JOURNAL OF INFORMATION POLICY 3 (2013): 464-484.

AN EMPIRICAL ANALYSIS OF INTERNET TOP-LEVEL DOMAIN POLICY

BY TOM NICHOLLS*

The process of introducing new top-level domain names (TLDs) for use on the Internet, which is managed by the Internet Corporation for Assigned Names and Numbers, is fraught with rhetoric about an increase in costs and abuses, or an increase in choice and reduction in scarcity. Mr. Nicholls argues that there has been very little empirical research on the potential impact of new TLDs, and analyzes issues of cost and scarcity in domain name registration practices. Mr. Nicholls concludes that some of the perceived threats of new TLDs have not manifested though others have, and that data-driven research is necessary.

INTRODUCTION

The Domain Name System (DNS) is a critical part of the global Internet infrastructure. Its authoritative root zone database, a master list of all valid top-level domain names (TLDs), is a single point of control and potentially of failure. The contents of the root have been managed by the Internet Corporation for Assigned Names and Numbers (ICANN), under contract to the U.S. Department of Commerce, since 1999. Ever since ICANN's creation, there has been extensive debate concerning the merits of adding a new series of TLDs to the root for global use (known as generic TLDs or gTLDs) – alongside the familiar .com, .org, and .net, these would include country code TLDs (ccTLDs) like .uk and .fr, and newer generic names such as .aero, .coop, and .biz.

The DNS underpins a multimillion dollar industry buying, selling, and maintaining domain names. Despite this, there has been little empirical research on the merits of expanding the number of generic TLDs. This is despite the unusual economic structure of the domain name business (very few industries have the option of costlessly extending the range of "products" available) and ICANN's unusual regulatory structure.

In domain name allocation, ICANN has a role variously described as "quasi-regulatory"[1] and as the "de-facto regulator."[2] ICANN has also never been far away from controversy. The circumstances of its creation were the subject of extensive debate, its initial flirtation with democratic processes was

* Doctoral candidate, Oxford Internet Institute, University of Oxford.

[1] Berkman Center for Internet and Society, "Accountability and Transparency at ICANN: An Independent Review," white paper, Oct. 20, 2010, accessed Nov. 11, 2013, http://www.icann.org/en/about/aoc-review/atrt/review-berkman-final-report-20oct10-en.pdf, 1.

[2] A. Michael Froomkin and Mark A. Lemley, "ICANN and Antitrust," *University of Illinois Law Review* 2003 (2003): 2.

quickly abandoned,[3] and its legitimacy was challenged by many governments who are dissatisfied with a U.S. corporation being responsible for the governance of the Internet's infrastructure.[4] As a consequence, decision-making structures in ICANN are complex, political, and plagued with controversy.[5]

This article aims to answer some of the unanswered empirical questions. It takes a random sample of domains registered in the gTLDs. Using data from the WHOIS ownership database, DNS records, and the contents of websites, the extent and cost of both defensive registration and domain squatting are quantified in each TLD. Implications for domain policy are then discussed.

BACKGROUND AND RATIONALE

> "It is not a secret that the questions of whether, how and when new gTLDs should be added have attracted a diversity of views, if not sharply divided views. At one end of the spectrum, certain Internet constituencies have maintained that the Internet should be an open system and that, at least in principle, any person should be able to introduce a new top-level domain, leaving the market to be the ultimate arbiter of its success. At the other end of the spectrum, some stakeholders have expressed strongly the view that no new gTLDs should be added, at least at this stage. Among the reasons in support of this latter position is a belief that there is currently no demonstrated need for additional name space and that adding new gTLDs will aggravate intellectual property problems and create consumer confusion."[6]

There are two competing conventional wisdoms in the gTLD expansion debate, identified by the World Intellectual Property Organization in the quotation above. One is that expansion would be a burden on trademark holders, with no benefit for consumers. The second is that artificial restriction of the number of top-level domains is harmful, and that the public would be better off with a much larger number of TLDs available for registration.[7] The ongoing debate between these two positions has been one of the factors delaying the process of gTLD expansion for twelve years. Although it was one of the first items on ICANN's agenda at that body's 1999 creation, final approval for a major expansion was only given in 2011.[8] Despite the introduction of new gTLDs in two small rounds in

---

[3] Andrew D. Murray, *The Regulation of Cyberspace: Control in the Online Environment* (Abingdon, U.K.: Routledge-Cavendish, 2007).
[4] Geoff Huston, "Opinion: ICANN, the ITU, WSIS, and Internet Governance," *The Internet Protocol Journal* 8, no. 1 (2005): 15-28.
[5] Milton L. Mueller, *Ruling the Root: Internet Governance and the Taming of Cyberspace* (Cambridge, MA: Massachusetts Institute of Technology Press, 2002).
[6] World Intellectual Property Organization, "The Management of Internet Names and Addresses: Intellectual Property Issues," Final Report of the WIPO Internet Domain Name Process, Apr. 30, 1999, accessed Nov. 11, 2013, http://www.wipo.int/amc/en/processes/process1/report/finalreport.html, ¶ 307.
[7] See for example Dennis Carlton, "Preliminary Report of Dennis Carlton Regarding Impact of New gTLDs on Consumer Welfare," white paper, Compass Lexecon, Mar. 2009, accessed Nov. 11, 2013, http://archive.icann.org/en/topics/new-gtlds/prelim-report-consumer-welfare-04mar09-en.pdf.
[8] Berkman Center for Internet and Society, 2.

the intervening period, very little research has been done on the impact of these changes, or the prospective impact of a broader expansion.[9] Although ICANN has now decided in favor of large-scale gTLD expansion, there is still an ongoing debate about its merits.

The U.S. Chamber of Commerce argues that "the proposed gTLD program […] will compel businesses to invest millions of dollars in defensive domain registrations and litigation."[10] This claim, along with others made in the gTLD debate, is empirically testable. To what extent are businesses currently investing money in defensive domain registrations, especially in the new TLDs created in the two previous rounds of ICANN expansion? What is the character of the domain name market? Is it characterized by scarcity of names, increasing the demand for expansion to new TLDs; or does it have plenty of room for expansion? Are there multiple good-faith users of similar names across different existing TLDs, suggesting that expansion would result in an increase in available names and little harm to existing registrants, or are the majority of co-registered names parasitic on the major user of that domain? Is there, in fact, suppressed demand for new TLDs at all? These are all questions that can be answered empirically.

Unfortunately, and despite the decade or more during which gTLD expansion has been considered, there is little empirical academic work on the impact of new TLDs. Because much of the work undertaken has been in industry studies and reports from international organizations, it will be helpful to summarize the main contributions here.

