UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

An Exploration of the Identifying Characteristics of Spam Campaign Address Lists

Permalink

https://escholarship.org/uc/item/05z017qc

Author

Gardner, Christopher Patrick

Publication Date

2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

An Exploration of the Identifying Characteristics of Spam Campaign Address Lists

A Thesis submitted in partial satisfaction of the requirements for the degree
Master of Science

in

Computer Science

by

Christopher Patrick Gardner

Committee in charge:

Stefan Savage, Chair
Kirill Levchenko
Geoffrey Voelker

2015

Copyright

Christopher Patrick Gardner, 2015

All rights reserved.

The Thesis of Christopher Patrick Gardner is approved and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2015

DEDICATION

To my loving wife Clair, without your seemingly infinite patience, support, and encouragement this never would have been accomplished.

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures vii

List of Tables ix

Acknowledgements x

Abstract of the Thesis xi

Chapter 1. Introduction 1

Chapter 2. Background 3

Chapter 3. Data Sources 7
3.1. Grum 8
3.2. MegaD 8
3.3. Pushdo 9
3.4. Rustock 10
3.5. Srizbi 11
3.6. Storm 11

Chapter 4. Methodology 12
4.1. Terminology 12
4.2. Sorting Methodology 14
4.3. Domain Distribution Analysis 15

Chapter 5. Determining the Address List Sources 16
5.1. Google Search 17
5.2. Invalid Addresses 20
5.3. Duplicated Addresses 21
5.4. Cross-List Shared Addresses 22
5.5. Cross-List Address Contiguity 24
5.6. Sampled Distinct Addresses 26

Chapter 6. Creating a Classifier 28
6.1. Cross-List Domain Distribution 29
6.1.1. Top-Level Domain Distributions 29
6.1.2. Country-Code Domain Distributions 36
6.1.3. Registered Domain Distributions 45
6.2. Sampling Domain Distribution 64
6.2.1. Grum 64
6.2.2. MegaD 66
6.2.3. Pushdo 72
6.2.4. Rustock 78
6.2.5. Srizbi 84
6.2.6. Storm 90
6.2.7. Storm (C&C) 97

Chapter 7. Conclusion 103

Appendix A. Domain List 108
A.1. Top-Level Domain 108
A.2. Country-Code Domain 109

References 121

LIST OF FIGURES

Figure 4.1: Basic Address Example 13
Figure 4.2: Registered Sub-Domain Address Example 13
Figure 4.3: Country-Code Address Example 13

Figure 5.1: Google Search Results 19
Figure 5.2: Consecutive Address Matches 25

Figure 6.1: Top-Level Domain Distributions 30
Figure 6.2: Top-Level Domain Difference vs. MegaD 31
Figure 6.3: Top-Level Domain Difference vs. Pushdo 32
Figure 6.4: Top-Level Domain Difference vs. Rustock 33
Figure 6.5: Top-Level Domain Difference vs. Srizbi 34
Figure 6.6: Top-Level Domain Difference vs. Storm 34
Figure 6.7: Top-Level Domain Difference vs. Storm (C&C) 35
Figure 6.8: Country-Code Domain Distributions 37
Figure 6.9: Curated Country-Code Domain Distributions 39
Figure 6.10: Curated Country-Code Domain Distribution Difference vs. MegaD . 40
Figure 6.11: Adjusted Curated Country-Code Domain Distribution Difference vs. Pushdo 41
Figure 6.12: Curated Country-Code Domain Distribution Difference vs. Rustock 42
Figure 6.13: Curated Country-Code Domain Distribution Difference vs. Srizbi . . 43
Figure 6.14: Curated Country-Code Domain Distribution Difference vs. Storm . 44
Figure 6.15: Curated Country-Code Domain Distribution Difference vs. Storm (C&C) 45
Figure 6.16: Curated Registered Domain Distributions 47
Figure 6.17: Adjusted Curated Registered Domain Distributions 48
Figure 6.18: Largest Registered Domain Distributions 49
Figure 6.19: Curated Registered Domain Distribution Difference vs. MegaD . . . 50
Figure 6.20: Largest Registered Domain Distribution Difference vs. MegaD . . . 51
Figure 6.21: Curated Registered Domain Distribution Difference vs. Pushdo . . . 52
Figure 6.22: Largest Registered Domain Distribution Difference vs. Pushdo . . . 53
Figure 6.23: Adjusted Registered Domain Distribution Difference vs. Rustock . . 54
Figure 6.24: Largest Registered Domain Distribution Difference vs. Rustock . . . 55
Figure 6.25: Adjusted Curated Registered Domain Distribution Difference vs. Srizbi 56
Figure 6.26: Largest Registered Domain Distribution Difference vs. Srizbi 57
Figure 6.27: Curated Registered Domain Distribution Difference vs. Storm 58
Figure 6.28: Adjusted Curated Registered Domain Distribution Difference vs. Storm 59
Figure 6.29: Largest Registered Domain Distribution Difference vs. Storm 60
Figure 6.30: Curated Registered Domain Distribution Difference vs. Storm (C&C) 61

Figure 6.31: Adjusted Curated Registered Domain Distribution Difference vs. Storm (C&C) 62
Figure 6.32: Largest Registered Domain Distribution Difference vs. Storm (C&C) 63
Figure 6.33: Grum Registered Domain Sampling Analysis 65
Figure 6.34: Grum Registered Domain Sampling Difference 66
Figure 6.35: MegaD Top-Level Domain Sampling Analysis 67
Figure 6.36: MegaD Top-Level Domain Sampling Difference 68
Figure 6.37: MegaD Curated Country-Code Domain Sampling Analysis 69
Figure 6.38: MegaD Curated Country-Code Domain Sampling Difference 70
Figure 6.39: MegaD Curated Registered Domain Sampling Analysis 71
Figure 6.40: MegaD Curated Registered Domain Sampling Difference 72
Figure 6.41: Pushdo Top-Level Domain Sampling Analysis 73
Figure 6.42: Pushdo Top-Level Domain Sampling Difference 74
Figure 6.43: Pushdo Curated Country-Code Domain Sampling Analysis 75
Figure 6.44: Pushdo Curated Country-Code Domain Sampling Difference 76
Figure 6.45: Pushdo Curated Registered Domain Sampling Analysis 77
Figure 6.46: Pushdo Curated Registered Domain Sampling Difference 78
Figure 6.47: Rustock Top-Level Domain Sampling Analysis 79
Figure 6.48: Rustock Top-Level Domain Sampling Difference 80
Figure 6.49: Rustock Curated Country-Code Domain Sampling Analysis 81
Figure 6.50: Rustock Curated Country-Code Domain Sampling Difference 82
Figure 6.51: Rustock Curated Registered Domain Sampling Analysis 83
Figure 6.52: Rustock Curated Registered Domain Sampling Difference 84
Figure 6.53: Srizbi Top-Level Domain Sampling Analysis 85
Figure 6.54: Srizbi Top-Level Domain Sampling Difference 86
Figure 6.55: Srizbi Curated Country-Code Domain Sampling Analysis 87
Figure 6.56: Srizbi Curated Country-Code Domain Sampling Difference 88
Figure 6.57: Srizbi Curated Registered Domain Sampling Analysis 89
Figure 6.58: Srizbi Curated Registered Domain Sampling Difference 90
Figure 6.59: Storm Top-Level Domain Sampling Analysis 91
Figure 6.60: Storm Top-Level Domain Sampling Difference 92
Figure 6.61: Storm Curated Country-Code Domain Sampling Analysis 93
Figure 6.62: Storm Curated Country-Code Domain Sampling Difference 94
Figure 6.63: Storm Curated Registered Domain Sampling Analysis 95
Figure 6.64: Storm Curated Registered Domain Sampling Difference 96
Figure 6.65: Storm (C&C) Top-Level Domain Sampling Analysis 97
Figure 6.66: Storm (C&C) Top-Level Domain Sampling Difference 98
Figure 6.67: Storm (C&C) Curated Country-Code Domain Sampling Analysis . . 99
Figure 6.68: Storm (C&C) Curated Country-Code Domain Sampling Difference 100
Figure 6.69: Storm (C&C) Curated Registered Domain Sampling Analysis 101
Figure 6.70: Storm (C&C) Curated Registered Domain Sampling Difference . . . 102

LIST OF TABLES

Table 5.1: Invalid Addresses 21
Table 5.2: Duplicated Addresses 22
Table 5.3: Cross-List Shared Address Counts 22
Table 5.4: Cross-List Shared Address Percentages 23
Table 5.5: Distinct Address Results 26

Table A.1: The Country-Code domains and their authorized sub-domains. 110

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Stefan Savage for his support as the chair of my committee. His guidance and insights through multiple drafts over several long years has been invaluable.

I would also like to acknowledge Kirill Levchenko who helped to frame and review many of the analytical techniques used in this paper in addition to serving as a member of my committee.

Finally, I would like to acknowledge the time and effort provided by Professor Geoffrey Voelker who also served as a member of my committee.

ABSTRACT OF THE THESIS

An Exploration of the Identifying Characteristics of Spam Campaign Address Lists

by

Christopher Patrick Gardner
Master of Science in Computer Science
University of California, San Diego, 2015
Stefan Savage, Chair

In this paper, email addresses targeted by several botnets including Grum, MegaD, Pushdo, Rustock, Srizbi, and Storm are analyzed for two goals. These addresses are organized into various lists which were gathered from these botnets. The first goal of this analysis is to determine how each botnet collected the addresses they send spam to solely from the addresses in the lists. This is performed using Google searches, by reviewing the duplicated and invalid addresses within each list, and by examining the addresses shared between lists.

The second goal is to determine if a classifier can be created from the domain distributions of the addresses in these lists. This classifier must be able to correctly identify the source botnet from a set of targeted addresses and must correctly distinguish between botnets. The top-level (TLD), country-code (ccTLD), and registered domain distributions will be used in this analysis.

Chapter 1

Introduction

Botnets are the most prevalent source of spam on the Internet today [1]. With email providing a vital and dominant method of communication in both the personal and business worlds, these spam messages have become increasingly disruptive. Today's spam can generally be divided into three categories: malicious payload delivery, phishing, or marketing. Malicious payloads generally take the form of binaries which takeover the target's computer and add it to the botnet or install a virus or other malware. Phishing messages work to convince the addressee to provide personal information such as financial data, login credentials, or other identifying private values. Lastly, marketing spam attempts to sell merchandise, often pharmaceuticals and designer goods, to its recipients via the unsolicited messages.

Even as spam detection and filtering software employed by Internet Service Providers (ISPs) and email operators improves and becomes more intelligent, spam campaigns continue to be successful. The economics of generating and distributing spam enable this success. Generating a variety of diverse messages using templates and randomization software which increases the difficulty of identifying messages from the same campaign are techniques used by several spammers [2]. Additionally, the cost of sending a large number of messages is negligible, especially when the resources used to transmit those messages are illegally acquired as they are in a botnet. This enables

botnets to continuously send millions of spam messages for very little cost and requires only negligible success rates to remain viable.

The low cost makes it acceptable to target addresses where the confidence of an active user being associated with an account is low. Unlike traditional mail-based marketing, addresses on the Internet are not public information; therefore, spammers must create their own lists of them to target. These lists are often built using values scraped from the hard drives and databases of infiltrated computers and networks, through web crawling and search engine based techniques, and using random generation methods, often utilizing dictionary techniques, to create possibly viable addresses. Sometimes transformations may be used to produce additional addresses from one known good address by appending other domains to its host.

Botnets may determine which addresses to prioritize for future campaigns by analyzing the server responses to their messages. Using error responses provided by some domains, such as the *address not found* message, botnets can weed out addresses which do not exist. Additionally, if a user clicks on a link from a spam email, or replies to the message, a botnet may mark that address as belonging to an active user by using click trajectories tied uniquely to the link in that particular user's message.

This paper analyzes the address lists botnets use to target their spam campaigns. Factors indicating how the addresses in the list were collected may be identified by examining the proportion of invalid, duplicated, or search-able addresses within a list. By comparing the contents of lists from several botnets, it should become clearer whether the botnets share or coordinate any portion of their lists. Also, by judging the similarities in the distribution of the address domains from the lists of different botnets and from the same botnet by using sampling, the ability to use these domain distributions to identify the source of a spam campaign from a subset of its targeted addresses will be determined.

Chapter 2

Background

There are several examples of investigations into the origins of the email addresses targeted in spam campaigns. These studies have generally used addresses seeded to the Internet, honeypots, and botnet infiltration to determine how addresses are gathered. Unlike this analysis which relies on information gathered solely from the botnets, each of the studies below used known values controlled by the researchers. Controlling the addresses analyzed and their distribution on the Internet, give clear insight into how those addresses were gathered by and used by botnets and other spammers. Studies relying on Internet-seeded addresses and honeypots provide insight into how web crawling techniques are used by spammers to find targets. A selection of these studies and their findings are described below. In this analysis of the botnet mailing lists, Google searches for addresses from each list will attempt to correlate these findings.

The United States Federal Trade Commission (FTC) has sponsored various studies on how email addresses are targeted by spammers using Internet seeding techniques. One study, done in 2002, seeded 250 unique addresses to 175 Internet locations including web pages, news groups, chat rooms, message boards, and online directories [3]. They found that 86% of the addresses seeded to a web site or newsgroup received spam and automated address harvesting software actively monitors chat rooms for addresses. Addresses posted to free personal web pages, message boards, and directories were

found to receive considerably less spam by the FTC, while addresses seeded to instant messaging user profiles, WHOIS domain registries, and online services for employment networking and dating received no spam during the study. This study did not use a set of control addresses to compare against their seeded results.

The Center for Democracy and Technology spent six months monitoring over 250 unique email addresses created for their study [4]. These addresses were seeded to web pages, USENET groups, and WHOIS database registries. Over 97% of the spam received during the study was sent to addresses posted to public web pages. The study also found that addresses on sites with more traffic and addresses that are available on the public Internet for longer periods of time were more likely to receive higher quantities of spam. During the course of the study, the mail servers used by the Center for Democracy and Technology encountered a brute force attack where a spammer attempted to email every possible combination of letters which formed a valid email address, mailing addresses which had never been created for their servers let alone seeded as a part of the study.

Similar studies were also run on two honeypots. The University of Regensburg in Germany used newly created addresses in a honeypot along with a control set [5]. The study found 47% of the email received by the honey pot was spam. 69.9% of the spam messages received by the Internet were received by seeded addresses, showing that approximately 30% of the spam received was from random guesses made by spammers. Further, 43% of the addresses seeded on the web by the study received spam.

Another honey pot study, run by Unspam, LLC in conjunction with the University of California Santa Cruz, studied the behavior of email address harvesters [6]. This study used Project Honey Pot, a collection of honey pots installed on machines worldwide and a central server which coordinate these honey pots; collecting data from the honey pots and sending addresses to the honey pots for their use. Links to the honey pots are added to web pages and are formatted to be invisible to human visitors but accessible by spiders and robots. When these links are followed, the honey pot records the IP address, user-agent, and referrer string of the visitor while returning a web page with

an embedded unique email address to the visitor. This unique address is tied to both the visitor information and the time the link was followed. These addresses are only sent from the honey pots once. The study's analysis of the visitors to the web page estimate that at least 5 percent of all automated traffic are spam harvesters and that the spammers put little effort into obscuring the identity of their harvesters. The study also found that a small number of harvesters were responsible for more than 50% of the spam received by the honeypot network.

Another method of analysis used to investigate botnets and their behaviors is infiltration. The Storm botnet was successfully infiltrated and analysis from University of California, San Diego's monitoring is used for several studies [6, 7, 8]. This infiltration was achieved by creating new proxy bot as required by infecting globally reachable hosts controlled by the researchers with the Storm botnet. One study analyzes the efficiency of and message lists employed by Storms spam campaigns [7]. This study uses UCSD's compromised proxy bots to track the spam campaigns run by Storm. The proxies determined that the spam sent by worker bots is only successfully delivered to its target 1/6th of the time. Most of these failures are caused by SMTP and DNS lookup errors. The proxies also recorded that of the 929,976 addresses they received from worker bots in harvest reports, only 463,580 (or approximately 50%) were unique. Further, 10% of the harvested addresses do not have a valid top-level domain, an issue most likely caused by poor pattern matching during the scan process of worker bots for possible email addresses.

Infiltration-based studies can provide greater insights into exactly how a botnet operates. The addresses spammed by the botnet can be gathered from the infiltrated nodes. In this case, a greater overall perspective of the addresses targeted is available because it is not limited to the subset of known addresses seeded by the researchers and also include invalid and duplicated addresses. This analysis will look for badly formatted and duplicated addresses in its lists.

This study attempts to determine what can be understood about a spam campaign in a more real-world situation than the studies described above. That is, where the

investigators of a campaign do not have any *a priori* knowledge, such as control over the addresses targeted, or an insider perspective of the spamming botnet meaning they have not already infiltrated the network.

Chapter 3

Data Sources

Several spam mailing lists collected from various botnets are used in this analysis. The majority of these lists were compiled from the University of Washington's Botlab system [8]. Botlab monitors the incoming and outgoing botnet activity on the University of Washington's network scanning incoming spam links and managing sandboxed bot nodes. Lists of targeted addresses identified by this system for the Grum, MegaD, Pushdo, Rustock, Srizbi, and Storm botnets are analyzed.

In addition to the lists compiled by Botlab, a list of addresses targeted by the Storm botnet is used. This list originates from the command and control (C&C) infiltration of this network by the University of California San Diego [9]. Three types of systems are used in the Storm network: worker bots, proxy bots, and master C&C servers. Worker bots request work and transmit spam based on orders received. Proxy bots transfer these requests and commands to master servers which generate commands for the workers and compile their resulting status reports. When a new bot is infected, the infection determines if the host can be reached externally. If the host can be reached, it will become a proxy; otherwise, it will become a worker. The master servers are directly controlled by Storm's operators and are used to control the network via the proxy bots. UCSD accomplished its infiltration of the Storm botnet's C&C system by purposefully infecting globally reachable systems under their control. To distinguish between

the lists generated in UCSD's analysis and Botlab's Storm address set the former will be referred to as the Storm (C&C) list.

A summary of the botnets that have address subsets used in this analysis is provided below. This provides background context regarding these spammers and may provide better understanding and additional insights in the analysis.

3.1 Grum

Grum was a relatively small botnet but sent a high volume of spam per infected host. The botnet was first detected in February 2008 [10]. By January 2009, the botnet had only around 100,000 infected hosts and was responsible for approximately 0.9% of all spam [11]. Six months later, Grum had infected approximately 600,000 to 900,000 computers and was responsible for around 6.0% of global spam [12].

Grum continued to grow and by September 2009 it had between an estimated 560,000 to 840,000 bots. Even though Grum was half the size of the largest botnets at this time, it sent the most spam at over 300 messages every minute per bot and was responsible for approximately 23% of all spam [13].

In April 2010, Grum grew to approximately 1 million bots and was responsible for just under 24% of spam. The Grum botnet had previously been primarily in Brazil, the United States, Korea, and Vietnam, but the majority of the increase was due to new infections in Russia and India [14]. By August 2010, Grum was still at a similar size, but the amount of spam it sent had decreased. Grum was responsible for approximately 16% of all spam that month and the majority of it came from infected computers in Vietnam, India, and Russia [15].

