www.iana.org/reports/2001/biz-info-report-25jun01.html, June 2001.

[24] Internet Corporation for Assigned Names and Numbers. Domain Name Supporting Organization formation concepts. http://www.icann.org/en/meetings/singapore/dnso-formation.htm, March 1999.

[25] Internet Corporation for Assigned Names and Numbers. ICANN accredits nnew top-level domains—.biz and .info registration process to begin this summer. http://www.icann.org/en/announcements/icann-pr15may01.htm, May 2001.

[26] M. L. Katz, G. L. Rosston, and T. Sullivan. An Economic Framework for the Analysis of the Expansion of Generic Top-Level Domain Names. http://www.icann.org/en/topics/new-gtlds/economic-analysis-of-new-gtlds-16jun10-en.pdf, June 2010.

[27] Letter from FTC to ICANN, May 27, 2015. http://domainincite.com/docs/Chairwoman%20Ramirez%20ICANN%20Response%20Letter%205.27.15.pdf.

[28] 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.

[29] Manwin Licensing International v. ICM Registry, et al. http://www.icann.org/en/news/litigation/manwin-v-icm.

[30] Manwin Licensing International's Complaint, Case CV 11-9514-PSG. http://www.icann.org/en/news/litigation/manwin-v-icm/complaint-16nov11-en.pdf.

[31] Monthly Registry Reports. http://www.icann.org/en/resources/registries/reports.

[32] T. Moore and B. Edelman. Measuring the Perpetrators and Funders of Typosquatting. In *Proceedings of the 14th International Conference on Financial Cryptography and Data Security*, 2010.

[33] Mozilla Firefox. http://www.mozilla.org/en-US/firefox/new.

[34] My Interview with Daniel Negari Addressing Reported Inflated .xyz Registrations. http://www.ricksblog.com/2014/06/interview-daniel-negari-addressing-inflated-xyz-registrations/.

[35] name.com Pricing for Common TLDs. https://web.archive.org/web/20141128024531/http://www.name.com/pricing.

[36] National Telecommunications and Information Administration. Statement of policy on the management of internet names and addresses. http://www.ntia.doc.gov/federal-register-notice/1998/statement-policy-management-internet-names-and-addresses, June 1998.

[37] .net Fees. https://www.icann.org/sites/default/files/tlds/net/net-fees-01feb15-en.pdf.

[38] NeuLevel, Inc. .biz proof of concept report to ICANN. http://forum.icann.org/lists/gnso-pro-wg/msg00020.html, 2004.

[39] New gTLD Applicant Guidebook. https://newgtlds.icann.org/en/APPLICANTS/AGB.

[40] New gTLD Current Application Status. https://gtldresult.icann.org/application-result/applicationstatus.

[41] New gTLD sales miss ICANN estimates by a mile. http://domainincite.com/18857-new-gtld-sales-miss-icann-estimates-by-a-mile.

[42] new gTLDs Launches. https://ntldstats.com/launch.

[43] ODP – Open Directory Project. http://www.dmoz.org, September 2011.

[44] PHPWhois. http://sourceforge.net/projects/phpwhois/.

[45] Porn domain .xxx blocks use of celebrity names. http://www.guardian.co.uk/technology/2011/sep/07/porn-domain-xxx-celebrity-names.

[46] Products and Suggested Retail Pricing. https://www.registry.sucks/products.

[47] .realtor Fact Sheet. http://www.realtor.org/sites/default/files/handouts-and-brochures/2014/DotREALTOR-Launch-Factsheet.pdf.

[48] Reconsideration Request 00-15. http://archive.icann.org/en/committees/reconsideration/rc00-15-1.htm.

[49] Registering .XXX Domain Names. http://support.godaddy.com/help/article/5789/registering-xxx-domain-names.

[50] .reise to start at $400k in no-reserve auction. http://domainincite.com/17988-reise-to-start-at-400k-in-no-reserve-auction.

[51] Renewal Trends: Day 26. http://www.donuts.domains/donuts-media/blog/renewal-trends-day-26.

[52] Request for Independent Review Process by ICM Registry, LLC. https://www.icann.org/en/system/files/files/icm-irp-request-06jun08-en.pdf.

[53] Rightside Analyst and Investor Day 2014, slide 104. http://edge.media-server.com/m/p/f9o6abq7.

[54] SIE Passive DNS. https://archive.farsightsecurity.com/SIE_Channel_202/.

[55] .sucks registrations begin soon–at up to $2,500 per domain. http://arstechnica.com/information-technology/2015/03/sucks-tld-to-accept-sunrise-registrations-soon-but-theyll-be-pricey/.

[56] T. Telegraph. Top 10 most expensive domain names. http://www.telegraph.co.uk/technology/news/7412544/Top-10-most-expensive-domain-names.html.

[57] The ABC&Ds about .XXX. http://www.icmregistry.com/about/policies/abcd/.

[58] 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.

[59] ToysRUs pays $5m for toys domain. http://news.bbc.co.uk/2/hi/technology/7923433.stm.

[60] Tweet by ICM Registry on the @dotXXX Twitter Account. https://twitter.com/dotXXX/status/144495755946762240.

[61] URIBL. http://uribl.com/about.shtml.

[62] US government finally lets ICANN go. http://www.zdnet.com/article/us-government-finally-lets-icann-go/.

[63] 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.

[64] J. Weinberg. Report (part one) of Working Group C of the Domain Name Supporting Organization Internet Corporation for Assigned Names and Numbers. http://www.dnso.org/dnso/notes/20000321.NCwgc-report.html.

[65] Why Donuts is revealing domain name renewal rates. http://domainnamewire.com/2015/03/31/why-donuts-is-revealing-domain-name-renewal-rates/.

[66] .XXX Adult Performer Program. http://performers.icmregistry.com/.

[67] .XXX Agreement Appendix 3: Zone File Access Agreement. http://www.icann.org/en/about/agreements/registries/xxx/appendix-3-31mar11-en.htm.

[68] .XXX Agreement Appendix 4: Registry Operator's Monthly Reports. http://www.icann.org/en/about/agreements/registries/xxx/appendix-4-31mar11-en.htm.

[69] .XXX Agreement Appendix 6: Schedule of Reserved Names. http://www.icann.org/en/about/agreements/registries/xxx/appendix-6-31mar11-en.htm.

[70] .XXX Announces Launch of Pre-Paid Adult Performer Program. http://www.icmregistry.com/press/xxx-announces-launch-of-pre-paid-adult-performer-program/.

[71] .XXX Launch Plan and Related Policies. http://www.icmregistry.com/about/policies/launch/.

[72] .XXX Registry Agreement. http://www.icann.org/en/about/agreements/registries/xxx/xxx-agreement-31mar11-en.htm.

[73] J. Zittrain and B. Edelman. Survey of Usage of the .BIZ TLD. http://cyber.law.harvard.edu/tlds/001/, June 2002.