Two papers by Dennis Carlton (commissioned by ICANN) review the economic justification for the expansion of gTLDs by looking at consumer welfare and the merits of a price cap for new gTLD registries.[11] The first paper largely applies theoretical microeconomics to gTLDs with no empirical analysis. The second paper, on consumer welfare, is somewhat more detailed but is still short on empirical data. Rather than calculating the extent of defensive registrations, Carlton simply writes that "claims that the introduction of new gTLDs will necessitate widespread defensive registrations appear to be exaggerated and are inconsistent with the oft-noted observation that there have been a limited number of registrations on gTLDs introduced in recent years."[12]

The arguments of many of the critical papers are themselves poorly grounded in the empirical sense. For example, a reply paper commissioned by AT&T argues that defensive registrations are a serious DNS policy issue by assessing a non-random sample of five multinational companies. These are then used to challenge both the cost of defensive registrations across the whole domain name space, and later Carlton's entire price cap report on the basis that defensive registrations (demonstrated to be

[9] Andrew Mack, "Andrew Mack Comments on Dennis Carlton Report Regarding Impact of New gTLDs on Consumer Welfare," Apr. 17, 2009, accessed Nov. 11, 2013, http://forum.icann.org/lists/competition-pricing-prelim/msg00019.html.
[10] Chamber of Commerce of the United States of America, "Comments on the New gTLD Applicant Guidebook on Behalf of the U.S. Chamber of Commerce," Dec. 15, 2008, accessed Nov. 11, 2013, http://forum.icann.org/lists/gtld-guide/pdfmcXyIEWj44.pdf, 1.
[11] Dennis Carlton, "Preliminary Analysis of Dennis Carlton Regarding Price Caps for New gTLD Internet Registries," white paper, Compass Lexecon, Mar. 2009, accessed Nov. 11, 2013, http://archive.icann.org/en/topics/new-gtlds/prelim-report-registry-price-caps-04mar09-en.pdf.
[12] Dennis Carlton, "Preliminary Report of Dennis Carlton Regarding Impact of New gTLDs on Consumer Welfare."

large-scale by the sample) are less price-sensitive than competitive ones.[13] Nevertheless, the submissions of Yahoo!, Mack, and several others are also deeply critical of Carlton's papers.[14]

After the Carlton papers were criticized for lack of empirical content, ICANN effectively acknowledged this as an outstanding gap by commissioning two further papers.[15] Nevertheless, skeptical market participants remained unhappy with the scope and quality of the analysis. The work has not been formally peer-reviewed and unfavorable comments have also come from ICANN's own Government Advisory Committee, the International Trademark Association, and members of the U.S. House of Representatives.[16]

Although there is little peer-reviewed empirical literature analyzing gTLD registrations, some work has been done by market participants. These take the common approach of studying a sample of domain names to assess different aspects of use. These studies offer variable coverage of ccTLDs and different approaches to sampling.

Stahura, looking to refute the argument that new gTLDs would be created to maximize takings from brand protection registrations, examined seven major gTLDs and only at the most congested part of the namespace.[17] Summit Strategy looked at scarcity and the extent to which the domains are "used" or simply redirect/don't resolve.[18] Katz et. al. explored scarcity in existing namespaces and the extent of defensive registrations, through an automated analysis backed up with a survey of registrants.[19] Unfortunately, the choice of domains for study is small and arbitrary, which limits the generalizability of the findings. Krueger and Van Couvering looked specifically at a group of 1043 "brand" domain strings, analyzing duplication across gTLDs, which is interesting for examining the IP rightsholders' situation from their own perspective but not for an overall view of DNS use.[20] A study by Halvorson

[13] Michael Kende, "Assessment of ICANN Preliminary Reports on Competition and Pricing," white paper, Analysys Mason, Apr. 17, 2009, accessed Nov. 11, 2013, http://forum.icann.org/lists/newgtlds-defensive-applications/pdfeFL72Sk5n5.pdf.
[14] J. Scott Evans, "Comments of Yahoo! Inc. to the Preliminary Report Regarding Impact of New gTLDs on Consumer Welfare," Apr. 17, 2009, accessed Nov. 11, 2013, http://forum.icann.org/lists/competition-pricing-prelim/pdfesWzs0oVj6.pdf; Mack.
[15] Michael L. Katz, Gregory L. Rosston, and Theresa Sullivan, "An Economic Framework for the Analysis of the Expansion of Generic Top-Level Domain Names," white paper, ICANN, June 2010, accessed Nov. 11, 2013, http://archive.icann.org/en/topics/new-gtlds/economic-analysis-of-new-gtlds-16jun10-en.pdf; Michael L. Katz, Gregory L. Rosston, and Theresa Sullivan, "Economic Considerations in the Expansion of Generic Top-Level Domain Names," white paper, ICANN, Dec. 2010, accessed Nov. 11, 2013, http://www.icann.org/en/topics/new-gtlds/phase-two-economic-considerations-03dec10-en.pdf.
[16] Berkman Center for Internet and Society, 69.
[17] Paul Stahura, "Analysis of Domain Names Registered across Multiple Existing TLDs and Implications for New gTLDs," CircleID, Feb. 2, 2009, accessed Nov. 11, 2013, http://www.circleid.com/posts/20090202_analysis_domain_names_registered_new_gtlds/.
[18] Summit Strategies International, "Evaluation of the New gTLDs: Policy and Legal Issues," white paper, ICANN, July 10, 2004, accessed Nov. 11, 2013, http://archive.icann.org/en/tlds/new-gtld-eval-31aug04.pdf.
[19] Katz, Rosston, and Sullivan, "An Economic Framework for the Analysis of the Expansion of Generic Top-Level Domain Names."
[20] Fred Krueger and Anthony Van Couvering, "An Analysis of Trademark Registration Data in New gTLDs," Working Paper 2012-2, Minds+Machines, Feb. 14, 2010, accessed Nov. 11, 2013, http://web.archive.org/web/20130314194928/http://www.mindsandmachines.com/wp-content/uploads/Analysis-of-Trademark-Registration-Data-in-New-gTLDs.pdf.

et al. is the strongest of those currently published – it is a careful analysis of the .biz domain with a good random sample and consideration of defensive registrations, scarcity, and parked domains.[21] It concludes that in many ways .biz resembles .com.

RESEARCH QUESTIONS AND METHOD

This article quantifies the use and impact of new gTLDs. It asks whether increasing the number of TLDs principally relieves scarcity in names and supports multiple good-faith independent uses of a name, or rather serves to increase defensive registration of domain names by existing holders, thereby increasing deadweight costs.

The following four questions are answered:

- Is there currently scarcity for new entrants to the domain name market?
- Where existing domain names are registered in several TLDs, what proportion are held by the same owner?
- Where domain names are registered in several TLDs by the same owners, what proportion are defensively registered?
- Where existing domain names are registered in several TLDs by different owners, what proportion are cybersquatted?

The data used are collected from a fully random sample of domain names that are registered in at least one gTLD. The process is summarized in Figure 1 below. Zone files, containing the complete list of registered domains, were collected from each gTLD registry on June 10, 2011.[22] The files were then merged to create a complete non-duplicative list of all strings registered in gTLDs on that date. A random sample of 10,000 of those (referred to here as "name strings") were selected.

[21] 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 *Passive and Active Measurement: 13th International Conference*, PAM 2012, ed. Nina Taft and Fabio Ricciato (Berlin: Springer-Verlag, 2012), 221-230.

[22] gTLDs were selected to construct the sample frame for two reasons. Firstly, the zone file data is available to interested parties as a consequence of registries' contracts with ICANN. The same is not true of most ccTLDs, which prohibit zone file access to third parties. Secondly, because this study aims to look at the impact of new gTLDs on the Internet, the universe of names of interest includes those that are registered in gTLDs, not those in the DNS as a whole. For the same reasons, .gov, .mil, .edu and .arpa were not collected. None are available to researchers: .gov and .mil are essentially different in scope to the rest of the gTLDs, being private to the U.S. government, and .arpa is an infrastructure domain for which the zone file would not be meaningful for this study.