3.2 MegaD

The Mega-D botnet (also known as Ozdok) rose to prominence after the Srizbi botnet was crippled by the take down of the McColo ISP in November 2008 [11]. After

the fall of Srizbi, Mega-D became the highest spamming botnet in January 2009, sending approximately 26 million messages per minute and approximately 589,000 messages per infected computer each day [11]. These high levels of spam continued through February, when Mega-D was responsible for 40% of all spam and had a higher output than other larger botnets [16].

On November 4th 2009, security researchers led a coordinated attack involving several ISPs against the Mega-D botnet, crippling it almost immediately [17]. However, by November 13th, the botnet resurfaced sending out large volumes of spam. The new Mega-D network was comprised of 95% previously unseen IP addresses. These fresh computers were most likely a collection of contingency sleeper bots held in reserve to recover from such a coordinated assault [17].

By April 2010, Mega-D had regained its stature and had the third highest overall output of all botnets, responsible for approximately 18% of all spam. While the botnet was still much smaller than other botnets, controlling approximately 240,000 infected computers, each of its infected bots had an extremely high output of 430 spam emails every minute [14]. The top infected regions for Mega-D also transitioned from Vietnam, Brazil, and India to Russia, Ukraine, and Kazakhstan [14].

The spam generated by the Mega-D network used short term disposable domains for 92.9% of its emails [18]. Most often several disposable domains would resolve to the same IP address. It has also been determined that Mega-D was one of the most multilingual botnets, with 6.4% of its spam messages using a language other than English. Mega-Ds spam languages included Chinese (0.6%), French (3.7%), and Spanish (2.0%). Additionally, 4.5% of MegaD's spam messages are sent in a non-identifiable language [19].

3.3 Pushdo

The Pushdo botnet was first seen in January 2007 [12] and is also known as Cutwail [18] or Pandex [11]. The botnet spread its infection through spoofed digital

greeting cards which include malicious links to install the Trojan [20]. Pushdo had a strong presence in European, Middle Eastern, African, South American, and Asian Pacific countries [21].

While Pushdo was the largest botnet with approximately more than one million infected bots by January 2009, it sent only 7.7% of all global spam [11]. By the end of May, Pushdo had grown to approximately encompass 1.5 to 2 million infected computers and was responsible for 35% of all spam [12].

3.4 Rustock

Rustock was a botnet that appeared on the security scene in 2007 [22]. The Trojan employed several anti-countermeasure features, including root-kit techniques, to prevent its detection and removal by anti-virus applications [22]. In early 2007, Rustock was one of the main image spam senders, using templates to create unique images on demand [2]. Later in 2007, Rustock began using free picture-hosting sites to provide a link that displayed when clicked or as a HTML encoded image [22].

In 2009 Rustock had grown to become the largest botnet in the world at approximately 1.3 to 1.9 million infected computers in September 2009 [13]. Although Rustock was the largest network during this time, it sent fewer messages per host than some of the smaller botnets and, therefore, was only responsible for approximately 10.0% of worldwide spam [13].

Rustock continued its growth, and in April 2010 it increased the number of bots it had by 300% to approximately 1.6 to 2.4 million bots, making it the largest botnet in the world [14]. This increase greatly changed the geographical distribution of the botnet which had previously had the most infections in Brazil. After the increase, both India and the United States surpassed Brazil as the most infected countries. At the same time, the network decreased the spam output of each bot by 65%; yet with the size of the network, Rustock was still responsible for over 32% of all spam, the highest volume by a single botnet at the time [14].

3.5 Srizbi

The Srizbi botnet was the emailing module of the Reactor Mailer spamware portal providing spamming services to spammers. While the Reactor Mailer software appeared in 2004, the Srizbi Trojan first appeared in mid-2007 [23]. By May 2008, Srizbi was responsible for over 40% of all spam including phishing schemes spoofing several different banks and financial organizations [24]. Srizbi continued to thrive throughout the year. When the Intercage ISP (a.k.a. Atrivo) was shutdown in September, Srizbi's output dropped briefly but quickly recovered [20]. By October, Srizbi was responsible for 50% of all global spam [2]. However, when the McColo Corporation ISP was taken down in November, the output of the botnet immediately dropped by 60% [20]. Srizbi never recovered after McColo was shut down.

3.6 Storm

The Storm Worm botnet (also known as Zhelatin, Peacomm, and Nuwar [25]) was first detected in January 2007 [26]. By the end of 2007 an estimated 1.8 million computers were infected worldwide [22]. By March 2008, Storm was responsible for 23% of all spam worldwide and sent an estimated 2.9 million messages containing links to its malware [27]. Also in 2008, the Storm botnet began partitioning its network into smaller discreet sub-networks which were rented out to other spammers and malware distributors [27].

The Storm botnet proliferated through email advertising virtual post cards, beta software, or YouTube videos. These emails contained a single link to an IP address of an infected machine within the botnet. The infected machine then redirected the victim to a back-end server in an attempt to infect them. No malicious code was contained within the emails and the links, message text, and subject lines were continually changed to prevent detection [25].

Chapter 4

Methodology

4.1 Terminology

For the purposes of this paper, an email address is comprised of four main parts: the user name, registered domain, country-code domain (ccTLD), and top-level domain (TLD). A basic example of this can be seen in Figure 4.1. The user name of an address will be the identifier to the left of the @ symbol and is unique within the domains specified to its right.

The right-most domain lexically is the topmost domain of the address and may either be a TLD or a ccTLD. An address cannot have both a TLD and a ccTLD as its topmost domain. However, the ccTLD may include a sub-domain associated with its registrar. These sub-domains will be to the immediate left of the two character ccTLD. Each country-code's registrar has a set of officially recognized sub-domains including the set of domains recognized as TLDs. The valid sub-domains for each country are listed in the Appendix in section A.

The registered domain is the first domain lexically to the left of the topmost domain (either a top-level or a country-code domain including its associated sub-domains). The registered domain may have sub-domains beneath it which will be to its left lexically.

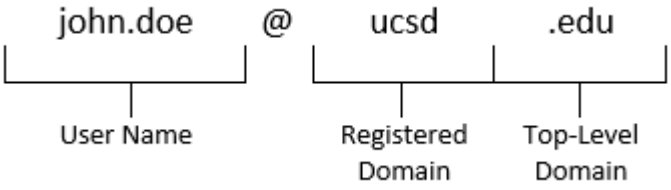


Figure 4.1: A basic email address with simple registered and top-level domains.

The second example in Figure 4.2 is similar to the basic email address from the first example but includes a sub-domain under the registered domain. For many of the experiments performed, the sub-domain will be removed from the address as part of the sorting process described later.

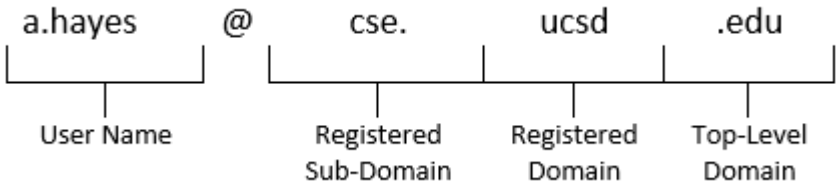


Figure 4.2: Example of an email address with an additional registered sub-domain.

The final example in Figure 4.3 also has a sub-domain, but in this case that sub-domain is part of the country-code domain. For experiments using sorted and verified email addresses, this sub-domain will be validated as a known good domain supported by the particular country. Once confirmed as a valid address during the sorting process, the country sub-domain will be removed from the address.

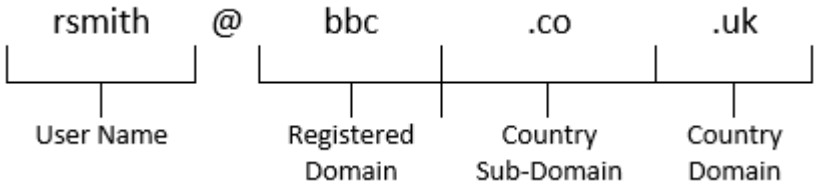


Figure 4.3: Example of an email address with a country-code domain including a sub-domain for that country.

4.2 Sorting Methodology

Several of the analysis algorithms and experiments rely on the address lists to be in a sorted order. By presorting the lists, the efficiency of these tests is improved. The sorting process also removes any invalid or duplicated addresses, eliminating bad values and decreasing the size of the lists for future analysis.

Two different sorting methods are used: the first is a standard sort of the addresses; the second method sorts the addresses by their registered domains. The standard sorting process sorts addresses via their components from right to left. In other words, addresses are sorted by their topmost domain to their bottommost domain, and then finally by the user name of the address. The registered sorting algorithm sorts addresses first by their registered domains followed by their topmost domains, then by any lower domains, and finally by the user names of the addresses. To handle the large sizes of several of the lists, these lists are broken into more manageable sized files which are then sorted together using a merge sort.

These two different sorting methods are used to meet the various resource and efficiency needs of the analysis routines run on the sorted results. The standard sorting is often the most straightforward and easily used. The majority of the analysis routines use this sorting for this reason. However, the registered domain sorting algorithm is necessary to meet the memory requirements of the domain distribution analysis discussed in the next section. With the scale of some of the lists, storing all the registered domains present in the list with a count of the number of entries for each domain can become prohibitive with regards to space. By using a list sorted primarily by the registered domain, when the next registered domain is encountered sequentially, the previous domain's count can be written to disk and discarded from memory.

As part of both sorting mechanisms, each email address is validated. The validation process ensures that an email address has a valid user name and registered domain comprised of valid symbols and alphanumeric characters. It ensures these two components are separated by a character. It also ensures that the address ends with a recog-

nized top-level or country-code domain. If the address ends with a country-code domain and a sub-domain is present, the sub-domain is checked against the list of known sub-domains for that country (see the Appendix in section A.2 for the complete list) or that it is one of the top-level domains. Any invalid addresses are removed and all valid addresses are sorted after any sub-domains in the address have been removed. Once sorted, the list is checked for any duplicated addresses. If found, the duplicates are counted and removed from the final list.

4.3 Domain Distribution Analysis

Sorted lists are run through a distribution analysis routine which breaks down the distribution of top level, registered, and country code domains within a list. This routine is run on lists which have been sorted by their registered domains. The sorting algorithm maintains a running count of the number of instances of each top-level and country-code domain encountered. Lists sorted by registered domains are necessary due to the number of different registered domains present in some of these lists. By relying on lists presorted by the registered domain, it is easy to count the number of each entries for each registered domain since all of those entries will be together in a block. This prevents the analysis process from having to store the different domains in memory and their counts simultaneously. It also eliminates the need for long lookups in a large table of domains. Because both the country-code and top-level domains have a limited subset of recognized values, it is easy to store these domain counts using hash tables.

For each level of domain, the analysis routine compiles a list of every value encountered and a count of the number of times that value was found. The recognized top level domains and country code domains, including country-specific sub-domains, are listed in the Appendix.

Chapter 5

Determining the Address List Sources

The first goal of this analysis is to postulate how each address list was formed and the source of the email addresses in it. By analyzing the addresses in each list and considering the addresses in the other lists, patterns may be found that indicate the sources used to create the list. While it is impossible to guarantee the accuracy of the conjectures produced by this analysis, it may help to shape the understanding of the formation of these lists and provide better context for attempting to build a classifier for each list.

These techniques may also provide insight into how botnets or spam campaigns determine their targets by analyzing the addresses which received emails. This analysis can be performed after the campaigns have started and may be completed before security researchers have enacted a method of infiltrating a botnet or spam campaign. While these infiltrations often provide a more accurate and greater view of the internal workings of an attack, preparing and executing each infiltration can be both time and resource intensive.

5.1 Google Search

A common method used by botnets and spammers for gathering active email addresses is to collect them from the Internet. These searches are typically performed either by using a web crawler to search websites for character strings which match a pattern describing an email address or by entering email addresses into a public search engine and determining if any matches were found.

The first technique assumes that an email address found on a website is in active use. The accuracy of this technique is heavily dependent on the implementation of the web crawler and the correctness of the string pattern used to search for addresses.

The second technique requires the searcher to generate valid email addresses to search for. This technique's usefulness may be limited by policies enacted by search engine providers to prevent automated tools from performing the type of scans used by this technique. Several of these search engines do allow the use of automated searches if a fee is paid by the searcher to the provider; however, spammers already control a large number of hosts with distinct Internet addresses controlled by their botnet and can bypass these limits by distributing their searches across this network. Another limiting factor of this technique is the correctness and sophistication of the address generation algorithm used to provide the terms to search on. If the algorithm generates addresses with invalid formats it will waste search throughput on useless searches and may fill the resulting address list with entries that are not actual email addresses but returned results indicating success. Alternatively, an unsophisticated algorithm may generate addresses which all have a correct format, but are filled with nonsensical user names or registered domains comprised of random characters. While these are valid addresses, they are less likely to result in a successful search or, if a search was successful, have an active user accessing the account.

Since the addresses collected by the first technique are dependent on the implementation of the web crawler used and the sites it searched, the address lists do not contain enough information to determine if a web crawler was used to collect their

addresses. However, the addresses in a list can be entered into a search engine to determine if they are on a public web site. Address lists which have a high percentage of addresses found using this method are more likely to have used one of the techniques described above to scrape its addresses from the Internet. This search experiment can only determine if an address is publicly available on the Internet. The experiment cannot determine which of the two techniques described above were used to collect addresses for the list; simply that it is likely that addresses were sourced from the Internet.

In this searching experiment, a sampling of 250 addresses from each of the address lists were selected and a manual Google search was performed with each. A search was successful only if it returned a result containing the exact search term was returned (search terms were surrounded by quotation marks to limit misleading results). It should be noted that a larger search may be performed using an automated searcher run on a distributed network with independent machines to bypass the per-IP address limits enforces by many search engines. This ability exceeded the resources available for this analysis necessitating a manual analysis. The results for these searches are shown below in Figure 5.1.

The highest percentage of successful address searches was found in the Storm address list with 35.2% of its searches returning a positive result. The Storm (C&C) and Grum lists share similar outcomes with 32.8% of the addresses in each resulting in successful searches. This indicates that a significant portion of the addresses in each of these three lists were collected from the Internet. The similar results for both Storm lists are also encouraging as differences would conflict with them originating from the same botnet.

In the remaining four lists it seems that while the Internet may have been used as a source of addresses, it was not a significant one. Just because an address can be searched for successfully on the Internet does not mean that it can only be found through an Internet search technique. This naturally inflates the percentage of successfully searched addresses with respect to the percentage that were actually collected using an Internet search based technique for each of the address lists.

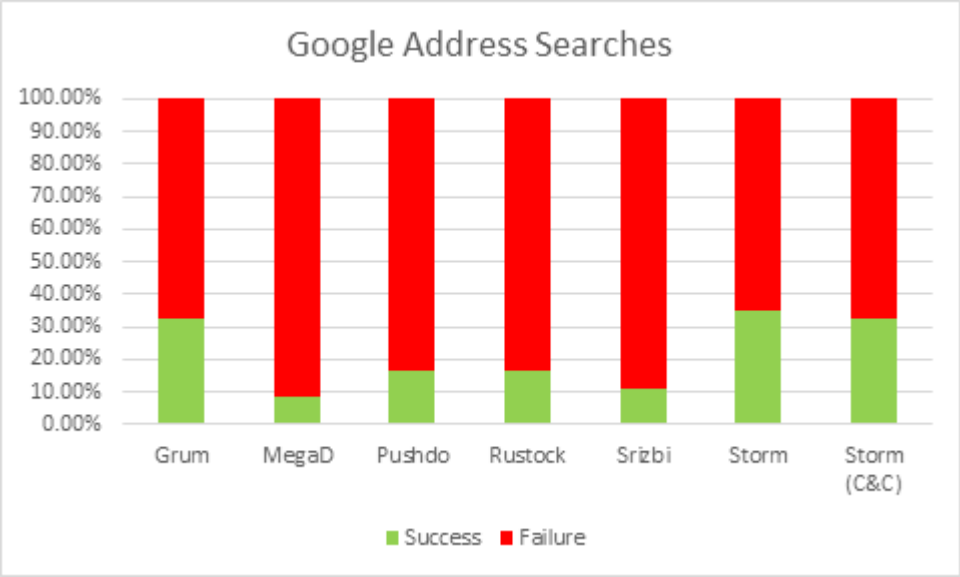


Figure 5.1: The results of the Google searches over 250 random email addresses from each address list. Represented as the percentage of the searches which succeeded and failed for each list.

It should be noted that on average, 25-30% of all email addresses can be found using crawling and searching techniques. This is due to the fact that approximately one in four users explicitly or implicitly make their addresses public by publishing on the web. Users may have accomplished this by using an archived mailing list, blog, forum, or other online public service. This implies that approximately 25% of them should be discoverable through web crawling and search techniques no matter the circumstances by which the addresses were gathered. Given these circumstances regarding the nature of addresses on the Internet and the inherent inability to determine how exactly each address is collected, it is not possible to determine how predominant the crawling technique was for each of these botnets. The results do reveal that the majority of addresses in each list are from sources other than Internet searches. The data shown in Figure 5.1 does indicate that some fraction of each list was collected through crawling or search techniques. However, it is also clear that this is well less than 60% of the addresses in all cases.

5.2 Invalid Addresses

When each list is run through the sorting algorithm, entries in the address list which do not comply with the formatting requirements of valid email addresses are removed as these addresses are likely from a poorly implemented automated process. The presence of invalid addresses does not point to a specific form of automated scraping (web crawling, address searching, hard drive scraping, etc.); the presence of invalid addresses does indicate that automated tools were used in the collection of addresses in the list and that these tools were likely not of high quality. Conversely, a lack of invalid addresses does not indicate that automated tools were not used, but it does indicate that if they were used that they were well implemented.

When the address lists are sorted, the invalid addresses are removed and the number encountered is counted for each list. Table 5.1 shows the percentage of entries in each address list that are not valid email addresses. The majority of the lists have an extremely low percentage of invalid addresses. Both Storm address lists have a considerably higher percentage of invalid addresses than the other lists, however, it is still an insubstantial percentage of the overall lists. The low percentage of invalid addresses in the lists indicates that at least one automated tool had a poor implementation or did not validate the addresses it found. This is likely in a hard drive scraping tool which may grab everything from the known location of the address book but does not validate the scraped entries. However, almost the entirety of each list is valid indicating that almost all the methods of gathering addresses are well implemented or have their findings validated before being used by the botnet.

As discussed previously, Storm was shown to scrape the hard drives of its infected machines to gather additional addresses to target [9]. This may account for the higher rate of invalid addresses found in both Storm lists. However, this scraping functionality may have been implemented correctly, validating addresses before adding them, and might therefore not be the cause of the increased invalidity.

Table 5.1: Outlines the percentage of each address list that do not have a valid email address format.

List Source	Total Addresses	Invalid Address Count	Invalid Address Percentage
Grum	982,888	20	0%
MegaD	177,952,799	20,190	0.01%
Pushdo	6,920,873	0	0%
Rustock	7,152,100	1,942	0.03%
Srizbi	49,439,079	8,417	0.02%
Storm	947,915	905	0.10%
Storm (C&C)	264,443,464	422,411	0.16%

5.3 Duplicated Addresses

The sorting algorithms run against the lists also find any duplicate entries within each list. A large number of duplicate entries indicate one of two things: the botnet did not organize and maintain its lists well enough to prevent duplication, or the method used to analyze the botnet by researchers allowed for duplicate entries in the resulting address lists. Duplicates created by researchers could be caused across multiple scans of the same botnet or if the researchers analyzed the botnet through a man-in-the-middle technique and saw the same address targeted across multiple different campaigns. In both these examples the implementation did not check for already existing entries in the results of the scans to create the duplicate entries. In the second example, the botnet may only store the address once, but it could send spam to that address multiple times causing the researchers to record the additional attacks as duplicate addresses in their resulting list.