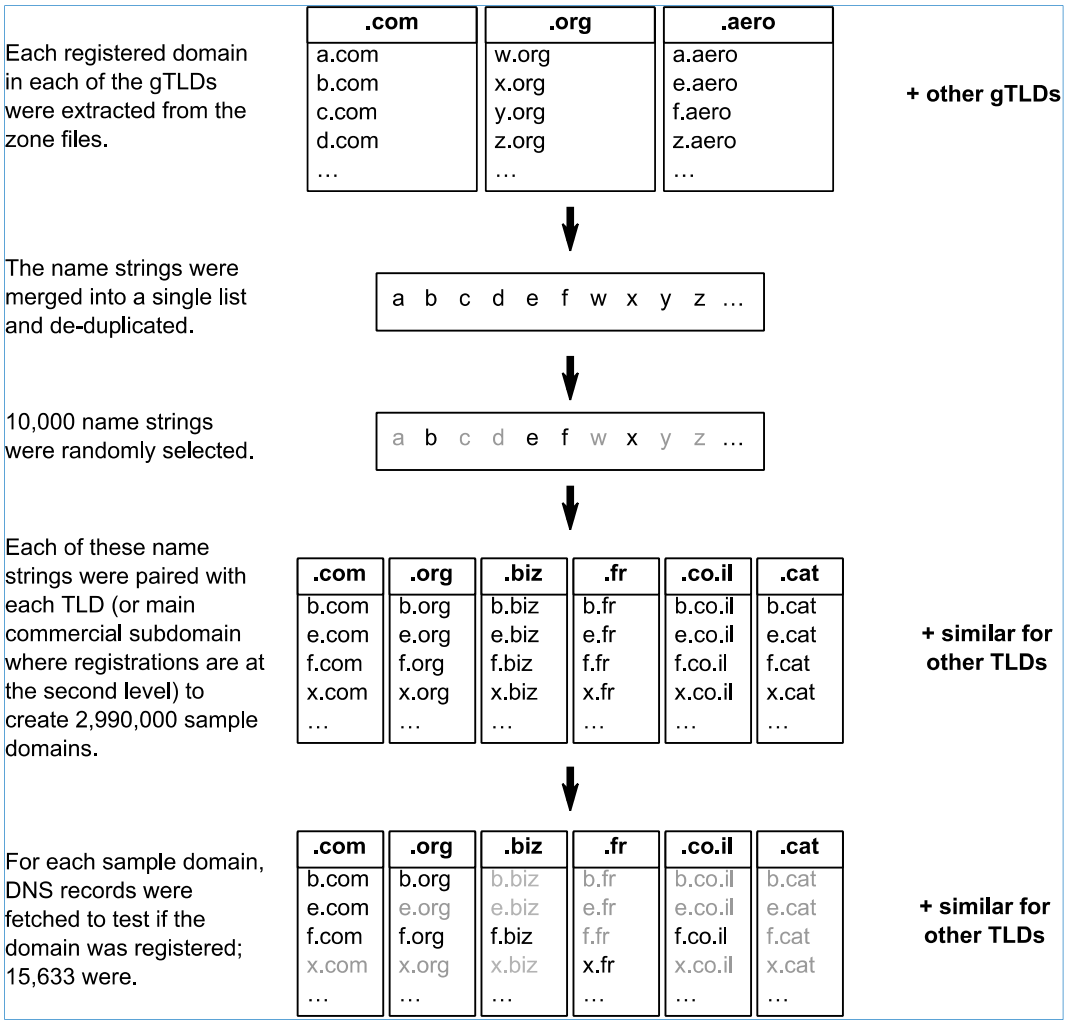

Figure 1: The process for sampling domain names

A list of the gTLDs and all country code TLDs (ccTLDs such as .uk and .ca) was then assembled. The Public Suffix List was used to identify those TLDs in which registration happens at the second level.[23] In these cases, the main commercial subdomain, such as .co.uk, was used instead. TLD policies in this regard were verified individually, where language permitted. In case of doubt, the domain in which any national Google site is registered was taken as a reasonable proxy. Each of these TLDs were then paired in turn with the name strings selected above to generate a final set of string.tld pairs (referred to here as "sample domains") for study. Each sample domain was then categorized by its TLD type. .com is treated as a unique class, as it is the most popular domain and that around which the gTLD expansion debate typically revolves. A distinction is made between the older gTLDs (.net, .org, .edu, and .int) and the newer gTLDs created in the earlier days of ICANN. The latter are the closest existing analogues to the new gTLD expansion proposed by ICANN, and their analysis therefore gives insight

[23] The Public Suffix List can be searched in various ways at the Mozilla Foundation site http://publicsuffix.org/list/.

into the possible effects of future expansion. Finally, ccTLDs are separated from the new internationalized ccTLDs.[24] At the time of the data collection these were still in an experimental stage and therefore had few registrations. See Appendix 1 at the end of this article for a full list of domains and their categories. DNS records were then checked for each of the 2,990,000 possible sample domains. 15,633 of these were registered and had at least one DNS record.[25]

For each .com sample domain with a valid DNS entry, an attempt was made to fetch WHOIS data and parse it for registrant contact information. This was successful in 2,168 cases (21.7%).[26] Each non-.com sample domain was then compared to its corresponding .com to try to identify shared ownership. Domains were recorded as co-owned if 50% or more of the following indicators were found:

- Mail eXchange DNS records pointing to the same mail server.
- DNS records hosted on the same DNS server.
- WHOIS records matching, separately for each among the registrant name, registrant organization, registrant address, registrant city, registrant zip code, registrant phone number, registrant e-mail address, technical contact name, technical contact phone number, and technical contact e-mail address. Fuzzy matches were made using a modified Levenshtein algorithm, to account for variations in formatting and typography: a threshold of 60% matching was established for each item.

The matching accuracy of the algorithm above was measured with an inter-coder reliability trial between the algorithm and the researcher. A random subset of one-sixteenth of the dataset was selected and the algorithm and the researcher independently coded each .com/non-.com pair as matching or non-matching. Krippendorf's alpha was selected as the test statistic with a conservative threshold of 0.85. This is widely used in Internet content classification, is suitable for binary data classification, and corrects for agreement by chance.[27] The results are in Table 1 below. Overall, inter-coder agreement was high: Krippendorf's alpha was 0.916, comfortably above the threshold.

Table 1: Results of automated domain ownership pilot study

		Matched by human coder?		
		false	**true**	**Total**
Matched by algorithm?	**false**	55	2	57
	true	0	14	14
	Total	55	16	71

[24] These are the new internationalized variants of ccTLDs, written in punycoded versions of local scripts. "Punycode" is an instance of a general encoding syntax by which a string of Unicode characters is transformed uniquely and reversibly into a smaller, restricted character set.
[25] The registration status of a further 1,370 was ambiguous, due to server errors when attempting to fetch DNS data.
[26] The major reasons for failure to generate .com ownership data were the use of third-party privacy services such as Domains by Proxy or the inability of the software to parse the reply (which is in an unstandardized free-form text format and varies by registrar).
[27] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken, "Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research Projects," working paper, June 1, 2010, accessed Nov. 13, 2013, http://astro.temple.edu/~lombard/reliability/.

For each sample domain with a valid DNS entry, whether or not WHOIS matching data were available, an attempt was made to fetch the front web page of the domain (both with and without a preceding "www."). This was successful in 13,741 cases. Where there was a redirect to a different domain name, this was also recorded.

Detecting web pages that are "parked" (a holding page, generally either serving ads or offering the domain for sale) is a difficult problem.[28] Here, a random sample of pages was manually classified into parked and non-parked categories by the researcher, using Google Translate for foreign language pages when it was unclear. These data were then used to train a machine-learning maximum entropy classifier using a natural language processing toolkit.[29] It was given no special assistance with non-English language pages. Overall accuracy, estimated using a standard leave-one-out cross-validation trial, was 0.831.[30] (Source code for the software used to collect and process the dataset is available from the author by request.)

FINDINGS

Is there currently scarcity for new entrants to the domain name market?

Scarcity is a tricky concept to test empirically, as different registrants are seeking different names. With 37^{63} domains theoretically available for registration in each TLD, the proportion of domains taken is essentially zero, however many million have been registered. In practice, scarcity claims tend to be made around notions of "short" or "good quality" names. Claims particularly center around .com, which has a much greater number of registered domains than any other TLD.

In theory it would be possible to develop a concept of "good quality" domain names that are available for registration, focusing on combinations of pronounceable words and syllables, possibly with the addition of short acronym-like strings, to reduce the scarcity-space to domains that are in some sense reasonable names for registration. In practice, this is intrinsically subjective (and also depends in large measure on the language in which the registrant expects the domain name to be read). This study takes an alternative approach, which is to focus on the name strings actually registered in at least one of the gTLDs. These strings have been shown to be sufficiently valuable to a registrant and to be worth an annual registration charge. Essentially, this study crowdsources a definition of good-quality names to the market that actually registers them.

---

[28] W-Shadow.com, "Detecting Parked Domains," Nov. 13, 2009, accessed Nov. 13, 2013, http://w-shadow.com/blog/2009/11/13/detecting-parked-domains/.

[29] The toolkit used was Natural Language Toolkit, version 0.9.5 (2009), developed by Edward Loper, Ewan Klein, and Steven Bird. See http://sourceforge.net/projects/nltk/.

[30] Gavin C. Cawley and Nicola L.C. Talbot, "Efficient Leave-One-Out Cross-Validation of Kernel Fisher Discriminant Classifiers," *Pattern Recognition* 36, no. 11 (2003): 2585-2592. Because the input is in multiple languages, this is a reasonable though not spectacular accuracy level. Restricting to English-language pages, or a larger training sample, would improve the classifier's accuracy. Improvements could also be made by training additional classifier features such as word bigraphs or stemming words in the data.

Even though all the sample domains have name strings that are registered in at least one domain, only 0.523% [95% CI: 0.515%-0.531%] of the sample domains were registered in the DNS overall (n = 2988630). For full details see Table 2 below.

Table 2: Proportion of domains registered, by type of TLD, excluding domains with ambiguous registration indications

	Type of TLD											
	.com		old gTLD		new gTLD		ccTLD		IDN (cc)TLD		Total	
DNS exists?	N	%	N	%	N	%	N	%	N	%	N	%
Not registered	704	7.5	37744	94.9	128923	99.3	2385824	99.8	419802	100.0	2972997	99.5
Registered	8637	92.5	2034	5.1	916	0.7	3858	0.2	188	0.0	15633	0.5
Total	9341	100.0	39778	100.0	129839	100.0	2389682	100.0	419990	100.0	2988630	100.0

This is particularly stark when graphed. Figure 2 below shows registration density broken down by the different categories of TLD.

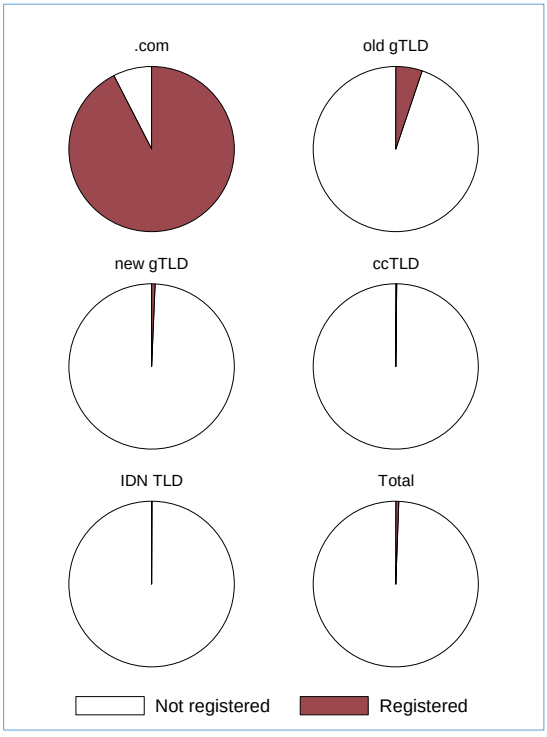

Figure 2: domains registered and unregistered, overall and by type of TLD

A related issue is that although domain names are not scarce, short names are often seen as more desirable and the scarcity of domains can also vary with the length of the name string, especially in .com. Many short, pronounceable domains have indeed been taken, but it is far from clear that length is a sufficiently important criterion for registrants overall for this to constitute domain scarcity.

Constructing a simple linear regression model using the length of the strings (n = 10000, mean = 13.5, sd = 5.39, range: 3-47) and the proportion of domain names registered with a given string (n = 10000, mean = 0.005, sd = 0.0077, range: 0-0.24) shows a very weak negative relationship: each additional character in the string is associated with a 0.0018 reduction in the proportion of domains registered ($p < 0.001$). Although the model and the string length term in it are both statistically significant, the effect size is very small (adjusted R^2 = 0.015). This suggests that domain length is greatly outweighed by other factors in influencing name string registration: the length of available name strings is not the principal issue for registrants and the limited number of short strings available in major TLDs is not enough, in itself, to conclude that the domain name market is characterized by scarcity, given the vast number of longer strings still available for registration.

Where existing domain names are registered in several TLDs, what proportion are held by the same owner?

The proportion of registered non-.com domains with the same owner as the corresponding registered .com was only 0.395 [95% CI: 0.364-0.427], or significantly less than half.[31] Nevertheless, there are important variations by domain. Figure 3 below shows the distribution of sample domains with matched owners in each class of TLD. The old and new gTLDs, with around 60% ownership matching, offer a quite different picture to the gTLDs with only 11%. This suggests that gTLDs may be more prone to defensive registrations, especially as gTLDs are less useful for geographically-differentiated service delivery than ccTLDs.

[31] Excluding domains for which attempting to match was not possible (see footnote 26 above). As with many other confidence intervals for analysis of the work of the ownership matcher and parking classifier, this is the confidence interval taking into account sampling error only – the other methodological quirks in the process mean the true confidence interval will be somewhat wider.