Most of the lists contain a low percentage of duplicate addresses as seen in Table 5.2. However, all lists, except for the Storm Botlab list, have a higher percentage of duplicate entries than invalid addresses. It is not clear if these duplicates were created by the botnets' collection methods or by the researchers' analysis techniques. If created by the botnets, duplication is a more prevalent issue for list maintenance than collecting invalid email addresses.

Table 5.2: Outlines the percentage of each address list made up by duplicated entries.

List Source	Total Addresses	Duplicate Address Count	Duplicate Address Percentage
Grum	982,888	8,540	0.87%
MegaD	177,952,799	221,209	0.12%
Pushdo	6,920,873	3,291	0.05%
Rustock	7,152,100	3,490	0.05%
Srizbi	49,439,079	29,670	0.06%
Storm	947,915	1	0%
Storm (C&C)	264,443,464	189,157,009	71.53%

The notable outlier in this is the Storm C&C list where the majority of the addresses are duplicates. The extraordinary percentage of duplicated entries, especially when compared with the other Storm list which has practically none, points to the conclusion that the method used to collect the address list by researchers was highly prone to duplication.

5.4 Cross-List Shared Addresses

Comparing the addresses between lists gives insight into the likelihood that lists shared the same source for a portion of their addresses. Table 5.3 shows the results of the comparisons between different sorted lists.

Table 5.3: Number of addresses shared between lists. The value at the intersection of each column and row is the number of addresses found in both lists. Comparisons which have over 250,000 duplicate addresses are in bold.

		Compared To...					
		Grum	MegaD	Pushdo	Rustock	Srizbi	Storm
Source	MegaD	7,525					
	Pushdo	1	**1,103,225**				
	Rustock	1,975	**524,649**	30,157			
	Srizbi	3,274	**3,223,868**	**261,716**	165,446		
	Storm	489	112,507	6,946	3,596	22,991	
	Storm (C&C)	12,196	**6,681,866**	**550,622**	**251,484**	**2,019,057**	67,172

From the table, it is clear that MegaD had many of the same addresses as Storm (C&C), Srizbi, Pushdo, and Rustock. MegaD shares a particularly large number of addresses with both the Storm (C&C) and Srizbi lists. This may be because the MegaD list has the largest number of valid unique addresses. The Storm (C&C) list also shares a large number of addresses with the Srizbi and Pushdo lists, possibly due to its size.

Where there are a large number of shared addresses, it may be indicative that a single source was used by multiple botnets. Table 5.4 reduces the influence in the differences in scale between the sizes of the lists by comparing the shared addresses as a percentage of the number of unique valid addresses in each list.

Table 5.4: Percentage of addresses shared between lists. The intersection shows the percentage of the Source list which is shared by the Compared To list. Comparisons which are over 5% of the list are in bold.

		Compared To...						
		Grum	MegaD	Pushdo	Rustock	Srizbi	Storm	Storm (C&C)
Source	Grum		0.77%	0%	0.20%	0.34%	0.05%	1.25%
	MegaD	0%		0.62%	0.30%	1.81%	0.06%	3.76%
	Pushdo	0%	**15.95%**		0.44%	3.78%	0.10%	**7.96%**
	Rustock	0.03%	**7.34%**	0.42%		2.32%	0.05%	3.52%
	Srizbi	0.01%	**6.53%**	0.53%	0.33%		0.05%	4.09%
	Storm	0.05%	**11.88%**	0.73%	0.38%	2.43%		**7.09%**
	Storm (C&C)	0.01%	**5.89%**	0.49%	0.22%	1.78%	0.06%	

As seen in the Table 5.4, the addresses in MegaD are a large portion of most other lists. Due to MegaD's size though, the shared addresses do not make up a significant portion of MegaD's addresses. While it does not have as large of an impact, the Storm C&C list also shares a large percentage of the addresses from the other lists. However, due to the size of both the MegaD and Storm C&C lists they have a much lower percentage of their own addresses shared with the other lists.

Over 7% of the Storm address list is shared with the Storm C&C list. Since both these lists were independently collected from the same botnet it is understandable that some crossover in the addresses was encountered.

It is clear that no list is entirely influenced or shared by another. The cases where

a sizable portion of a list is shared by another may be caused by both lists sharing the same source for that portion of the list or by one list selling a portion of its addresses to another list. Alternatively, these cases seem to coincide with the lists with the most addresses which may mean the percentage of shared addresses is simply a coincidence caused by scale.

5.5 Cross-List Address Contiguity

Comparing two lists for groups of contiguous addresses that are the same can also help determine if the lists share the same source. If two lists have large groupings of the same consecutive addresses, it is likely these lists use the same source for a portion of their addresses or that one botnet shares a portion of its internal address list directly with another botnet.

This method is performed on raw unsorted lists which have not been changed since they were collected. Each comparison is a time consuming process as each address in one list is searched for in the second list. If found, the next address in the first list is compared against the next address in the second. This sequential comparison is continued until a mismatch is detected. The search algorithm then returns to the initial address and continues searching the second list until the end of that list is reached. Then the next address from the first list is retrieved and the search through the second list starts over again. Only matches with two or more consecutive addresses are recorded.

This brute-force scanning process is required and cannot be optimized for several reasons. First, because the lists are not sorted, each address from the first list has to be checked against every address in the second. Second, because there are duplicated addresses in these lists, the first match may not be the longest sequential match and therefore the search has to continue through the entire second list for each address in the first list. Third, long contiguous matches which include invalid addresses may be better at identifying an identical source shared by two lists and therefore cannot be removed to decrease the search space of the lists.

It should be noted, that the method used by the original researchers to compile the address lists may not represent how the addresses are stored internally by a botnet. If the addresses were collected by intercepting emails sent by a bot or by intercepting command and control messages the order the addresses were received by the researcher may be altered from the internal list of addresses.

Due to the time consuming nature of this analysis, only a limited number of list comparisons were run. Each comparison is against at least one of the smaller lists to limit the runtime of the analysis. Compared to the number of addresses present in even the smallest lists (see Table 5.2) the results from this contiguity analysis are negligible. When the number of consecutive addresses exceeds two, this becomes even less prevalent. These results do not indicate any shared source between the lists.

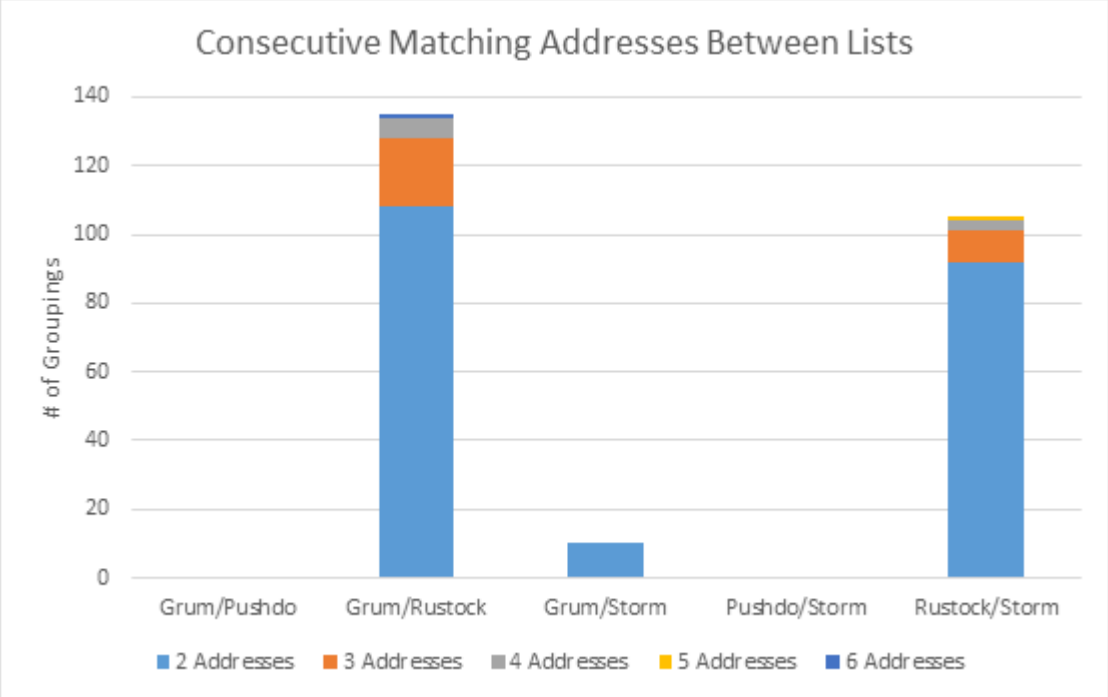


Figure 5.2: The number of consecutive matches of varying size found between address lists. The y-axis represents the number of time a match of a particular size was found. The x-axis list which two lists were compared. The different colors in the legend specify how many consecutive addresses were found in a match.

5.6 Sampled Distinct Addresses

After looking at how addresses are shared between the sources, it is also important to see how they contain addresses which are entirely unique to each list. For this analysis, fifty thousand addresses are randomly sampled from the unsorted raw version of each list. The addresses are left unmodified and were run through a basic lexical sort. Then, any duplicate entries within the list are counted and removed. This is followed by searching, and removing if found, any duplicates in the other lists. The remaining addresses are distinct throughout the samples. The results of this analysis are shown below in Table 5.5.

Table 5.5: The results of the distinctness analysis of the 50,000 address samples. Results include the number of distinct addresses found across the samples and the percentage each sample the distinct addresses form.

	Distinct Address Count	Distinct Address Percentage
Grum	49,895	99.790%
MegaD	49,992	99.984%
Pushdo	49,992	99.984%
Rustock	49,994	99.988%
Srizbi	49,998	99.996%
Storm	49,892	99.784%
Storm (C&C)	47,939	95.878%

Almost all of the addresses in these samples are unique across the entire set of 350,000 addresses. Out of these addresses, only 2298 (0.657%) are not distinct. The Storm (C&C) list shows the most duplicated or shared addresses, but this is unsurprising and corresponds with the results seen previously for this list As shown in the shared address analysis in Section 5.4, these lists share a significant portion of their address, at a much higher rate than shown here. This may have been affected by the sampling process. This may have been affected by the lack of normalization. By not removing invalid addresses and sub-domains below the registered domain, more distinct, and complicated, entries may be found because they are less likely to be duplicated. This complication can come from both sub-domains which are generally uncommon in email

addresses or from having invalid formatting which is unlikely to be repeated.

Chapter 6

Creating a Classifier

The second goal of this analysis is to determine if a classifier can be created using information only available from a list of targeted addresses. Ideally, the distributions of the different domains within the lists can be used in a classification system. The analysis shown below is based on the distributions of top-level, country-code, and registered domains.

In the first set of experiments, the distributions are compared across lists from different sources. In these cases, the desired results are to see large differences in the distributions showing that these values can be used to accurately distinguish between spam sources.

In the second set of experiments, each list will be divided into various sized sub-lists and the distribution analysis will be run again on each of these sub-lists. These distributions will then be compared against the original list's domain distributions. The desired results are the sub-lists will show similar comparisons. This will correspond to a positive identification of the same source for the different subsets.

If these experiments show the desired results, it may be possible to create a classifier with a signature for each botnet or spam source. Then, once a large enough list of targeted email addresses for a new spam campaign has been compiled, it may be used to correctly identify the source of the campaign.

6.1 Cross-List Domain Distribution

The first aspect of the domain distribution based classifier that will be tested is to verify that the distributions can correctly distinguish between lists from different sources. This is performed using the domain distributions gathered in the process described in Section 4.3.

After being sorted, the distributions for each list are then compared against the other lists for each domain type (top-level, country-code, and registered). To more clearly show the resulting differences and similarities, the proportional difference, also known as percentage difference, for each comparison is calculated for each list against the others.

The desired results are large differences for the comparisons showing that the different sources can be distinguished by the domain distributions. The only caveat to this is for the two Storm lists. Since both lists are from the same source, the desired results are to see almost identical domain distributions.

6.1.1 Top-Level Domain Distributions

Out of all the domain categories examined, the Top-Level domains have the fewest possible options. Unsurprisingly, as shown in Figure 6.1, the *.com* domain is the most dominant of all these domains. The *.net*, *.org*, and *.edu* domains also share a reasonable percentage of the addresses for each list while all other domains are insignificant in comparison to these four.

The Grum list contains only *.com* addresses, making it an outlier for comparison with the other lists.

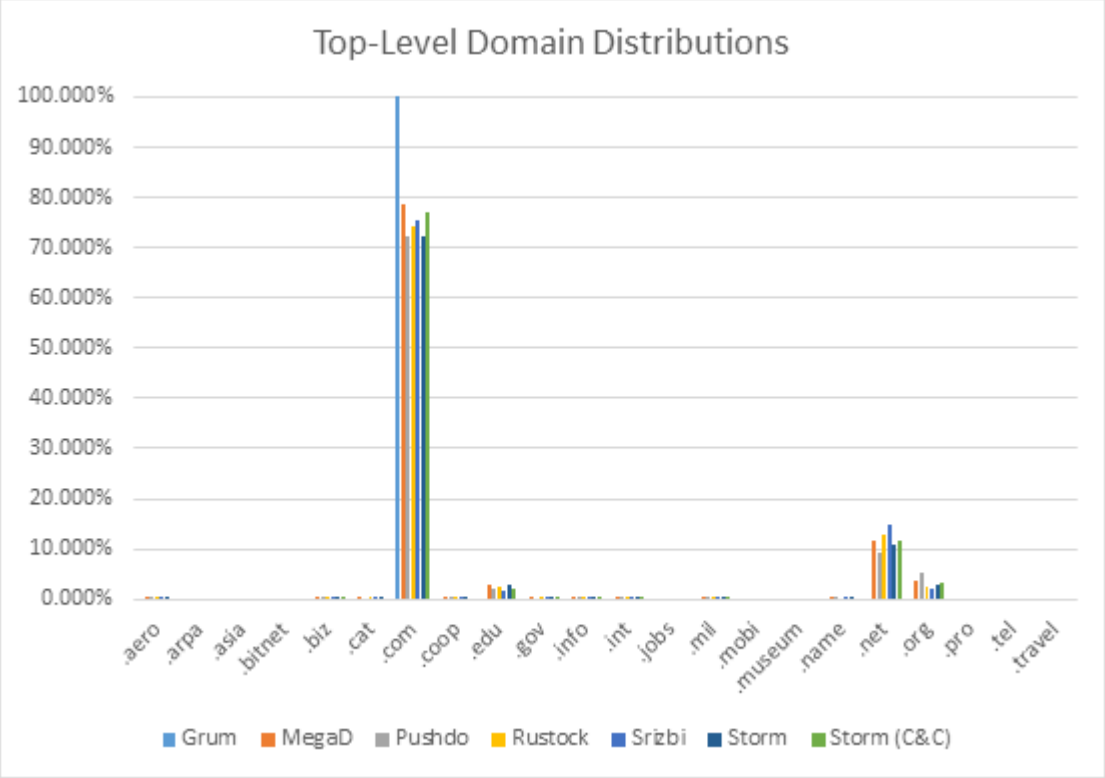


Figure 6.1: The percentage of addresses for each list which end in each of the Top-Level domains listed along the x-axis.

To more clearly show the similarities and differences in the domain distributions between the lists, each list is compared against the others and the percentage difference between the lists is calculated. As stated previously, the greater the differences between the lists, the more likely a working classifier can be built based around the domain distributions. Due to the insignificance of the other Top-Level domains, only the top four domains (*.com*, *.edu*, *.net*, and *.org*) are compared. These differences are shown in the figures below.

Figure 6.2 shows the percentage difference between MegaD and the other lists. It is immediately clear that the Grum list can easily be differentiated. This is not unexpected as the Grum sampling only contains addresses with the *.com* domain. While there are no other differences as drastic as Grum's, each other list exceeds a 20% difference for at least one of the domains and the Pushdo and Srizbi lists exceed this for three

of the four domains.

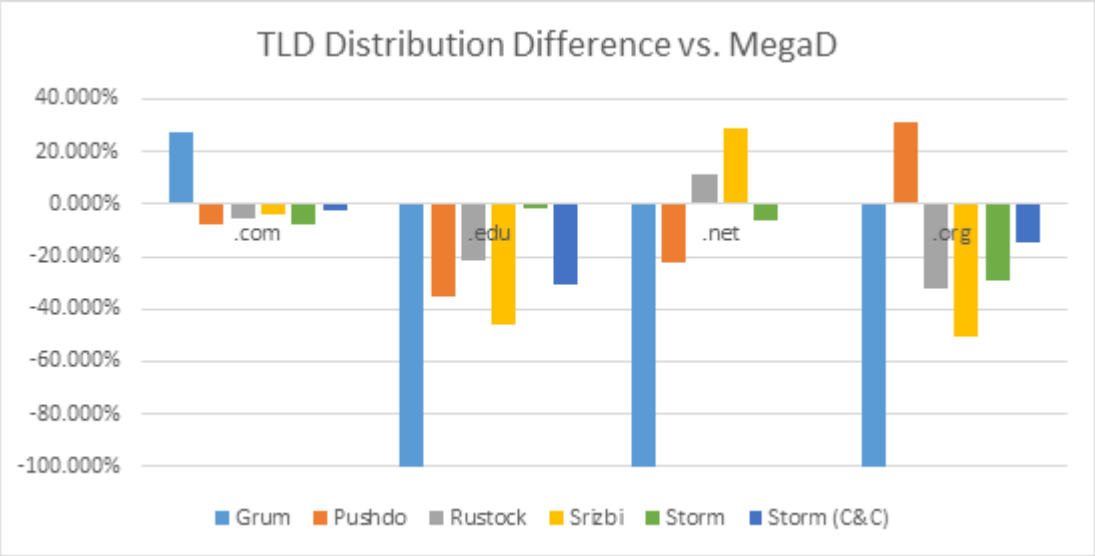


Figure 6.2: TLD Distribution Percentage Difference compared to MegaD.

The *.com* domain shows very little distinction between the lists, other than the previously mentioned Grum outlier, which is unsurprising as that domain dominates the distribution of each of the lists. Both Storm lists have similar distributions to MegaD, only exceeding a 20% difference for one domain each.

The comparisons between Pushdo's Top-Level domain distributions and the other lists are shown in Figure 6.3. It is clear Pushdo has more obvious differences than MegaD's results. Grum has the same differentiations as seen in comparison to MegaD for the same reasons. The *.com* domain again has little distinction between the other lists. However, each of these other lists have at least two domains which exceed a 20% difference and all but the Storm (C&C) list exceed 40% for at least one domain. This shows that it is much easier to distinguish the Pushdo list from the other botnet lists than MegaD was able to from the distribution of the top-level domains for that list.