Figure 3: Ownership-matchable domains for which the owner was the same as the .com, by TLD class

Where domain names are registered in several TLDs by the same owners, what proportion are defensively registered?

It is possible to produce a much more informed estimate of defensive registration frequency by looking at domain usage, as well as domain ownership. There is, unsurprisingly, no existing academic model to follow for identifying precisely when a domain is defensively registered, so this article advances a new definition. A value-added domain is defined as one that has a website that is: 1) not parked, and 2) not simply redirected to content in a different TLD.

A defensively-registered domain is held by the same owner as the .com domain that is not value-added according to this definition.[32] This is designed to exclude domains not multiply-registered by the dominant owner and also domains that are owned in multiple, but for which productive use is being made. Amazon, for example, hosts country-specific stores at many of its domain names at amazon.tld. These registrations are not principally defensive.

For the 370 non-.com domains with ownership matching that of the .com domain, the proportion that had no value-added use (and hence were defensively registered) was 0.905 [95% CI: 0.871-0.933]; a large majority of domains multiply-registered by the same owners are defensive registrations. These

[32] Data is also available to see if a DNS MX record is established, allowing the domain to be used to receive mail. However, a configured mail server is not in itself evidence of value-added uses: RFC 2142 strongly encourages hostmaster@domain to be available whatever the use of the domain, so well-configured domains could have an MX record without any real-world usage of the domain for e-mail. See Dave Crocker, "RFC 2142: Mailbox Names for Common Services, Roles and Functions," Network Working Group, May 1997, accessed Nov. 13, 2013, http://tools.ietf.org/html/rfc2142.

defensive registrations are a substantial fraction of all registrations in the sub-sample: for non-.com domains where the ownership is, in principle, matchable, the proportion of defensive registrations is 0.350 [95% CI: 0.319-0.381].

An estimated lower bound of the total annual financial cost of defensive registrations was calculated at $265 million as follows:

$$totalcost = \frac{totalstrings}{sampledstrings} \cdot propmatchable^{-1} \cdot \sum defcost$$

Where:

- *totalstrings* is the total number of unique name strings registered in all gTLDs (102,098,333 at the time of this study);
- *sampledstrings* is the number of unique name strings in this study's sample (10,000);
- *propmatachable* is the proportion of individual domains in the study with a corresponding .com against which ownership can be matched (0.2168);
- *defcost* is the cost of each defensive registration identified in the study.[33]

This is necessarily a rough estimate – it ignores all additional administrative costs beyond the registration fee; it will under-count some defensively-registered domains for which WHOIS data is not available or not accurate; and there are unquantifiable uncertainties around the aggregate accuracy of the content classifier, ownership matcher, and interpretation of redirection. In addition, it will not identify all domains defensively-registered against typosquatting[34] – only that proportion of registrations parked or redirected to a different TLD are counted as defensively registered by this study, while defensive registration against typosquatting can also feature redirects within the same TLD (ammazon.com to amazon.com, for example).

The proportion of defensive registrations is particularly high for the gTLDs, at 0.53 [95%CI: 0.45-0.61]. This is much higher than the ccTLDs (see Figure 4 below). No defensive registrations were identified in IDN TLDs in this study.

[33] For each TLD with identified defensive registrations, an annual registration price was obtained on July 20, 2013 from godaddy.com, a major registrar. Where the TLD was not registrable at godaddy.com (generally ccTLDs) a price was obtained from a local registrar. Annual prices in U.S. dollars are: .biz $15.17, .ca $12.99, .ch $18.06 (17CHF), .co $29.99, .co.il $24.95, .co.uk $9.99, .cz $19.99, .info $10.17, .lv $13.79 (€10.50), .me $19.99, .mobi $18.17, .net $17.17, .org $18.17, .tv $39.99, .us $19.99.

[34] Typosquatting is the practice of registering domains that are common misspellings or homophones of existing domains, to profit from the traffic of confused or careless web users.

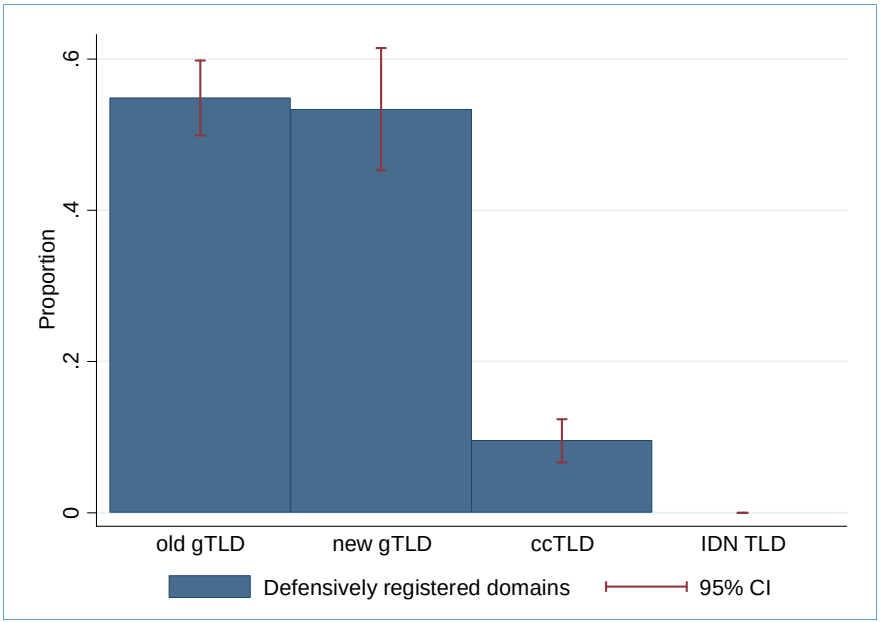

Figure 4: Proportion of defensive registrations, by class of TLD

Breaking down the data, it is possible to directly estimate the cost of defensive registrations in the new gTLDs at $46.2 million per year, the bulk of which ($25.4 million) is from .info and the balance roughly split between .biz and .mobi, with no contribution from other TLDs. The relatively small proportion of the total cost attributable to the new gTLDs is shown most clearly in Figure 5 below.

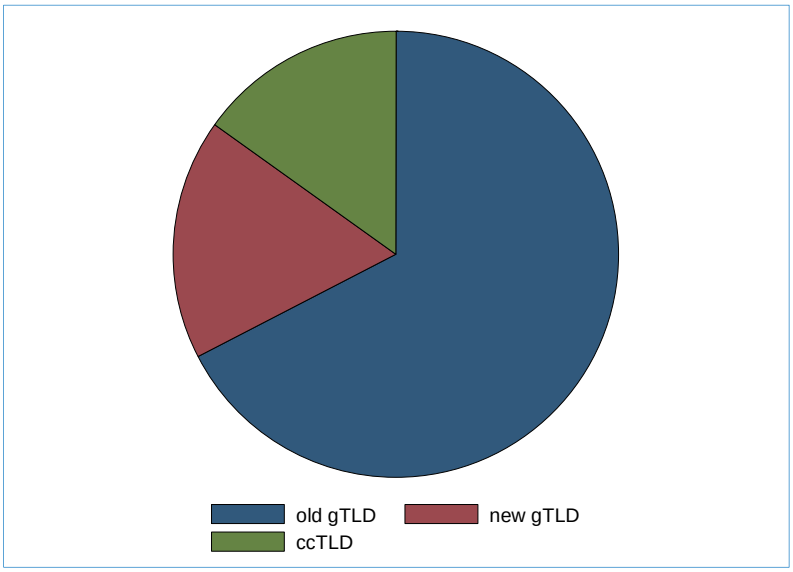

Figure 5: Costs of defensive registration per year, by class of TLD, in U.S. dollars

All the defensive registrations identified in gTLDs in this study are in "open" gTLDs such as .org, .net, .mobi, .biz, and .info. The smaller TLDs with restrictive registration requirements, such as .edu, .cat, and .aero, have too small a number of sample domains actually registered (as few as zero out of the 10,000 in each TLD tested). Although this prevents an accurate assessment of the proportion of defensive registrations in these domains, it does strongly imply that the scale of the issue in absolute terms is negligible compared to the open domains.

Given that names appear not to be defensively registered in all possible domains, the pattern of registration observed here makes intuitive sense. Not all organizations holding domain names are international in scope and not all would want or need registration in every ccTLD, any more than they would register their trademarks in every national trademark office. In the gTLDs, by contrast, the higher rates correspond to those open gTLDs which are in some sense intended to be a substitute or complement to .com (such as .org, .biz, and .info), rather than those with a tight subject focus and restricted registration requirements (such as .aero and .edu).

Where existing domain names are registered in several TLDs by different owners, what proportion are cybersquatted?

The proportion that are cybersquatted[35] can be tested in an analogous way to defensive registration, but by looking at domains that did not have matching owners despite being eligible for matching. For the 543 non-.com sample domains with different ownership to the .com domain, the proportion of domains that had value-added use was only 0.414 [95% CI: 0.373-0.457]. This proportion is roughly equal across new and old gTLDs and ccTLDs.[36] As before, this will not necessarily include all cybersquatters, if the target of the squatting has not registered the corresponding .com.

It would, however, be wrong to assume that all non-value-added domains with different owners were necessarily squatted for the purposes of extracting payment from the holder of the .com. Firstly, many names in .com themselves lack value-added uses: only 0.425 [95% CI: 0.415-0.435] of all .com domains are value-added by this study's definition, with many of the remainder being speculative registrations by "domainers."[37] Secondly, in addition to those piggybacking on other name string users, there are sizable numbers of registrants in the sample domains that are sharing name strings and making legitimate use of them.

[35] Cybersquatting is the practice of registering domains relating to others' trademarks or domain names, either to profit from misdirected web traffic or to attempt to resell the domain to the trademark holder at a profit.
[36] No analysis was possible for IDN TLDs, as there were no common strings between the IDN TLD and the .com dataset.
[37] Domainers are speculators in the domain name market who purchase domains and hold them (often monetizing a domain by hosting advert-displaying parking pages) while hoping they appreciate in value.

DISCUSSION

The Need for New gTLDs: Scarcity, Defensive Registrations, and Deadweight Costs

> "Getting a .com domain is almost impossible nowadays. Odds are someone else has already registered it, because there are close to 100m registered .com domains."[38]

Proponents of new gTLDs argue that names are scarce in .com, and that new domains are therefore needed to provide an alternative venue for registration. Both halves of this argument are difficult to sustain. Only a tiny fraction of the possible number of names in .com are, in fact, registered. Furthermore, this article has shown only a very limited relationship between domain registration rates and name string lengths. Although not all names are equally valuable, many popular domains have names that are not intrinsically meaningful to potential customers but have been successfully branded – major sites such as Amazon, Slashdot, and Google amongst others. That good domain names are scarce is therefore a difficult claim to justify.

Even if the first part of the argument is true, however, it would still be a mistake to assume that the introduction of new gTLDs will relieve scarcity. A central question raised by this study is the extent to which registrations in different TLDs can substitute for each other. After all, over twenty years of access to .net and .org and several years of access to the new gTLDs like .biz and .info have not resulted in those domains filling up to anything like the extent of .com. Even so, there is clearly a continuum of substitutability: the more generally-focused gTLDs (like .org, .biz, and .info) are, in particular, more likely to be defensively registered than the more focused ccTLDs and specialist gTLDs (like .travel and .asia). This suggests either that a popularity contest is at work, and less-prominent domains are not seen as visible enough to be good substitutes for .com, or that it is in similarity to .com as a global general purpose TLD that other domains become seen as reasonable substitutes for it.

In any case, .com is clearly preferred by registrants for some inherent quality – it is a domain apart. As such, it is a distinct market, and if the existing domains are insufficiently attractive to induce registrants to prefer them over a .com registration where the .com is available, it is not clear that newly created gTLDs will fare any better. This uniqueness seems to extend also to defensive registration and cybersquatting:

> "Defensive registrations are a real phenomenon in .com. 100% of the 1043 brands and brand variations are registered in .com. Our earlier study on UDRP filings suggests that this is where the vast majority of cybersquatting also takes place."[39]

Although the judgment as to whether an available name is "good quality" is intrinsically subjective, one benefit of this study's approach to sampling is that it consists of names that have actually been registered (and paid for) by someone who thinks they will be useful. This is a reasonable proxy for

[38] Quote by Diego Calle, in Arturo Wallace, "Colombia Markets its .co Domain as Internet Opens Up," *BBC News*, Aug. 2, 2011, accessed Nov. 13, 2013, http://www.bbc.co.uk/news/world-latin-america-14305361.
[39] Krueger and Van Couvering, 5.

quality, relying on the revealed preferences of market participants rather than the judgment of the researcher. That so many presumptively good names within the sample domains remain available in non-.com TLDs suggests that it is difficult to see the need on scarcity grounds for more names when the existing alternative TLDs are so sparsely filled. If lack of new gTLDs was creating a significant problem for the Internet overall, it is likely that one of the many alternative root proposals would have gained more traction. As the expansion of gTLDs will not in itself relieve any of the scarcity in .com, there may be little consumer surplus in doing so, especially as there is a thriving market in secondary registration of lapsed .com domains by speculators for click-through traffic. If a shorter or easier-to-pronounce name in a new gTLD is a good substitute for an existing .com, switching costs[40] and the likelihood of speculative re-registration prevent the release of .com names for reuse.

Although there may be little consumer benefit from scarcity reduction, this does not mean that there would be no social benefits to the introduction of new gTLDs. Many of the recent proposals have been aimed at a limited community of interest, rather than the general Internet population, and have aimed to produce names that are better than existing TLDs for those groups. For example, .coop limits registration to genuine co-operatives, and .cat to those using the Catalan language. These communities of interest may well be better served by such specialist provision, and the empirical research in this study has demonstrated that the aggregate number of non-value-added registrations is very low in such domains. In many ways, this suggests that ICANN should have perhaps focused on expansion in these sponsored TLDs (sTLDs) instead of large gTLDs.