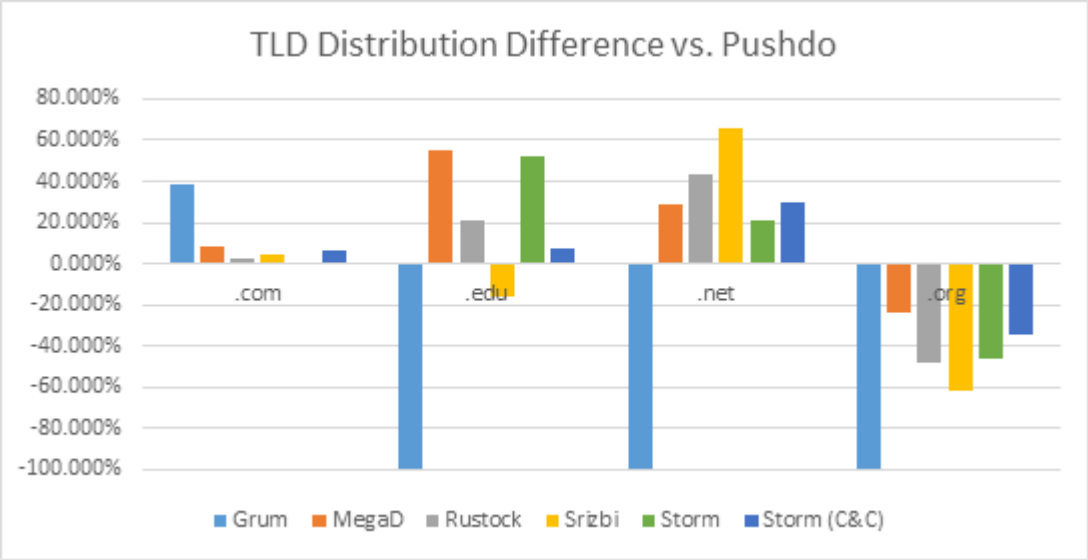


Figure 6.3: TLD Distribution Percentage Difference compared to Pushdo.

The Rustock list is not as clearly distinguishable as the Pushdo list. The top-level domain differentiation between it and the other lists are shown in Figure 6.4. While each list exceeds a 20% difference at least once, only the *.org* domain shows more extreme divergences between the lists. While it may be possible to differentiate between Rustock and the MegaD, Pushdo, and Srizbi lists based on their top-level domain distributions, both Storm lists only show larger divergences for a single domain with each limiting the ability to distinguish the different lists by top-level domain alone.

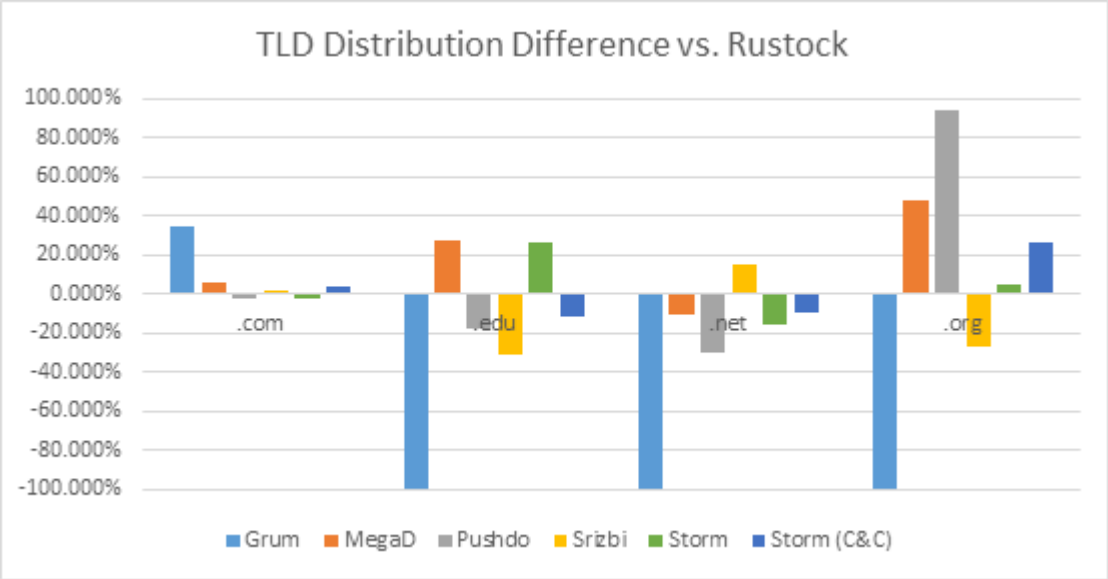


Figure 6.4: TLD Distribution Percentage Difference compared to Rustock.

Figure 6.5 clearly shows that Srizbi can be easily identified in comparison to the other lists by top-level domain alone. Other than the *.com* domain, almost every other domain comparison with the other lists exceeds a 20% difference and several are greater than 50%.

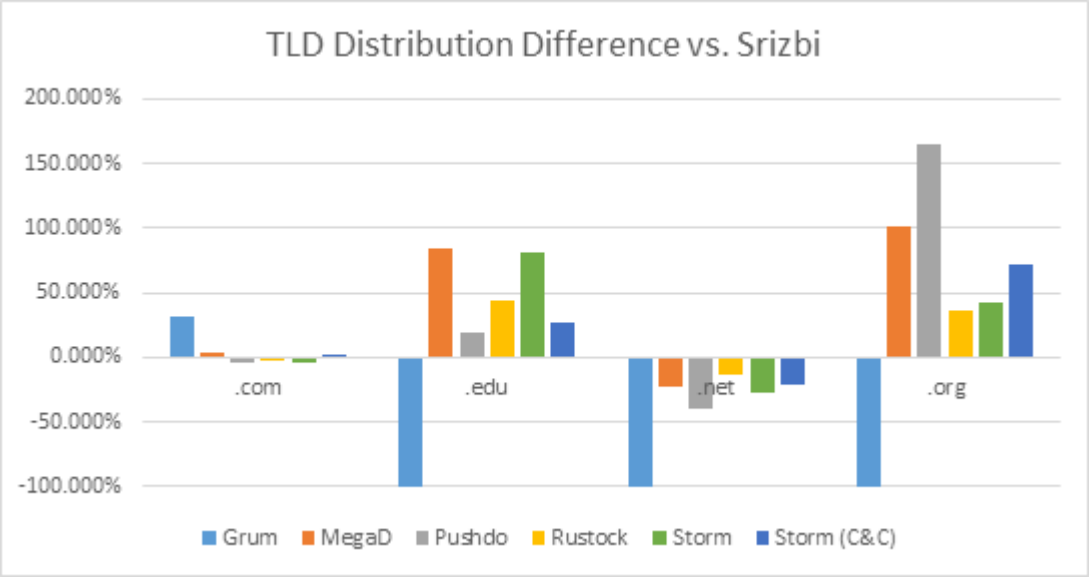


Figure 6.5: TLD Distribution Percentage Difference compared to Srizbi.

The first Storm list's comparisons are shown in Figure 6.6. In this case the desired outcome is to have divergences with all of the lists except the other Storm list. If the other Storm list does not have a similar top-level domain distribution, a reliable classifier will not be able to be devised from this domain type.

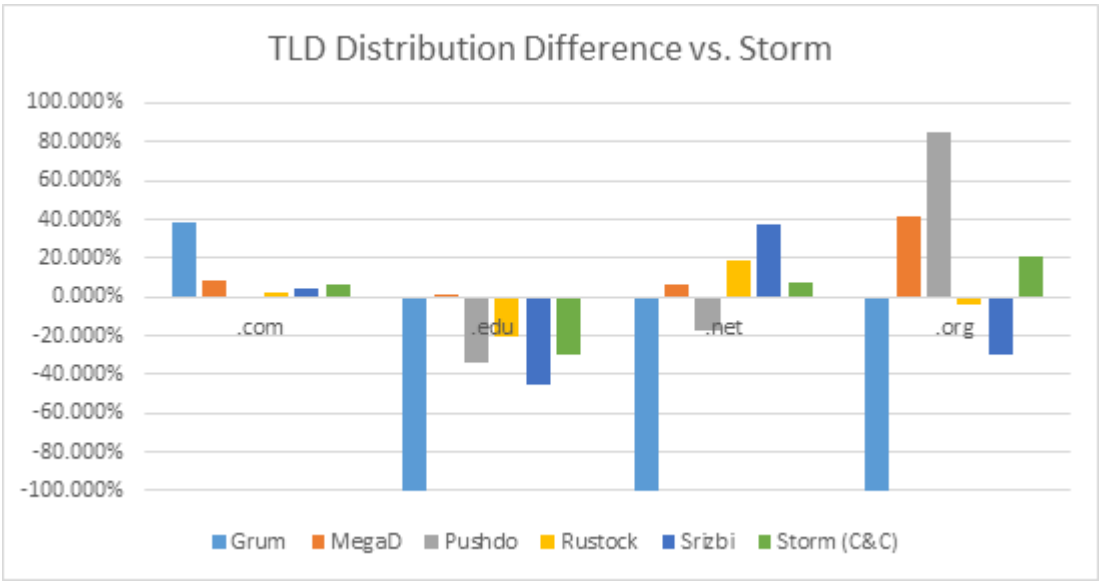


Figure 6.6: TLD Distribution Percentage Difference compared to Storm.

Figure 6.6 shows that most of the lists may be distinguished from the Storm list with Pushdo and Srizbi both exceeding a 20% difference for at least two domains. However, the MegaD list is similar to this Storm list for all domains except the *.org* domain where it does have an outlier greater than 40% and the Rustock list only reaches the 20% mark for a single domain. The other Storm list also shows more disparities than desired as it exceeds a 20% difference for both the *.edu* and *.org* domains.

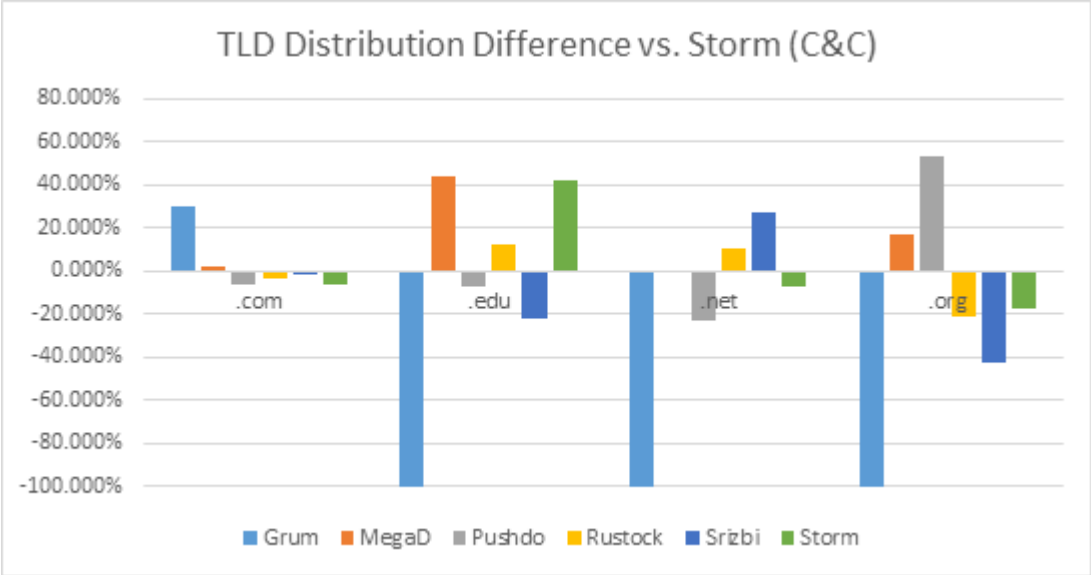


Figure 6.7: TLD Distribution Percentage Difference compared to Storm (C&C).

The final comparison is against the other Storm list and is shown in Figure 6.7. Like the first Storm list, it has several divergences with regards to the Pushdo and Srizbi lists. The MegaD list again only has a single dissimilar domain but in this case it is with the *.edu* domain instead of the *.org* domain. Also, the Rustock domain's differences again exceed for only a single domain, but this is a different domain than the first Storm list. As stated previously, the Storm lists show their discrepancies in the top-level domain distribution for several domains.

Reviewing the comparisons of all the lists shows that while the top-level domain may be able to distinguish one list from another in some instances, it is not a strong indicator. While clear distinctions can be made between Pushdo and the other lists,

there are more cases where a list may be distinguishable from one or two, but not all other lists. To be a good parameter of a classifier, the lists would need to be clearly differentiated in all cases.

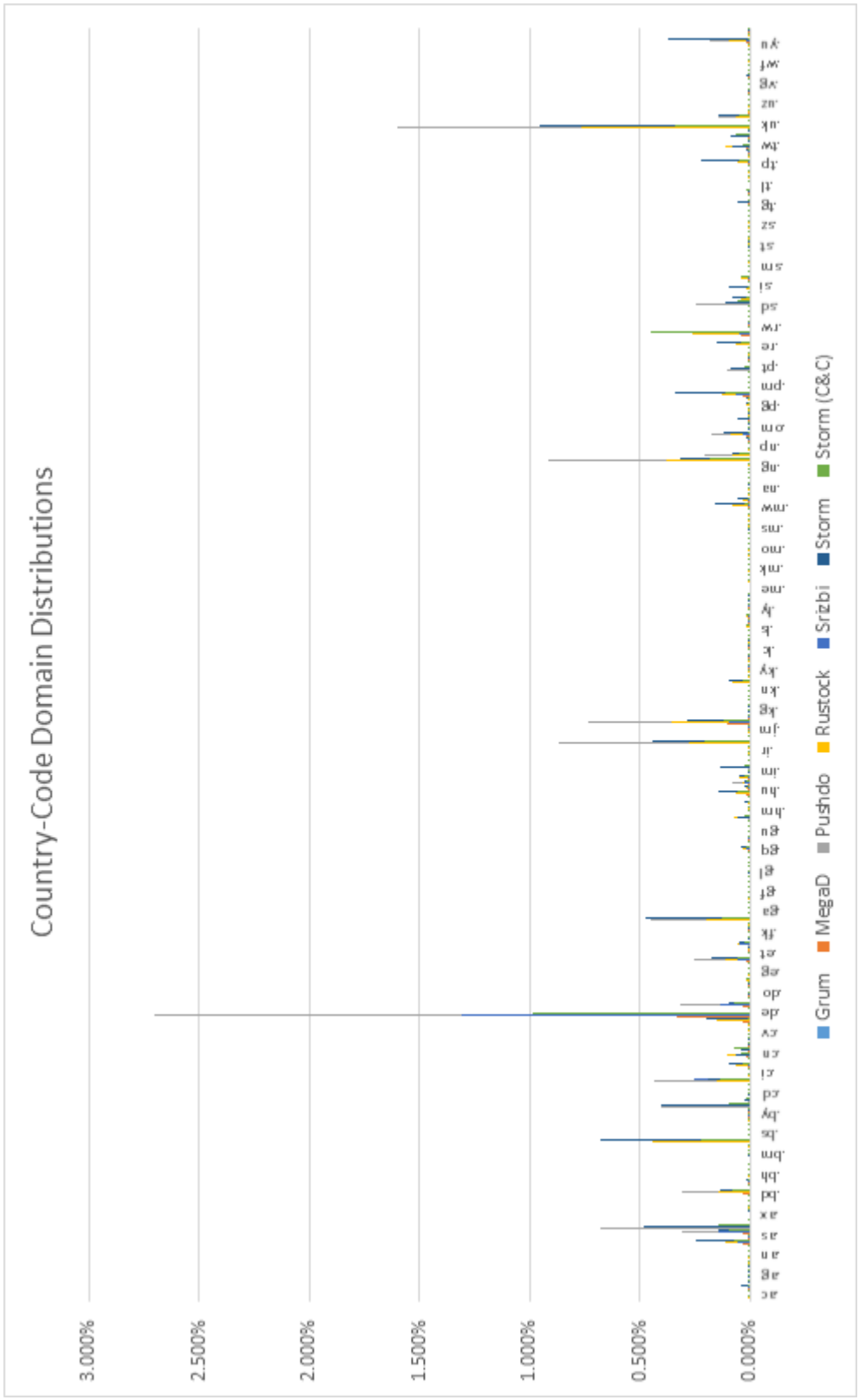
Also showing the limited worth of the top-level domain distributions as part of a classifier are the results of the Storm list comparisons. To be a good classifier, lists from the same source should not have distinguishing results. The top-level domain distribution comparisons between both the Storm lists show dissimilar results making it difficult to identify both lists as belonging to the same botnet from this factor alone.

6.1.2 Country-Code Domain Distributions

The country-code domains provide a much larger space for the comparative analysis between the lists to work in than the top-level domains. While still bound to a set collection of recognized domains available (see the full list in the Appendix), there are more values than there were for the top-level domains. Additionally, while some countries have a greater presence on the Internet than others, and therefore more domains registered to those country-code domains, none of them dominate the domain-space as *.com*, *.net*, and *.org* did for the top-level domains. This should hopefully mean more variety is seen in the domain distribution between lists making the country-code domain distribution a better candidate for building a classifier.

The country-code domain distributions for all the lists are shown below in Figure 6.8. With so many possible domains and with many not making up a sizable percentage of any list, this figure yields few conclusions. One item of note is that the Grum list has no country-code domains in its entire list. It will therefore not be relevant for this analysis.

Figure 6.8: The percentage of addresses for each list which end in each of the Country-Code domains listed along the x-axis.


Country-Code Domain Distributions
3.000%
2.500%
2.000%
1.500%
1.000%
0.500%
0.000%
Grum
MegaD
Pushdo
Rustock
Srizbi
Storm
Storm (C&C)

To remove any noise which may be created by domains with consistently negligible percentages of occurrences, the country domains analyzed will be limited. Figure 6.9 shows a subset of the distributions for country-code domains where each domain must make up at least 0.5% of one of the lists' addresses. Limiting the domains to those which appear in at least 0.5% of a list's addresses removes most of the country-code domains. Of the eleven domains left, only two appear in more than 1% of the addresses for a list.

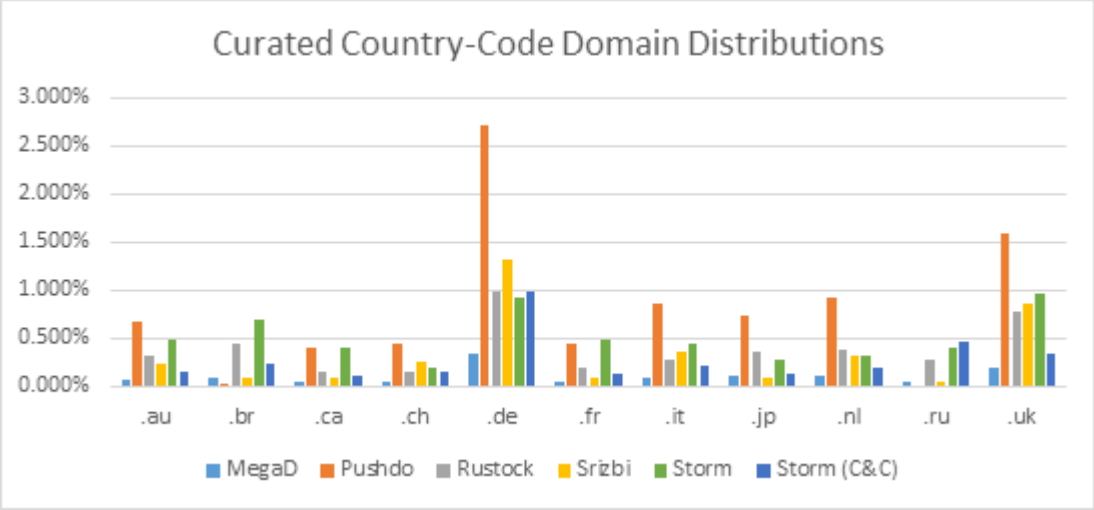


Figure 6.9: The percentage of addresses for each list which end in each of the Country-Code domains listed along the x-axis. This list is a subset of the previous figure and only shows those domains where at least one of the lists has over 0.5% of its addresses using that country-code domain.