This study has roughly quantified the extent of defensive registrations, both in proportional and absolute terms. The idea that domain owners religiously register their brands across all domains is thoroughly disproved – many ccTLDs and IDN TLDs have zero registrations within the sample domains, and some small gTLDs have very few. If any brands are registered globally, they are a sufficiently small proportion of the global domain name market to be invisible in this sample, and should not carry disproportionate weight in ICANN decision-making. Nevertheless, cross-ownership matches, coupled with rates of redirects and parking for such domains, demonstrate that cost of defensive registration is generally high in absolute terms: $265 million and potentially one-third of registered domains. This is not a trivial amount of money and, although it is unlikely that the creation of new gTLDs will proportionally increase this sum, the deadweight costs of these registrations should be quantified and taken into account.

Implications for Domain Name Policy

This study has demonstrated the benefits of a more data-driven approach. There is a clear distinction here between general-purpose gTLDs, which offer the widest scope for use but if widely used also come at significant potential cost in defensive registrations; and smaller scale sTLDs, which potentially offer larger benefits to a smaller audience with lower externalities. If the number of new domains is not to be limited by ICANN, this suggests that the maximum consumer surplus could come from a

[40] Milton L. Mueller, Yuri Park, Jongsu Lee, and Tai-Yoo Kim, "Digital Identity: How Users Value the Attributes of Online Identifiers," *Information Economics and Policy* 18, no. 4 (Nov. 2006): 405-422.

large number of specialist domains rather than promoting a series of general-purpose domains like .web.

Secondly, cybersquatting and defensive registration, though not as prevalent as sometimes claimed by trademark holders, is a non-trivial issue and has sizable costs attached. Future domain name policy should quantify these costs in more detail and work out how they can best be mitigated without closing down the whole domain space to future expansion.

Ultimately, this debate raises questions about what the DNS is for. Is it simply a technical database mapping strings to IP addresses? If so, arbitrary restrictions on the number of TLDs make no sense. Alternatively, are domain names intangible property imbued with meaning, for which decisions should be taken on the basis of private economic advantage? Or is part of the difficulty that domain names do not fit neatly into pre-existing categories? This question lies at the heart of many of the debates around TLD policy, but no answer is shared between the different parties.

Further Research Possibilities

Analysis of individual ccTLDs, rather than ccTLDs as a group, has been restricted in this study, as has detailed work on small-scale gTLDs and sTLDs. Despite the large volume of data collected, registration rates in some domains are so low that no registered domains at all were found in the sample. A study focused on ccTLDs and sTLDs would not only need to be much larger in scale, but would need to carefully consider an appropriate sampling frame to gather domain strings registered only in ccTLDs, given that ccTLD zone files are not routinely available.

In principle, a longitudinal study could also be constructed, either retrospectively using historical WHOIS information or prospectively by starting a data collection program, to look at the changing impact of new TLDs over time. Analysis of the behavior of the DNS root has, until now, been largely static, while the domain name market itself is notably dynamic.

A comparative analysis of domain data would also be useful, separating trademarks and non-trademarks as separate categories. Although in principle, registered trademark data (at least) is in the public domain in the United States, such an approach is currently difficult. The interface provided to the U.S. Patent and Trademark Office trademarks database is session-based and difficult to automatically script.[41] Although the raw data is also available,[42] it is not in searchable form and it would be a major undertaking to make it ready for automated classification. It would nevertheless be helpful in the future to process this data and address head-on some of the claims made about the role of trademarks in the domain name market. Validating the results of Krueger and Van Couvering's informal study of trademark data,[43] in particular, would be valuable further research.

[41] Data can be searched at United States Patent Office, Trademark Electronic Search System (TESS), http://tess2.uspto.gov/.
[42] Data can be searched via Google at USPTO Bulk Downloads: Trademarks, http://www.google.com/googlebooks/uspto-trademarks.html.
[43] Krueger and Van Couvering.

Finally, estimates of defensive registrations and cybersquatting could be made more accurate by future research taking account of typosquatting and defensive registrations against it. In principle, a similar approach could be taken, with the additional step of homophone- and typo-matching domains in the sample against the original zone file lists, to identify cases where similar-sounding or similar-spelling domains have been registered by the same or different registrants.

CONCLUSION

It is not accurate to claim that the consequence of new gTLDs will be either an increase in costs and abuses, or an increase in choice and reduction in scarcity. It is very likely, in fact, to be both. In place of ICANN's stakeholders' implicit assumptions about the underlying state of the domain name market, it would be a step forward to develop a more multifaceted model of domain behavior that tries to quantify rather than assert the impacts. To be sure, this study does not by itself do this. It is a small-scale piece of work with substantial uncertainties. But it has made some progress in establishing a richer and deeper understanding of domain name registrations and suggests that a more data-supported approach would be beneficial on the next occasion when ICANN pursues controversial policy changes.

APPENDIX 1: TLDS AS CLASSIFIED FOR THIS STUDY

Domain Classifications				
.com	**old gTLDs**	**new gTLDs**	**ccTLDs**	**IDN (cc)TLDs**
.com	.net .org .edu .int	.asia .biz .info .jobs .mobi .name .tel .travel .cat .coop .aero .museum .pro	.ac .ad .ae .af .ag .ai .al .am .an .co.ao .aq .com.ar .as .at .com.au .aw .ax .az .ba .bb .com.bd .be .bf .bg .com.bh .bi .bj .bm .com.bn .bo .com.br .bs .bt .co.bw .by .bz .ca .cc .cd .cf .cg .ch .ci .co.ck .cl .cm .cn .co .cr .cu .cv .cx .com.cy .cz .de .dj .dk .dm .do .dz .ec .ee .com.eg .com.er .es .com.et .eu .fi .com.fj .co.fk .fm .fo .fr .ga .gd .ge .gf .gg .com.gh .gi .gl .gm .com.gn .gp .gq .gr .gs .com.gt .com.gu .gw .gy .hk .hm .hn .hr .ht .hu .co.id .ie .co.il .im .in .io .iq .ir .is .it .je .com.jm .jo .jp .co.ke .kg .com.kh .ki .km .kn .com.kp .co.kr .com.kw .ky .kz .la .com.lb .lc .li .lk .com.lr .co.ls .lt .lu .lv .ly .ma .mc .md .me .mg .mh .mk .ml .com.mm .mn .mo .mp .mq .mr .ms .com.mt .mu .mv .mw .com.mx .my .co.mz .na .nc .ne .nf .com.ng .com.ni .nl .no .com.np .nr .nu .co.nz .com.om .com.pa .com.pe .pf .com.pg .ph .pk .pl .pn .pr .ps .pt .pw .com.py .com.qa .re .ro .rs .ru .rw .com.sa .com.sb .sc .sd .se .sg .sh .si .sk .sl .sm .sn .so .sr .st .su .com.sv .sy .co.sz .tc .td .tf .tg .co.th .tj .tk .tl .tm .tn .to .com.tr .tt .tv .tw .co.tz .ua .ug .co.uk .us .com.uy .uz .va .vc .com.ve .vg .vi .vn .vu .ws .co.za .co.zm .co.zw	.xn--3e0b707e .xn--45brj9c .xn--54b7fta0cc .xn--90a3ac .xn-- clchc0ea0b2g2a9gcd .xn--fiqs8s .xn--fiqz9s .xn--fpcrj9c3d .xn--fzc2c9e2c .xn--gecrj9c .xn--h2brj9c .xn--j1amh .xn--j6w193g .xn--kprw13d .xn--kpry57d .xn--lgbbat1ad8j .xn--mgb2ddes .xn--mgb9awbf .xn--mgba3a4f16a .xn--mgba3a4fra .xn--mgbaam7a8h .xn--mgbayh7gpa .xn--mgbbh1a71e .xn--mgbc0a9azcg .xn-- mgberp4a5d4a87g .xn--mgberp4a5d4ar .xn--mgbqly7c0a67fbc .xn--mgbqly7cvafr .xn--mgbtf8fl .xn--nnx388a .xn--node .xn--o3cw4h .xn--ogbpf8fl .xn--p1ai .xn--pgbs0dh .xn--s9brj9c .xn--wgbh1c .xn--wgbl6a .xn--xkc2al3hye2a .xn--xkc2dl3a5ee0h .xn--yfro4i67o .xn--ygbi2ammx

BIBLIOGRAPHY

Berkman Center for Internet and Society. "Accountability and Transparency at ICANN: An Independent Review." White paper, Oct. 20, 2010. Accessed Nov. 11, 2013, http://www.icann.org/en/about/aoc-review/atrt/review-berkman-final-report-20oct10-en.pdf.

Carlton, Dennis. "Preliminary Analysis of Dennis Carlton Regarding Price Caps for New gTLD Internet Registries." White paper, Compass Lexecon, Mar. 2009. Accessed Nov. 11, 2013, http://archive.icann.org/en/topics/new-gtlds/prelim-report-registry-price-caps-04mar09-en.pdf.

———. "Preliminary Report of Dennis Carlton Regarding Impact of New gTLDs on Consumer Welfare." White paper, Compass Lexecon, Mar. 2009. Accessed Nov. 11, 2013, http://archive.icann.org/en/topics/new-gtlds/prelim-report-consumer-welfare-04mar09-en.pdf.

Cawley, Gavin C. and Nicola L.C. Talbot. "Efficient Leave-One-Out Cross-Validation of Kernel Fisher Discriminant Classifiers." *Pattern Recognition* 36, no. 11 (2003): 2585-2592.

Chamber of Commerce of the United States of America. "Comments on the New gTLD Applicant Guidebook on Behalf of the U.S. Chamber of Commerce," Dec. 15, 2008. Accessed Nov. 11, 2013, http://forum.icann.org/lists/gtld-guide/pdfmcXyIEWj44.pdf.

Crocker, Dave. "RFC 2142: Mailbox Names for Common Services, Roles and Functions." Network Working Group, May 1997. Accessed Nov. 13, 2013, http://tools.ietf.org/html/rfc2142.

Evans, J. Scott. "Comments of Yahoo! Inc. to the Preliminary Report Regarding Impact of New gTLDs on Consumer Welfare," Apr. 17, 2009. Accessed Nov. 11, 2013, http://forum.icann.org/lists/competition-pricing-prelim/pdfesWzs0oVj6.pdf.

Froomkin, A. Michael and Mark A. Lemley. "ICANN and Antitrust." *University of Illinois Law Review* 2003 (2003): 1-76.

Halvorson, Tristan, Janos Szurdi, Gregor Maier, Mark Felegyhazi, Christian Kreibich, Nicholas Weaver, Kirill Levchenko, and Vern Paxson. "The BIZ Top-Level Domain: Ten Years Later." In *Passive and Active Measurement: 13th International Conference, PAM 2012*, edited by Nina Taft and Fabio Ricciato, 221-230. Berlin: Springer-Verlag, 2012.

Huston, Geoff. "Opinion: ICANN, the ITU, WSIS, and Internet Governance." *The Internet Protocol Journal* 8, no. 1 (2005): 15-28.

Katz, Michael L., Gregory L. Rosston, and Theresa Sullivan. "An Economic Framework for the Analysis of the Expansion of Generic Top-Level Domain Names." White paper, ICANN, June 2010. Accessed Nov. 11, 2013, http://archive.icann.org/en/topics/new-gtlds/economic-analysis-of-new-gtlds-16jun10-en.pdf.

———. "Economic Considerations in the Expansion of Generic Top-Level Domain Names." White paper, ICANN, Dec. 2010. Accessed Nov. 11, 2013, http://www.icann.org/en/topics/new-gtlds/phase-two-economic-considerations-03dec10-en.pdf.

Kende, Michael. "Assessment of ICANN Preliminary Reports on Competition and Pricing." White paper, Analysys Mason, Apr. 17, 2009. Accessed Nov. 11, 2013, http://forum.icann.org/lists/newgtlds-defensive-applications/pdfeFL72Sk5n5.pdf.

Krueger, Fred and Anthony Van Couvering. "An Analysis of Trademark Registration Data in New gTLDs." Working Paper 2012-2, Minds+Machines, Feb. 14, 2010. Accessed Nov. 11, 2013, http://web.archive.org/web/20130314194928/http://www.mindsandmachines.com/wp-content/uploads/Analysis-of-Trademark-Registration-Data-in-New-gTLDs.pdf.

Lombard, Matthew, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. "Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research Projects." Working paper, June 1, 2010. Accessed Nov. 13, 2013, http://astro.temple.edu/~lombard/reliability/.

Mack, Andrew. "Andrew Mack Comments on Dennis Carlton Report Regarding Impact of New gTLDs on Consumer Welfare," Apr. 17, 2009. Accessed Nov. 11, 2013, http://forum.icann.org/lists/competition-pricing-prelim/msg00019.html.

Mueller, Milton L. *Ruling the Root: Internet Governance and the Taming of Cyberspace*. Cambridge, MA: Massachusetts Institute of Technology Press, 2002.

Mueller, Milton L., Yuri Park, Jongsu Lee, and Tai-Yoo Kim. "Digital Identity: How Users Value the Attributes of Online Identifiers." *Information Economics and Policy* 18, no. 4 (Nov. 2006): 405-422.

Murray, Andrew D. *The Regulation of Cyberspace: Control in the Online Environment*. Abingdon, U.K.: Routledge-Cavendish, 2007.

Stahura, Paul. "Analysis of Domain Names Registered across Multiple Existing TLDs and Implications for New gTLDs." CircleID, Feb. 2, 2009. Accessed Nov. 11, 2013, http://www.circleid.com/posts/20090202_analysis_domain_names_registered_new_gtlds/.

Summit Strategies International. "Evaluation of the New gTLDs: Policy and Legal Issues." White paper, ICANN, July 10, 2004. Accessed Nov. 11, 2013, http://archive.icann.org/en/tlds/new-gtld-eval-31aug04.pdf.

Wallace, Arturo. "Colombia Markets its .co Domain as Internet Opens Up." *BBC News*, Aug. 2, 2011. Accessed Nov. 13, 2013, http://www.bbc.co.uk/news/world-latin-america-14305361.

World Intellectual Property Organization. "The Management of Internet Names and Addresses: Intellectual Property Issues." Final Report of the WIPO Internet Domain Name Process, Apr. 30, 1999. Accessed Nov. 11, 2013, http://www.wipo.int/amc/en/processes/process1/report/finalreport.html.

W-Shadow.com. "Detecting Parked Domains," Nov. 13, 2009. Accessed Nov. 13, 2013, http://w-shadow.com/blog/2009/11/13/detecting-parked-domains/.