The comparison results shown in Figure 6.10 show that the country-code domains in general have a much higher percent differential than the top-level domains did. Because they are a much smaller percentage of the list than the top-level domains, minor fluctuations result in a much greater percent difference from these comparisons.

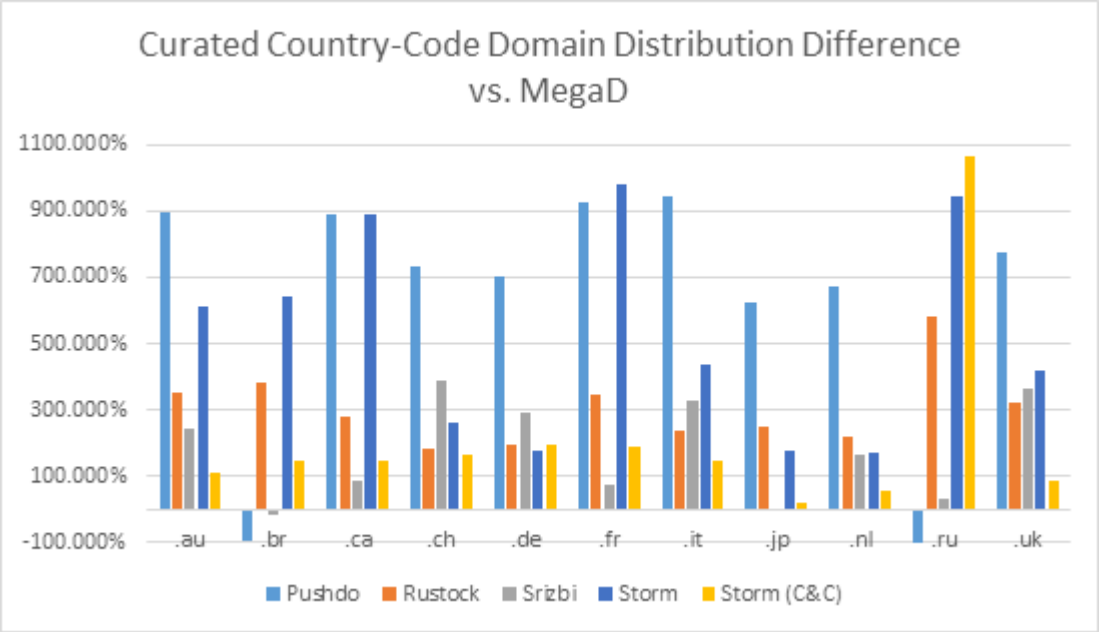


Figure 6.10: Country-Code Distribution Percentage Difference for domains with greater than 0.5% of a list's addresses compared to MegaD.

The percentage differentials for MegaD only drop below a 50% difference with four domain comparisons (three of which were with Srizbi). Only two domain comparisons fell below 20% with *.br* reaching -16.3% percent difference with Srizbi and a -6.9% percent difference for *.jp* with Srizbi. Even with these small differences, it is clear that the differences for these country-code domains easily distinguish MegaD from the other lists.

Comparing Pushdo to the other lists, it is readily apparent that the percent differences for the *.br* domain are so great they overshadow the differences of all the other domains. To better show the comparisons for the other domains, the *.br* domain is removed in the proportional differences for Pushdo shown in Figure 6.11.

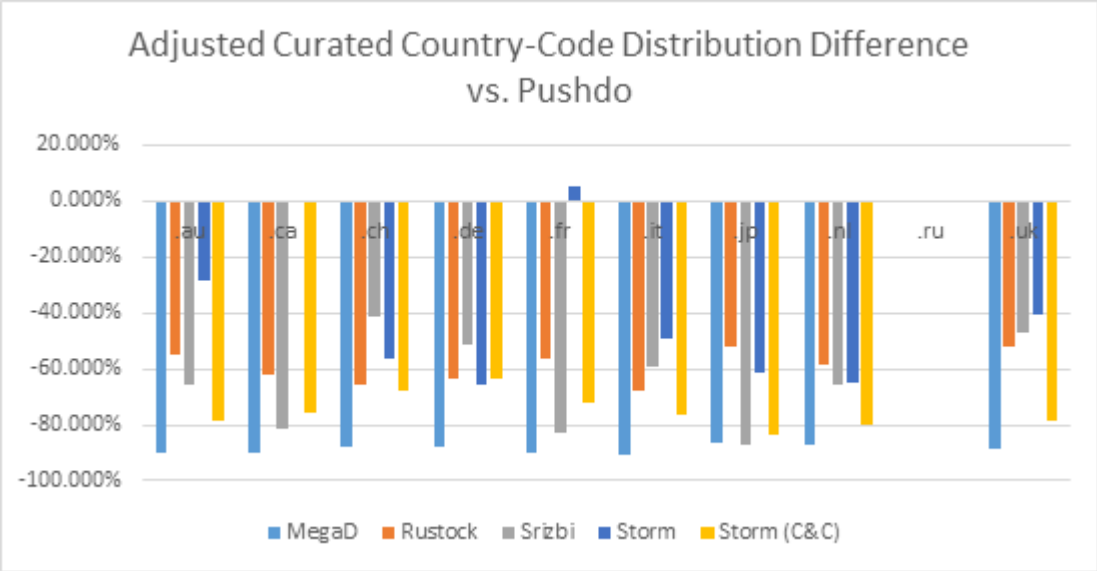


Figure 6.11: Country-Code Distribution Percentage Difference for domains with greater than 0.5% of a list's addresses compared to Pushdo where *.br* has been removed

Pushdo has no *.ru* addresses, limiting a reasonable percentage difference from being found. Each of the other lists have *.ru* addresses making up between 0.039% to 0.454% of their total addresses. Of all the other domains, only two, both in the Storm comparison, have less than a 20% difference.

Rustock does not have the drastic differences some of the other lists have shown with the country-code domains, but it still shows several distinct differences. All of the domains for MegaD and Pushdo have greater than a 50% difference. Both Srizbi and Storm only have two domains each which fall below a 20% difference while Storm (C&C) has three out of the eleven domains.

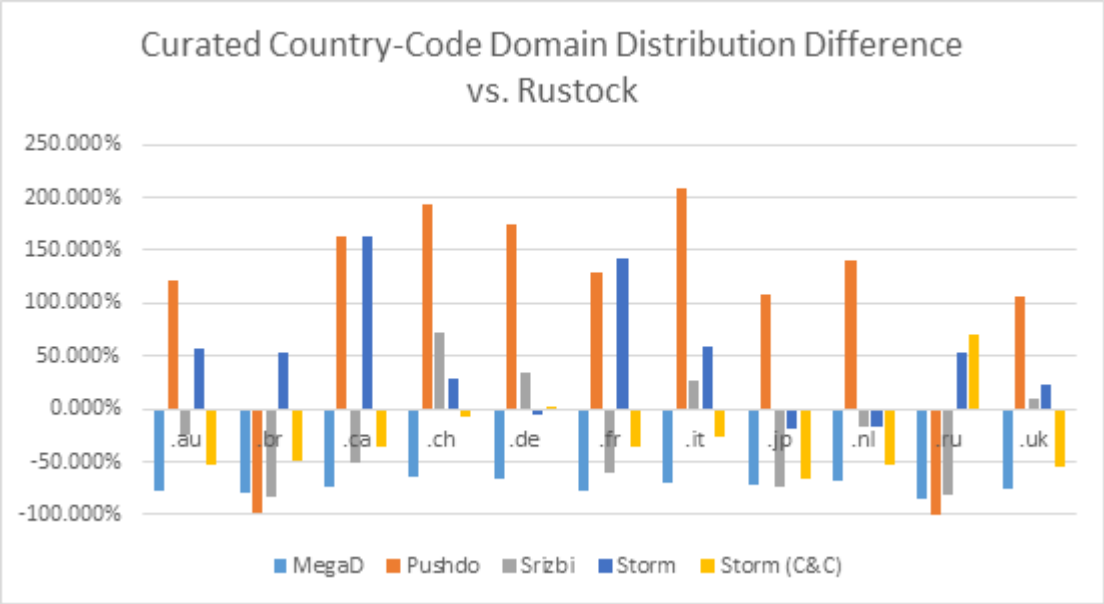


Figure 6.12: Country-Code Distribution Percentage Difference for domains with greater than 0.5% of a list's addresses compared to Rustock.

The Srizbi comparative analysis results are shown in Figure 6.13. Pushdo, Rustock, and both Storm lists all approach or exceed a 500% difference with Srizbi for at least one domain. While MegaD does not have these dramatic results, only two of its domains have less than a 20% difference.

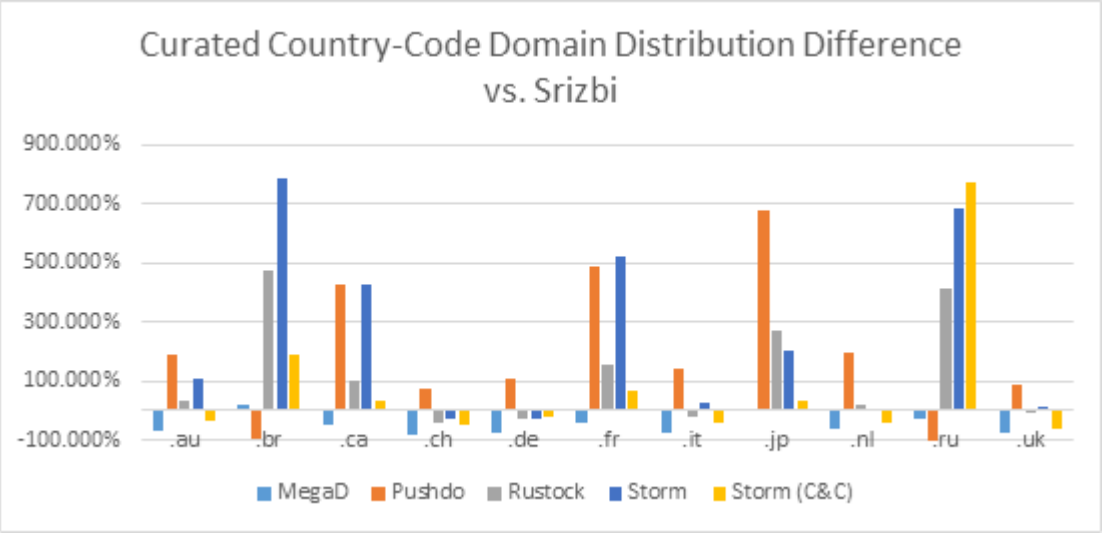


Figure 6.13: Country-Code Distribution Percentage Difference for domains with greater than 0.5% of a list's addresses compared to Srizbi.

The first Storm list also shows distinctions in this analysis, but it is not as clearly distinguished as some of the other lists have been. All of MegaD's domains compare at well above a 50% difference. Both Pushdo and Srizbi have two domains which fall under a 20% difference while Rustock has three. In general, all these lists have enough differences to distinguish themselves using country-code domains from the first Storm list.

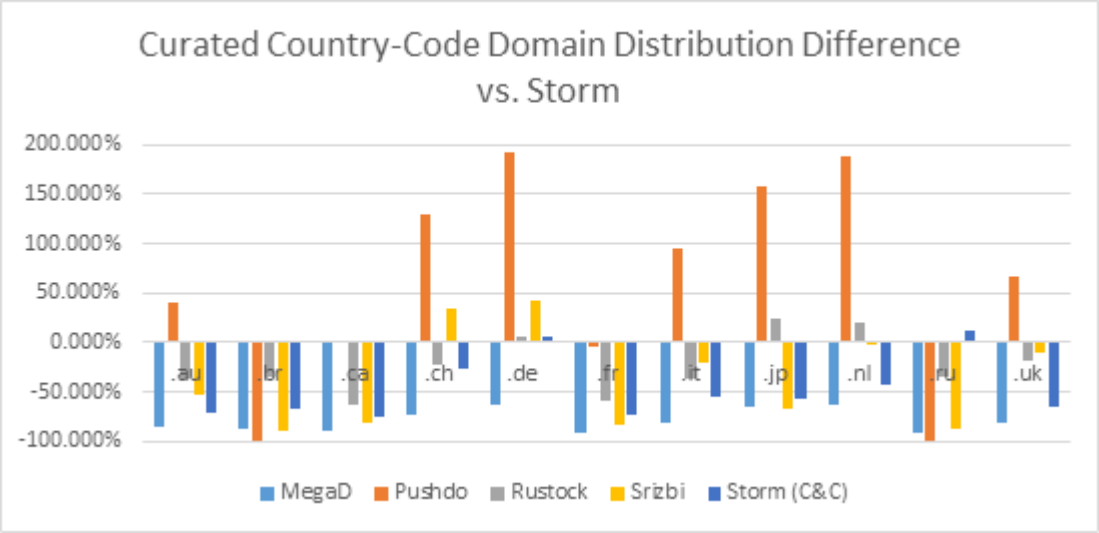


Figure 6.14: Country-Code Distribution Percentage Difference for domains with greater than 0.5% of a list's addresses compared to Storm.

With regards to the second Storm list, ideally, minimal differences will be seen. Instead only two domains have less than a 20% difference and seven of the eleven domains have greater than a 50% difference. Thus the country-code domain distributions between the two Storm lists for these prevalent domains are distinct between the two lists.

The final difference comparison for the country-code domains is for the second Storm list. This list has fewer similar comparisons than the previous Storm list with MegaD only have a single domain under 20% and with two such domains for the Rustock list. The degree of the differences is also greater than the previous Storm list, with several differences surpassing 300

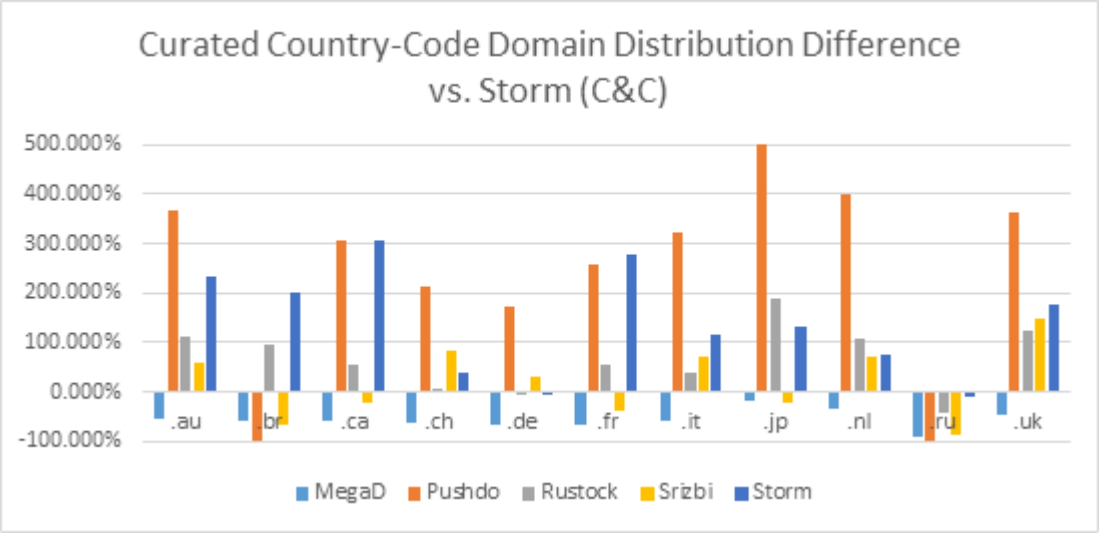


Figure 6.15: Country-Code Distribution Percentage Difference for domains with greater than 0.5% of a list's addresses compared to Storm (C&C).

In general, the country-code domain analysis shows that the majority of distribution comparisons between lists correctly show divergent results which distinguish the two lists as being from different sources. In the cases where a list has several similar results, the other comparisons are enough both in quantity and size to still show the lists as different.

These results follow the behavior needed to build a classifier as they accurately distinguish between lists from different sources. However, when the source is the same for two lists (as seen with the Storm lists) the country-code domains analyzed do not show comparable results and cannot be used to identify both lists as sharing the same source. Therefore, the county-code domain distributions may be useful for distinguishing lists from different sources, but they cannot correctly classify lists from the same origin.

6.1.3 Registered Domain Distributions

The final domain type analyzed is the registered domain name of the addresses. Unlike the previous two domain types, the registered domain space is not bound to a set

of specific values. This creates a diverse distribution set to work with.

Due to the enormous name space for the registered domains, some lists have over one million unique registered domains in them, using the full set of registered domains encountered for the distribution analysis is not feasible. This is reinforced when as only 10-15 of these domains are present in over 0.5% of the addresses for each list.

To resolve this, two more manageable sets of data are used in the analysis of the registered domains. The first set is comprised of the top ten most prevalent registered domains from each list resulting in twenty-eight different domains. This set is referred to below as the curated set of registered domains. The second set is referred to as the largest registered domains and contains every registered domain which occurs in over 1% of the addresses for at least one list. This second collection only has nine domains showing just how few websites make up a significant part of these lists.

The registered domain distributions for all the lists are shown below in Figure 6.16 for the curated subset of domains. It is clear that the Grum *.hotmail* domain percentage is drastically larger than every other entry, skewing this graph. The only other registered domain present in the Grum list is the *.msn* domain which is found in over 11% of its addresses. By removing the Grum results from the graph, a clearer view of the registered domain distributions for the other lists can be seen in Figure 6.17.

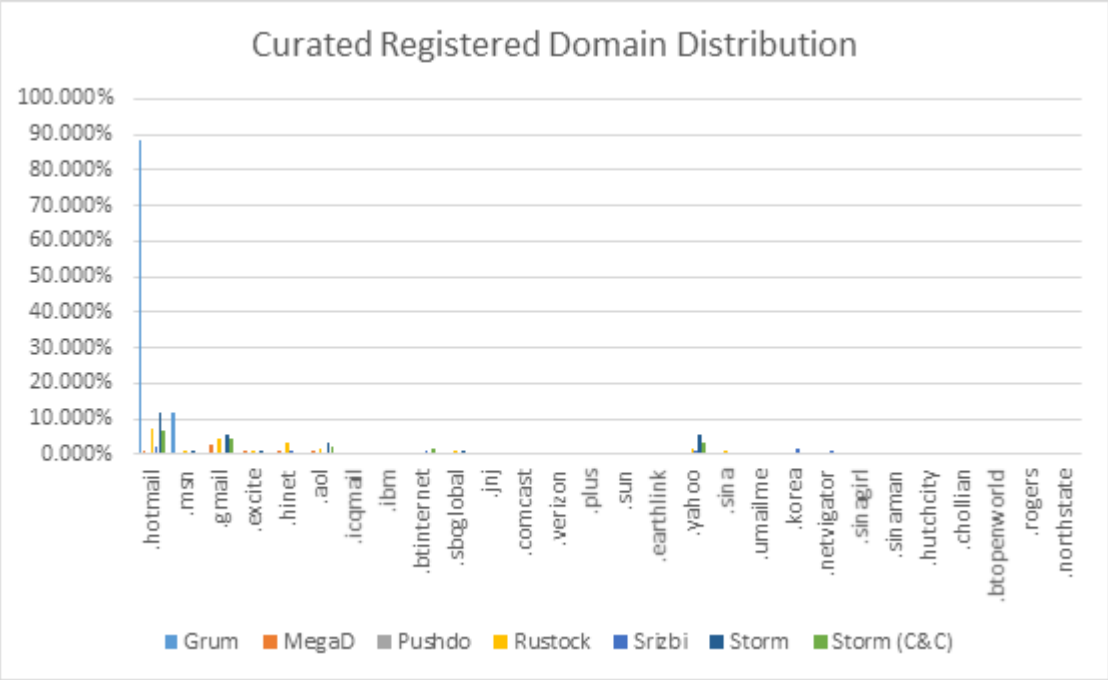


Figure 6.16: The percentage of addresses for each list which contain the registered domains listed along the x-axis. The set of registered domains are the top 10 most prevalent domains from each of the respective lists.

In this adjusted view of the curated registered domain distributions it is unsurprisingly shown that *.hotmail*, *.gmail*, and *.yahoo* all have significant portions of the domain distribution for these lists. As three of the major free email services on the Internet, it is only natural that they would be a common occurrence in the targeted addresses.

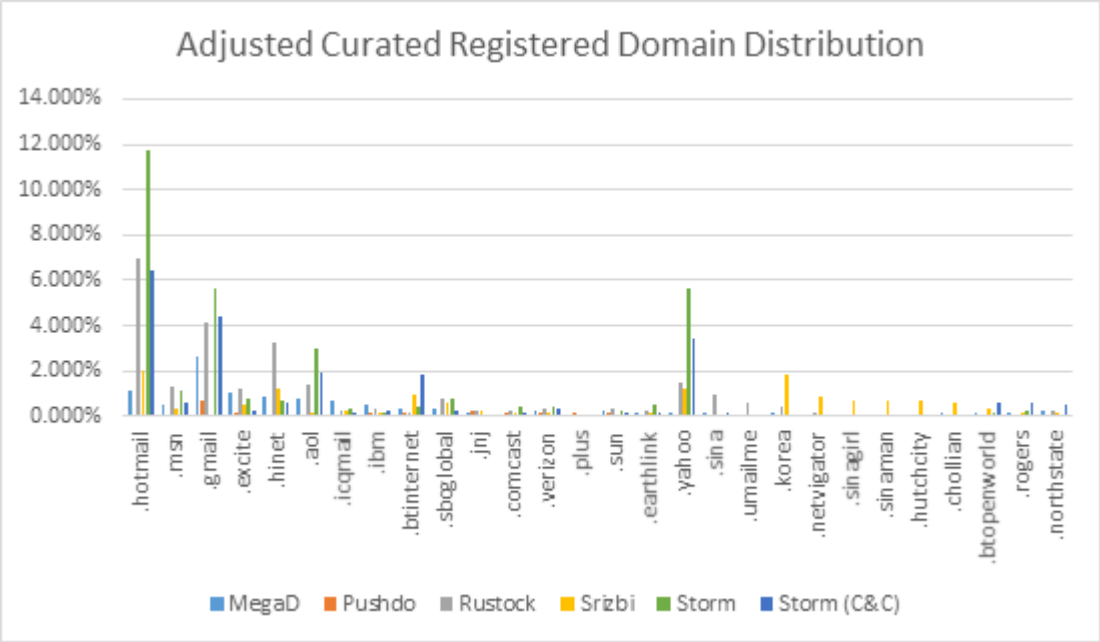


Figure 6.17: The percentage of addresses for each list containing a registered domain listed along the x-axis. This adjusted figure does not contain the Grum list.

Figure 6.18 shows the most influential domains in even greater detail by limiting them to a subset which exceeds a 1% address occurrence in at least one list. In addition to the previously mentioned domains, the remaining domains in this list were all popular Internet Service Providers (*.aol*, *.msn*, and *.hinet*) or personal content networks (*.excite* and *.aol*) which may account for their significant numbers.

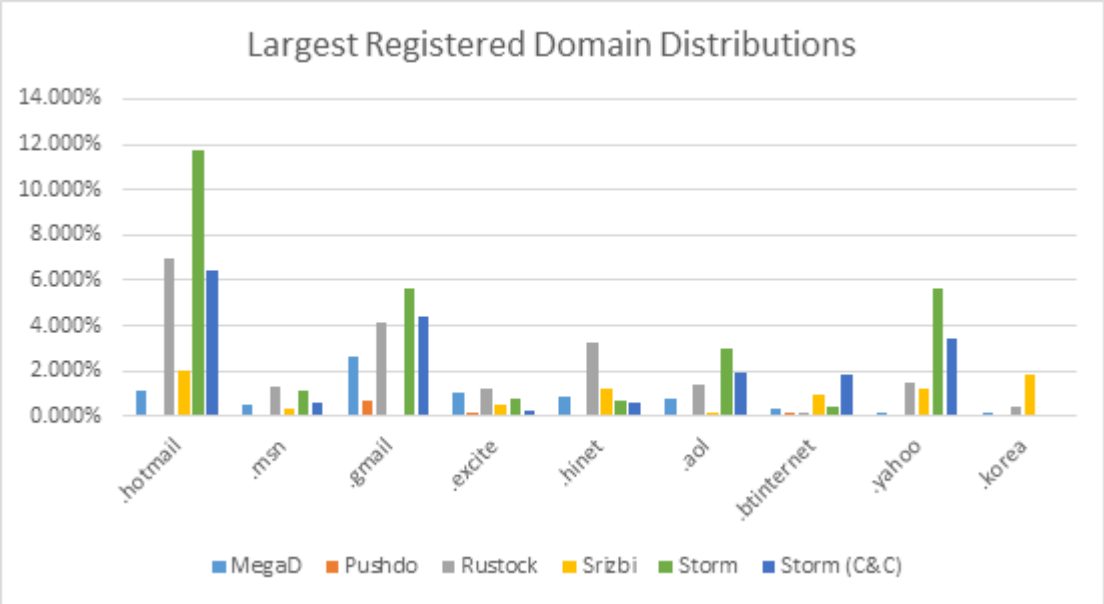


Figure 6.18: The percentage of addresses for each list which contain the registered domains listed along the x-axis. The set of registered domains are those which are encountered in over 1% of the addresses from at least one list. Grum is not included in this graph as its distribution skews the y-axis in comparison to the other lists.

Analyzing the proportional difference in the distribution between lists will again help to determine if these domains can be used to help distinguish between lists. This analysis is not shown for the comparison against the Grum list as it was so obviously distinguishable from the other lists with only two registered domains present in the entire list.

The results of the proportional difference with regards to MegaD for the curated domains can be seen in Figure 6.19. It is clear that Srizbi is significantly different from Pushdo for several domains. Pushdo only has seven comparisons which fall below a 20% difference and only two of these fall below 10%. Only one of these questionable comparisons is for a domain found in the largest domain subset shown in Figure 6.20. In this case the *.excite* domain has an approximately 11% difference in comparison to the Rustock list. With both these subsets, the MegaD list shows clear differences with the other lists.

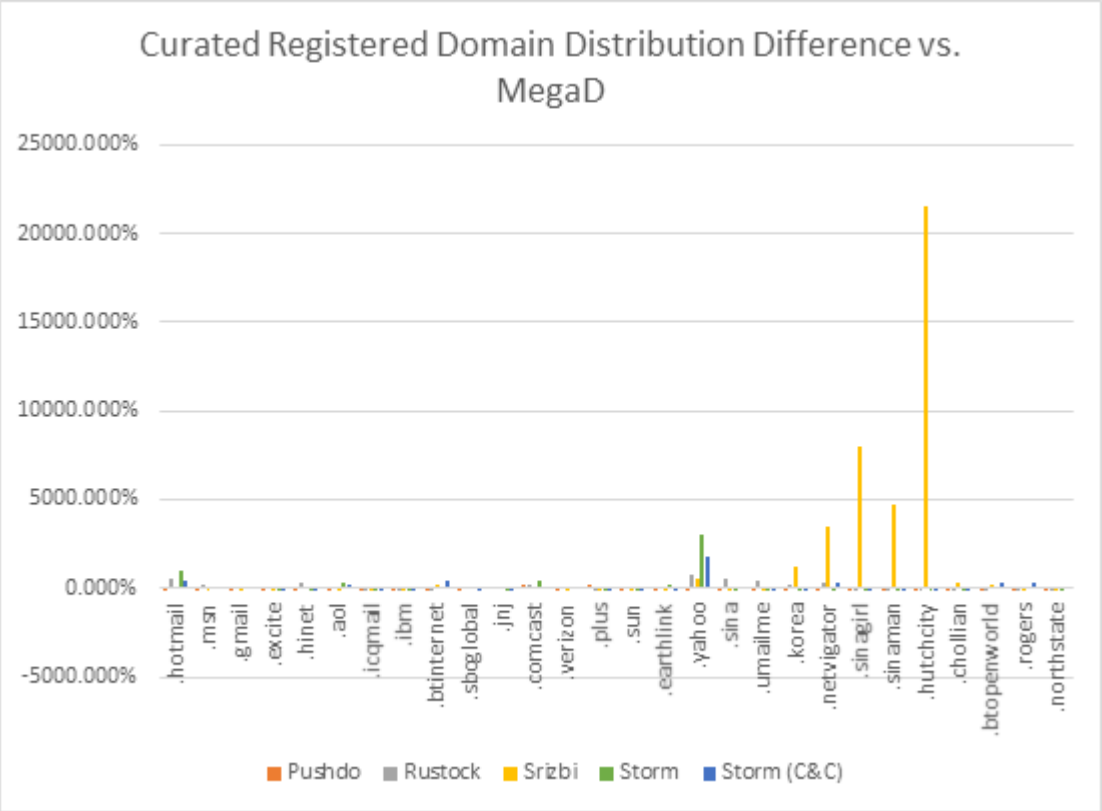


Figure 6.19: Registered Domain Distribution Percentage Difference for top-10 domains from all lists compared to MegaD.

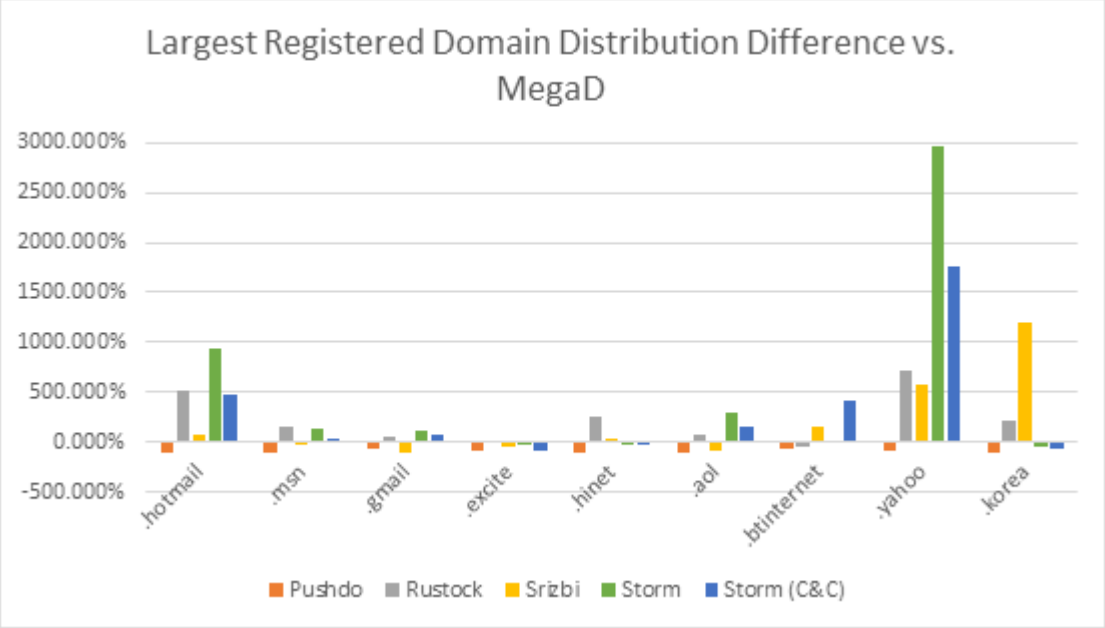


Figure 6.20: Registered Domain Distribution Percentage Difference for domains with greater than 1% of a list's addresses compared to MegaD.

The Pushdo lists comparisons to the other lists shows even more drastic difference, as seen in Figure 6.21. While Pushdo has nine comparisons that fall below a 20% difference, these similarities are overridden by the exceedingly large differences found for other domains. Several comparisons show over a 100,000% difference. The results for the *.hotmail* domain are misleading in the figures for Pushdo. Since Pushdo has no addresses with the *.hotmail* registered domain, the percentage difference is undefined.

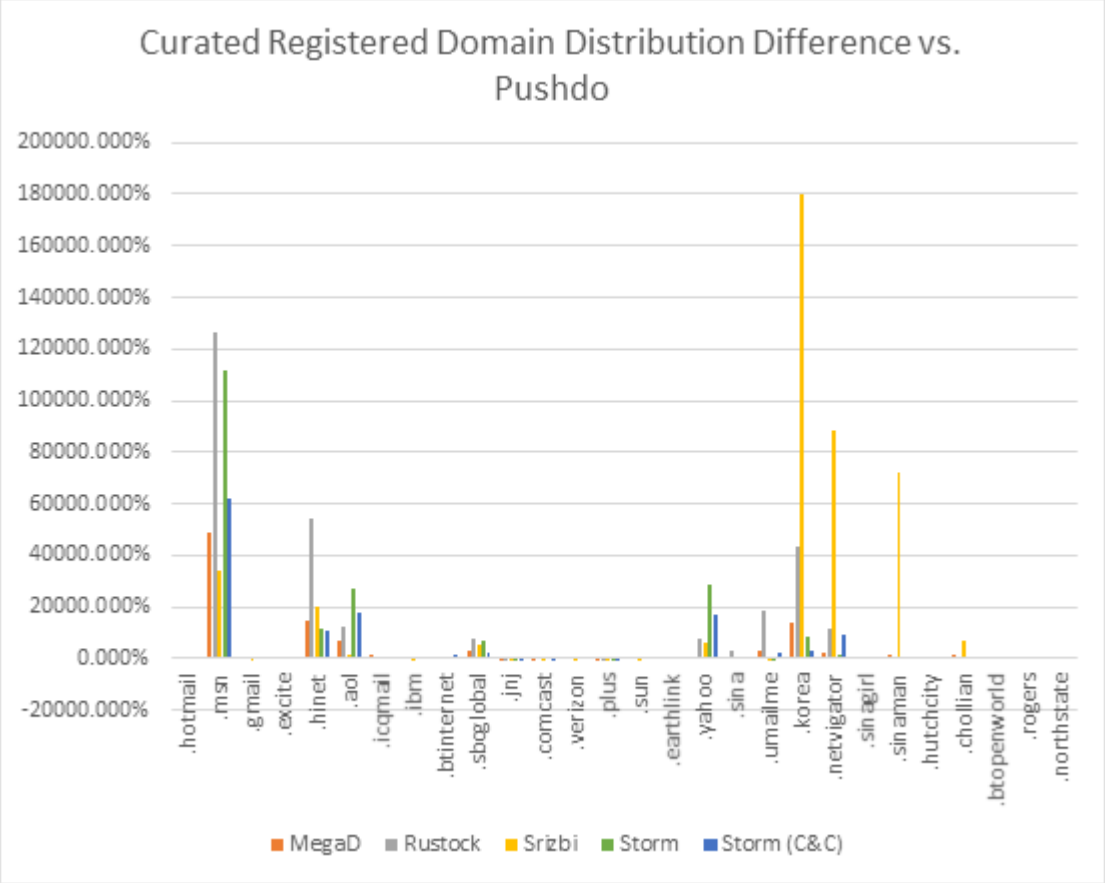


Figure 6.21: Registered Domain Distribution Percentage Difference for top-10 domains from all lists compared to Pushdo.

The subset of largest domains unsurprisingly also shows the strong distinctions between Pushdo and the other lists. While *.btinternet* looks similar, it only is for the Rustock list with a 20% proportional difference. The other lists range from a 170% difference to over a 1200% difference for this domain. Additionally, each list has at least three of the domains from the subset which show over a 10,000% difference.

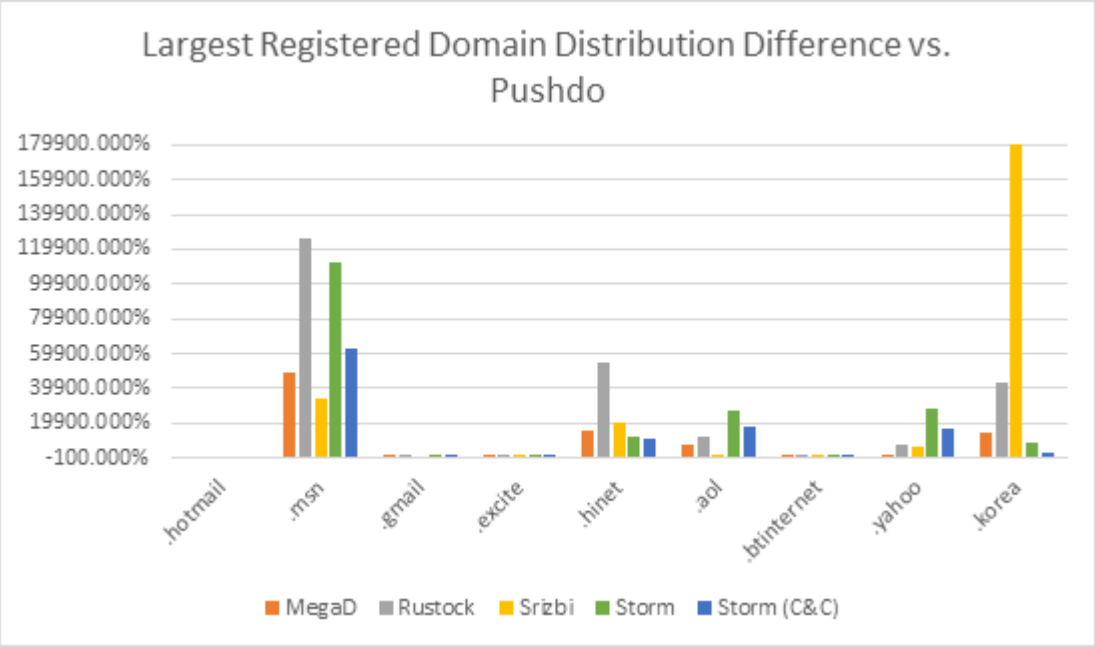


Figure 6.22: Registered Domain Distribution Percentage Difference for domains with greater than 1% of a list's addresses compared to Pushdo.

The Rustock domain shows more similarities with the other lists in the registered domain distributions than any of the previous lists have shown. However, the comparisons for the *.sinagirl*, *.sinaman*, and *.hutchcity* domains with the Srizbi list are considerably larger than the other results. These three differences are removed to better show the other comparisons in Figure 6.23.

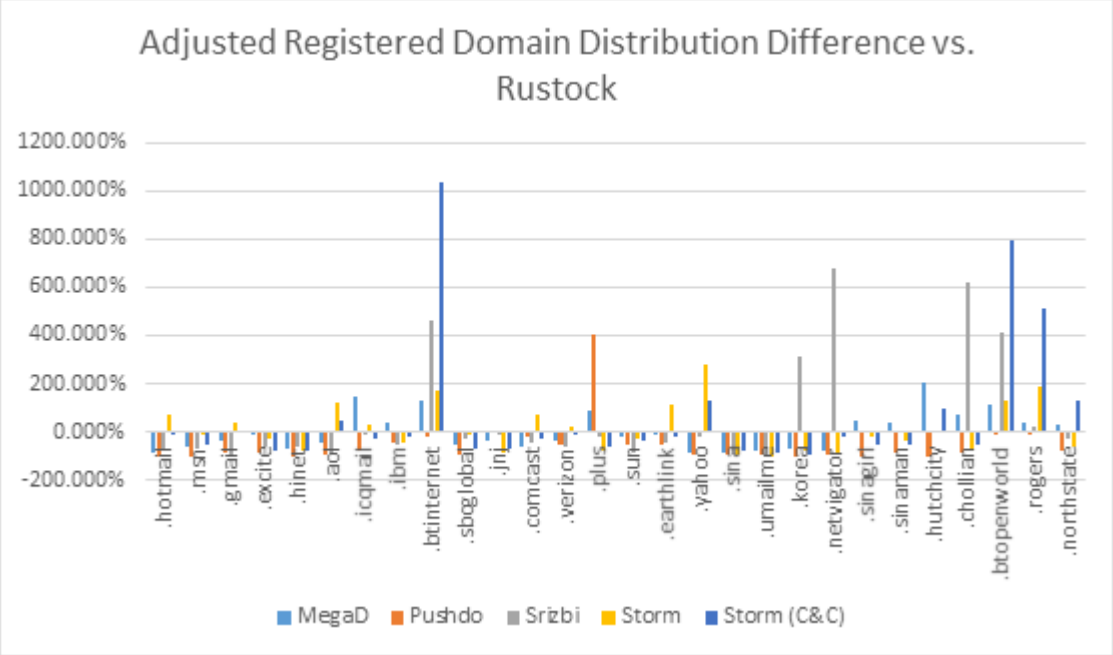


Figure 6.23: Registered Domain Distribution Percentage Difference for top-10 domains from all lists compared to Rustock. The comparisons for *.sinagirl*, *.sinaman*, and *.hutchcity* for the Srizbi list have been removed to better show the other comparisons.

Rustock has nineteen comparisons which fall within only a 20% difference and ten of these are below 10%. The Srizbi and both Storm lists each have four of these less distinguished comparisons and Pushdo has five. A large portion of the comparisons with the Rustock list result in large negative differences showing that these domains make up a larger percentage of Rustock in general. The Srizbi and Storm (C&C) lists each also each have several domains with over 300% more addresses than Rustock.

While these minor issues continue to be present in the largest domain subset, they are less significant. As seen in Figure 6.24, only three domain comparisons are below 20% difference with an additional three comparisons below 10%. Only the Storm (C&C) list has two of these less distinct entries whereas the other lists each have one. None of them occur for the same domain.

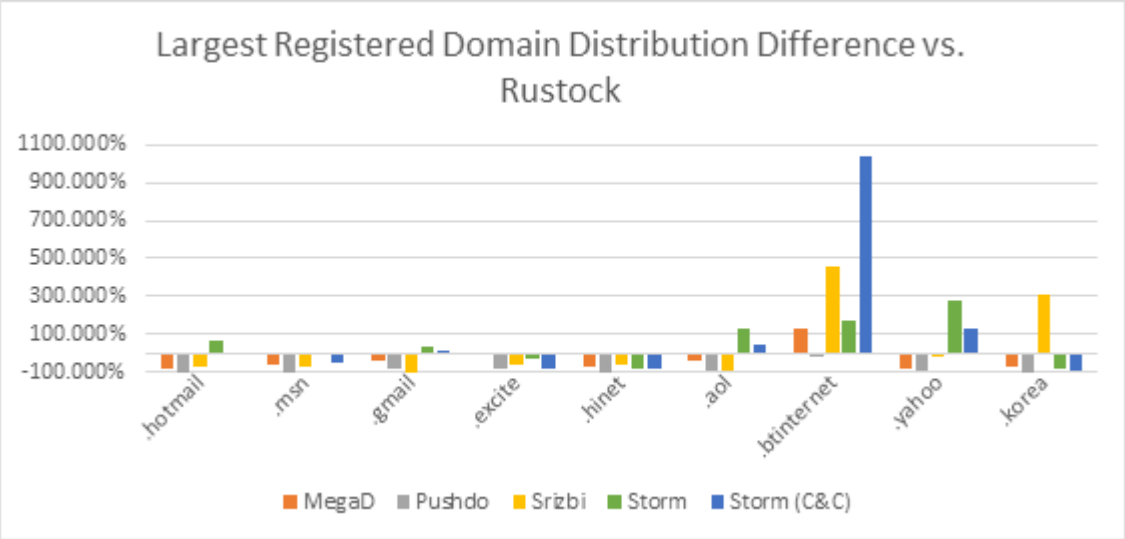


Figure 6.24: Registered Domain Distribution Percentage Difference for domains with greater than 1% of a list's addresses compared to Rustock.

Although there are more similarities present in the Rustock comparisons than previously seen, there are still differences in each subset to allow the lists to be distinguished. This is especially true with the Srizbi list which has multiple extreme differences.

Like the Rustock comparisons, the Srizbi proportional differences for the curated subset contains a few significantly different comparisons. Specifically, the comparisons for the *.gmail* domain are largely different for each list. By removing this domain from the chart in Figure 6.25 a clearer picture of the other domain comparisons can be seen.

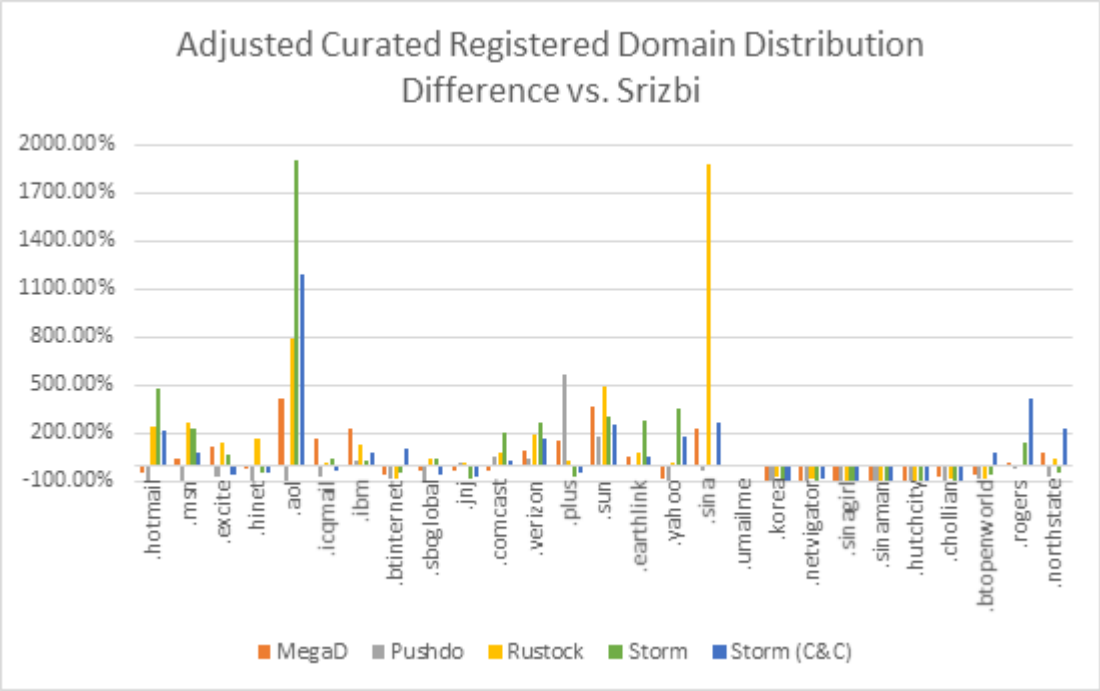


Figure 6.25: Registered Domain Distribution Percentage Difference for top-10 domains from all lists compared to Srizbi. The *.gmail* domain has been removed to better show the other comparisons.

Only seven domains have lower than a 20% difference. Three of these comparisons are with the Rustock domain. Additionally, the *.umailme* domain does not occur within the Srizbi list resulting in a deceptive entry in the graph.

The distinctions between Srizbi and the other lists are even clearer in the largest domain subset, shown in Figure 6.26. In this case, no comparison falls under the 20% threshold. Even without the obvious differences seen in the *.gmail* domain, it is clear that the Srizbi list can be easily distinguished from the other lists using the registered domains.

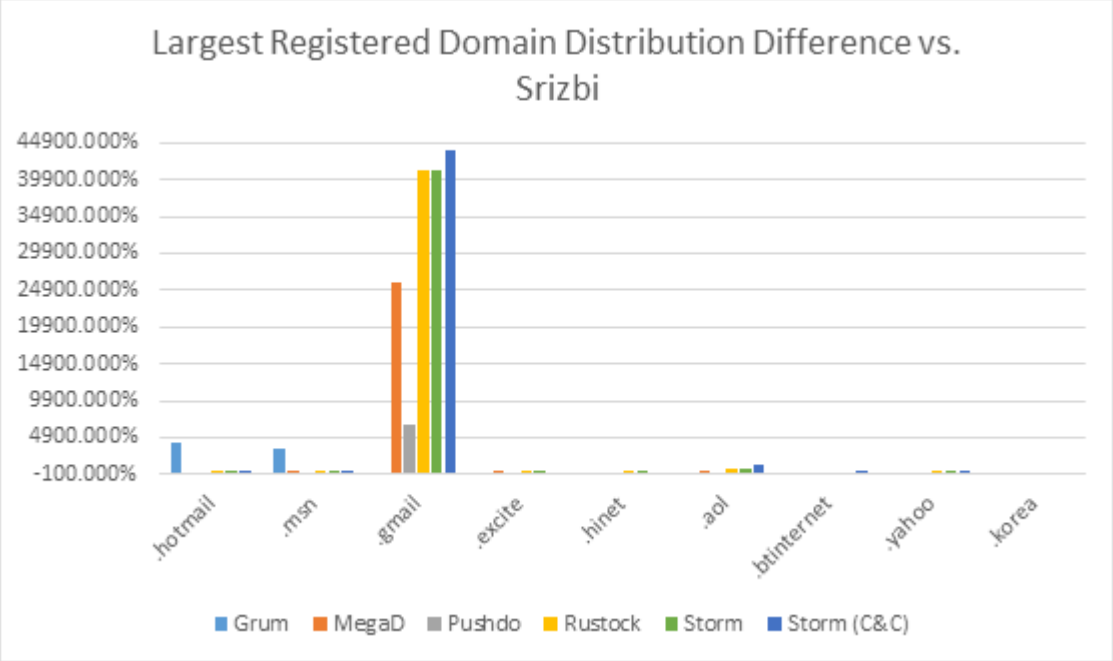


Figure 6.26: Registered Domain Distribution Percentage Difference for domains with greater than 1% of a list's addresses compared to Srizbi.

Like several of the other lists, the comparisons with the first Storm list had several drastic results, some with well over a 10,000% difference. These extreme differences again make the other results difficult to see in Figure 6.27. By eliminating those results which surpass over a 5,000% difference from the chart, a clearer view of the other comparisons can be seen in Figure 6.31.

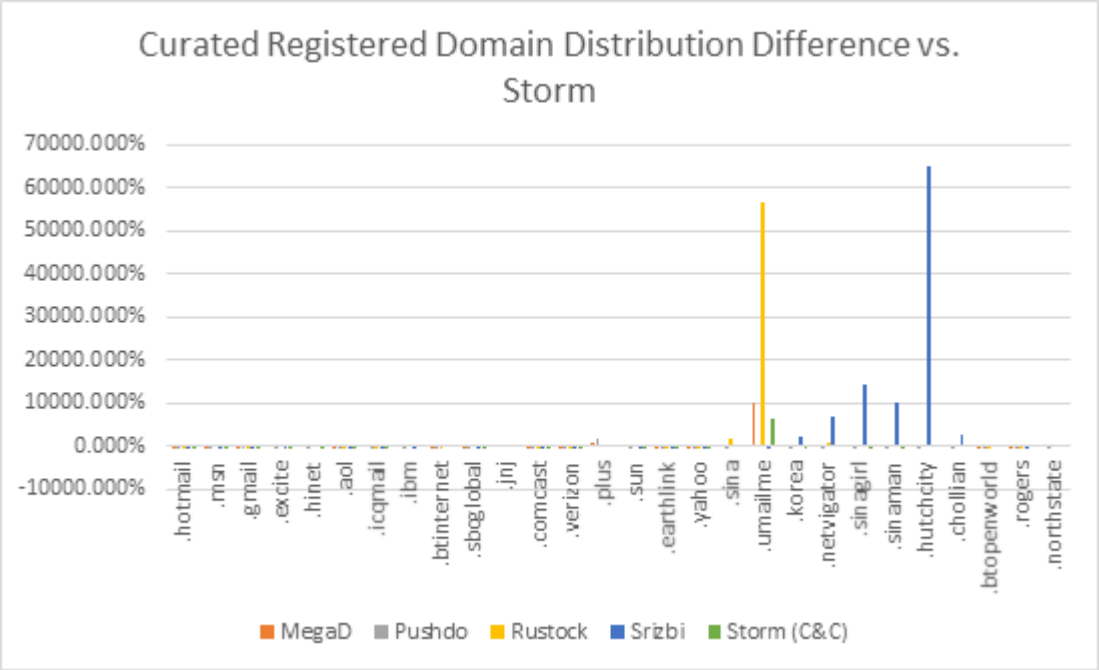


Figure 6.27: Registered Domain Distribution Percentage Difference for top-10 domains from all lists compared to Storm.

The adjusted graph shows that there are still several strong differences for each of the lists. There are several smaller results with twelve comparisons falling under a 20% difference and five under 10%. The Rustock domain has the most similarities with five out of the twenty-eight domains from these lower results.

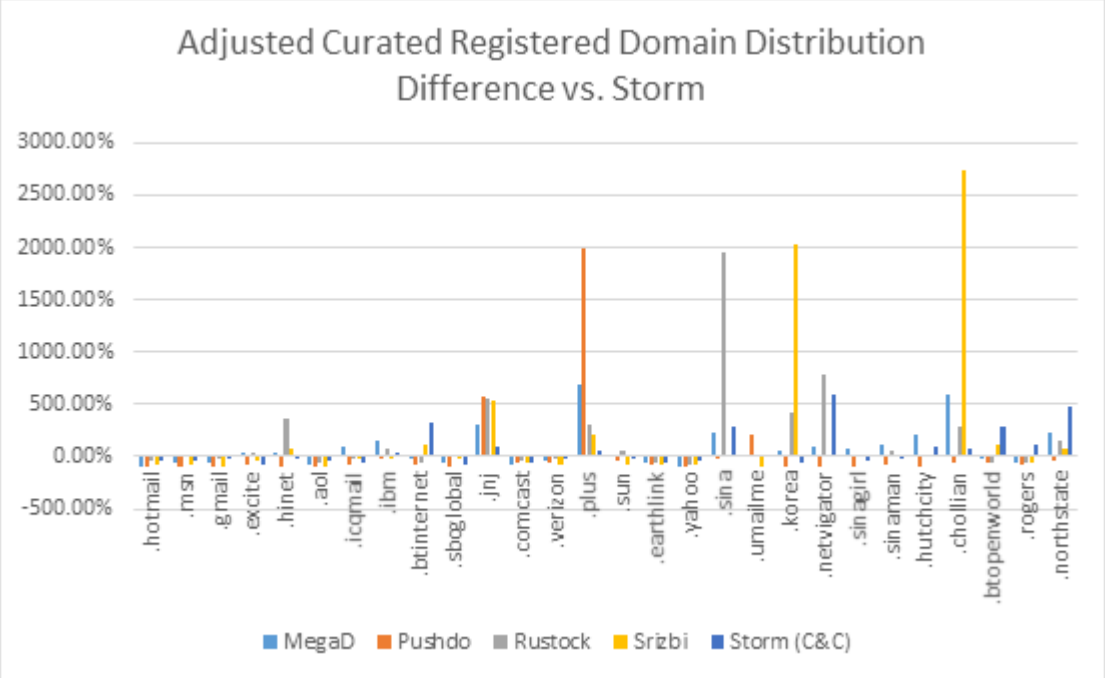


Figure 6.28: Registered Domain Distribution Percentage Difference for top-10 domains from all lists compared to Storm. The results which exceed a 5,000% difference have been removed to better show the other differences.

Ideally both Storm lists would show few differences to indicate that the registered domains selected in this subset could identify both lists as originating from the same source. However, only two domains have less than a 20% difference in this comparison. Additionally there are several domains which have well over a 100% difference and one with a 6,400% difference. This shows the curated subset of registered domains does not correctly relate the two Storm lists.

The subset of largest registered domains show similar results. Many of the extreme differences have been eliminated, but there are still only three domains which have under a 20% difference. The comparison with the other Storm list has better results here, with the greatest difference only just exceeding 300% and every other comparison resulting in a under a 100% difference. Still, only a single comparison falls below 20%.

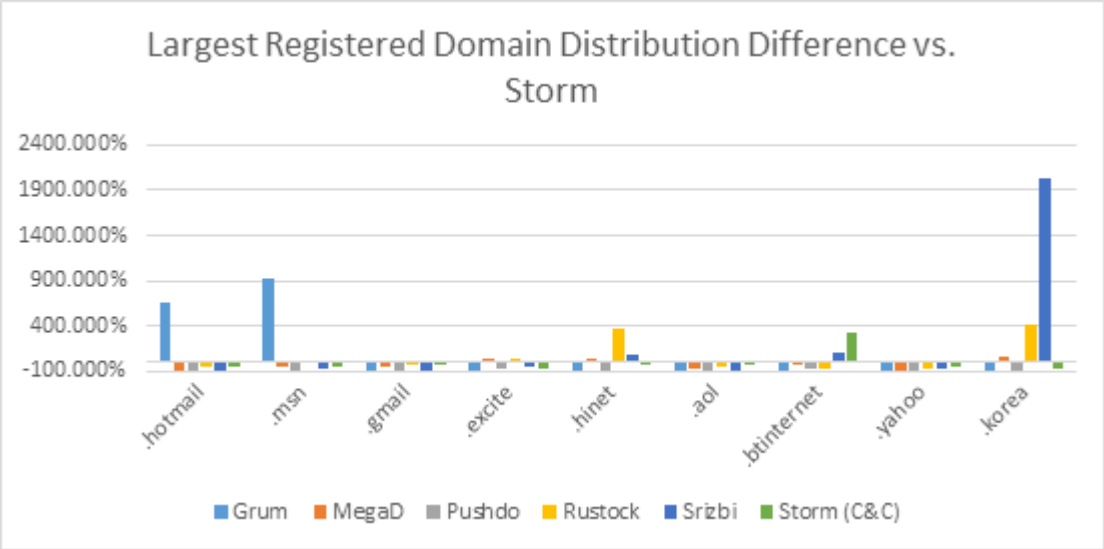


Figure 6.29: Registered Domain Distribution Percentage Difference for domains with greater than 1% of a list's addresses compared to Storm.

The final proportional difference comparison across lists is against the Storm (C&C) list. Like the previous Storm list, there are several large differences found which shroud the other results. The original comparison results can be seen in Figure 6.30. The differences between Srizbi and the Storm (C&C) lists are immediately apparent with several domains showing over a 5,000% difference all the way up to over a 32,000% difference.

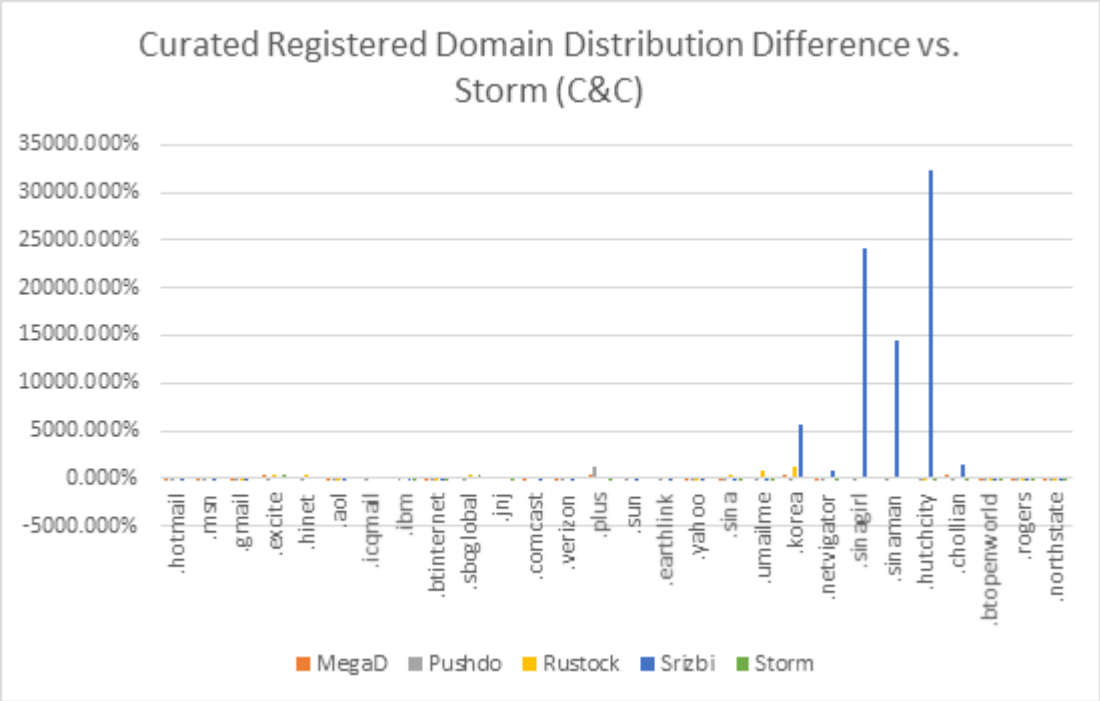


Figure 6.30: Registered Domain Distribution Percentage Difference for top-10 domains from all lists compared to Storm (C&C).

By removing these comparisons in Figure 6.31, the other results can be more clearly seen. Each other list have results which show large differences between the lists and Storm (C&C). There are only eight domains which have a result under 20% and only three which fall below a 10% difference. Only two similar comparisons occur between the two Storm lists, with the rest of the domains distinguishing the lists as unalike.

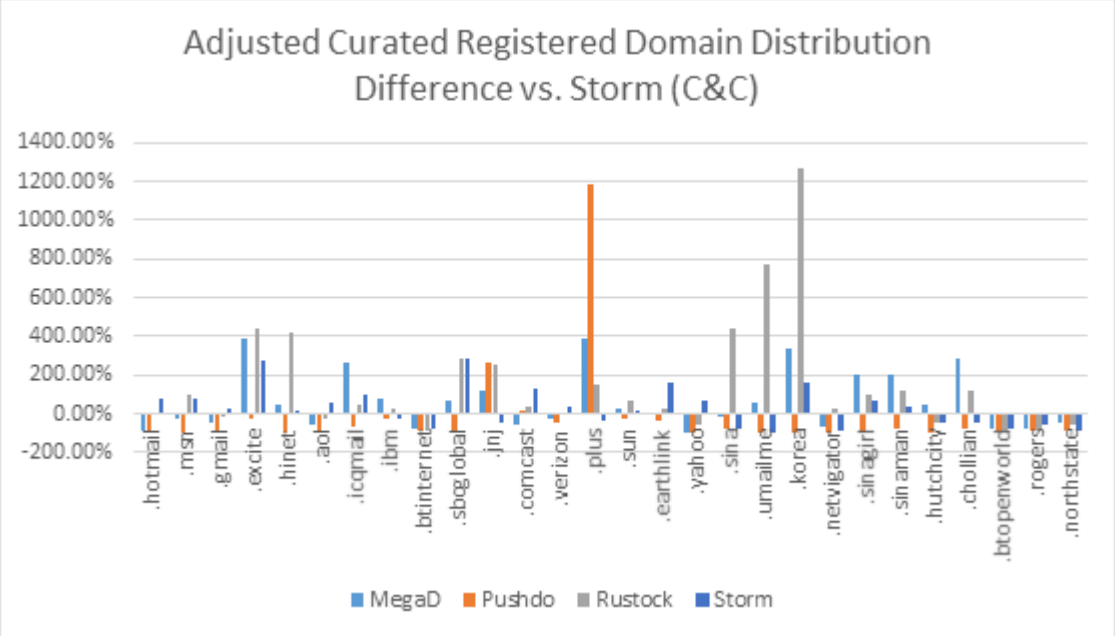


Figure 6.31: Registered Domain Distribution Percentage Difference for top-10 domains from all lists compared to Storm (C&C). The comparisons with the Srizbi list which exceed 5,000% have been removed to better show the other comparisons.

The results for the subset of largest registered domains also show few similarities between Storm (C&C) and the other lists. Only three comparisons fall below 20% with only one occurring with the other Storm list. The other results easily distinguish Storm (C&C) from the other lists.

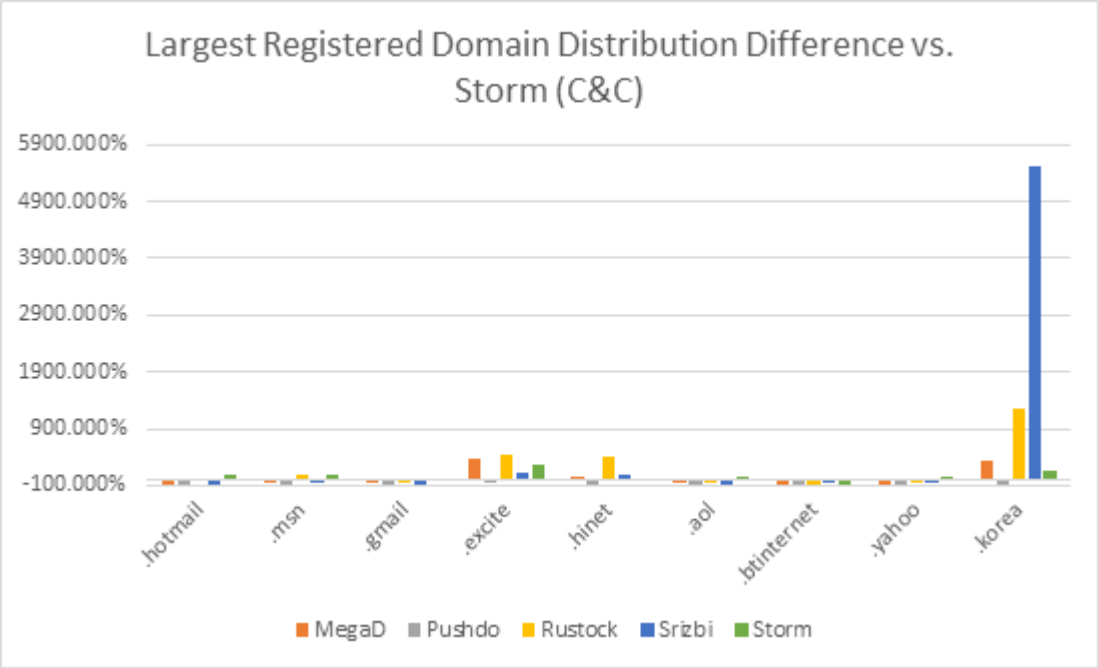


Figure 6.32: Registered Domain Distribution Percentage Difference for domains with greater than 1% of a list's addresses compared to Storm (C&C).

The results of the registered domain distribution analysis reveal two major observations. First, for both subsets of domains, there were very few similar results across all the comparisons. This shows that the registered domain distributions do distinguish between the different sources well. There are very few instances where two lists had several similar results, and in these cases there are often several drastic differences in the other comparisons to counteract these similarities. This is desired to distinguish between two address sets from different sources.

The second observation is similar but shows that the registered domains may not help build a classifier. As seen in the first observation, the comparisons were primarily different which is what is expected when comparing across two different sources. However, when comparing two different lists from the same source the results must be similar if a classifier can correctly identify that source for both lists. In the case of the two Storm lists, the comparisons produced mostly dissimilar results. Thus, while the registered domains successfully distinguished between lists from different sources, they

fail to correctly relate two distinct lists from the same source.

6.2 Sampling Domain Distribution

The second aspect of domain distribution based classification analysis is the ability to correctly identify if lists belong to the same source. This examination is performed using the domain distributions gathered through the process described in Section 4.3.

In these experiments, each unsorted raw list is divided into sub-lists. These sub-lists are comprised of different percentages of the addresses from the original list and are gathered consecutively. For example, the first sub-list is comprised of 50% of the addresses from the beginning of the original list. The second sub-list has 25% of the addresses from the original list and is taken immediately starting after the first sub-list ends. This is continued until sub-lists containing 12.5%, 6.25%, 3.125%, and 1.5625% of the original list's addresses are formed.

Each of these sub-lists are then sorted and run through the previously discussed domain distribution analysis routine. The resulting domain distributions for each sub-list are used to find their proportional differences with the original list.

The goal of these results is to see near identical distributions in the sub-lists to the original list. This will show these domain distributions can accurately classify different subsets from the same source.

6.2.1 Grum

The Grum list is an outlier from the other lists. As already mentioned, it only contains the *.com* top-level domain and two registered domains. Because it has so little variance, it is unlikely that this analysis will show many inconsistencies. The absence of a variety of values for both the top-level and country-code domains eliminates the need to include them in these comparisons.

The results of segmenting the Grum list are shown for the registered domain

distributions in Figure 6.33. There are only two domains available for analysis, and both make up a sizable portion of the list. Dividing the list into subsets shows that this caused little variance in the resulting subset distributions. This can be clearly seen below in the proportional difference charted in Figure 6.34.

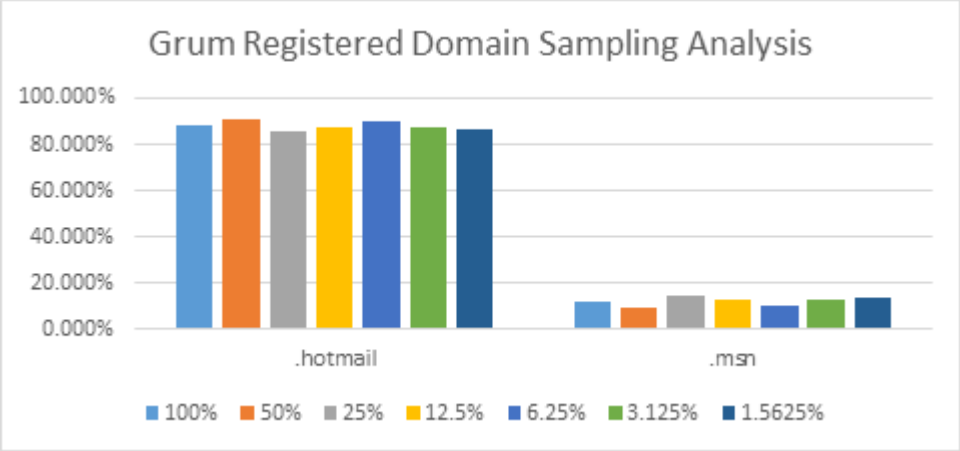


Figure 6.33: The percentage of addresses for each sampling subset of the Grum list containing a registered domain listed along the x-axis. The percentage of the original Grum list's addresses each subset contains is shown in the legend.

The subset comprised of 25% of the original Grum list's addresses has an approximately 25% difference for the *.msn* domain. This is the only comparison that differs by more than 20% with respect to the original list. Each Grum subset is relatively similar to the original list.

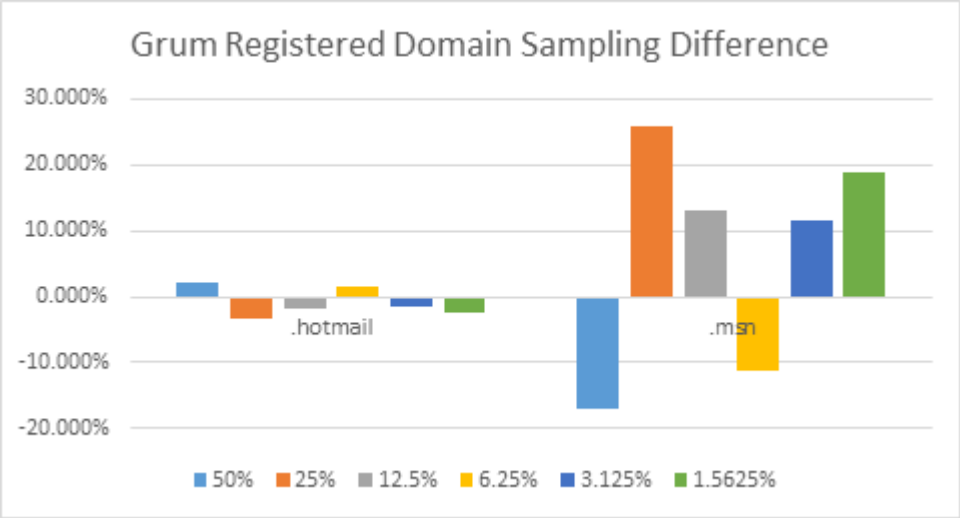


Figure 6.34: The percentage difference for all Grum's registered domains compared against the distributions of the entire list. The percentage of the original Grum list's addresses each subset contains is shown in the legend.

With the lack of variety found, it is unlikely that the original Grum list is an accurate representation of the addresses Grum routinely targets. Additionally, with only two data points available for comparison (the *.hotmail* and *.msn* domains) it is not surprising that little distinction is found in these results.

6.2.2 MegaD

The first set of domains analyzed for the MegaD list are the top-level domains. The distributions which result from the sampling segmentation of the original list are seen in Figure 6.35. As expected, the *.com*, *.net*, *.org*, and *.edu* domains are still the most prevalent top-level domains for all the subsets. At first glance below, while the *.com* domain shows some more variance in the distribution percentages, especially for the smaller subsets, the other three domains seem to be relatively consistent across the subsets.

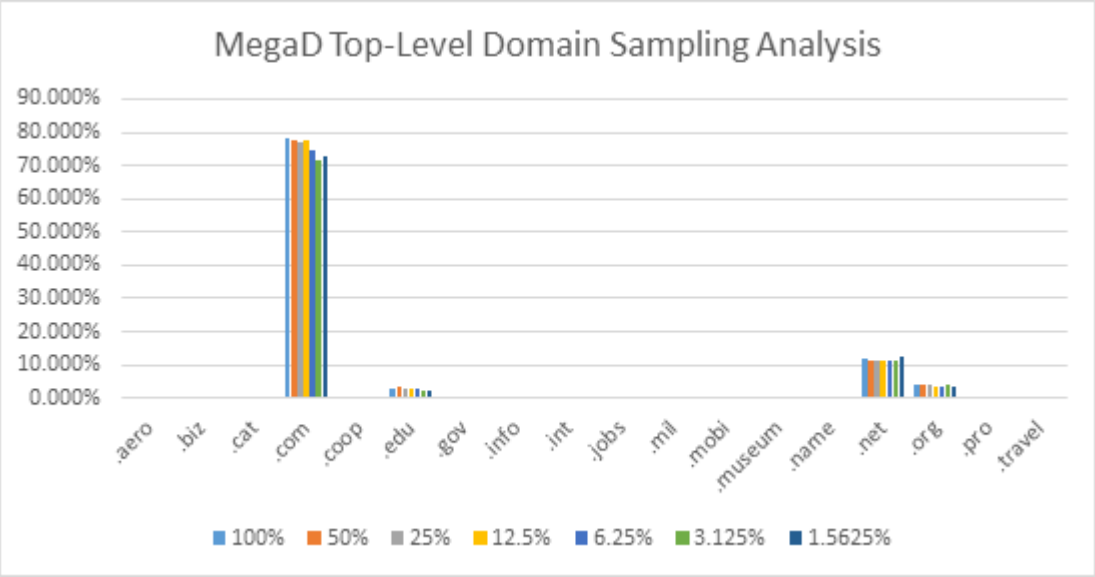


Figure 6.35: The percentage of addresses for each sampling subset of the MegaD list which end in each of the Top-Level domains listed along the x-axis. The percentage of the original MegaD list's addresses each subset contains is shown in the legend.

Figure 6.36 shows a much clearer view of the distribution changes between the subsets and the original list for the top-level domains. None of the comparisons exceed a 20% difference and many fall well below 10%. This shows that for the MegaD list, the top-level domain distributions seem to have relatively consistent values, at least with regards to these subsets.

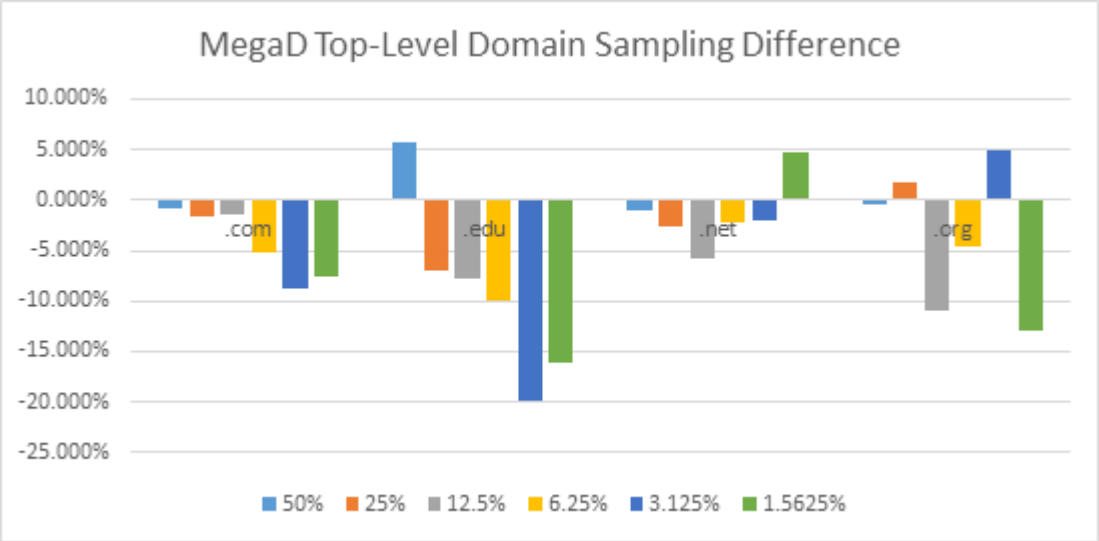


Figure 6.36: The percentage difference for MegaD's top-level domains compared against the distributions of the entire list. Only the four dominant domains are listed along the x-axis. The percentage of the original MegaD list's addresses each subset contains is shown in the legend.

The next set of distributions analyzed are for the country-code domains. For this analysis, the same curated subset of country-code domains from the cross-list domain distribution analysis are used. The distributions for the subsets are shown in Figure 6.37.

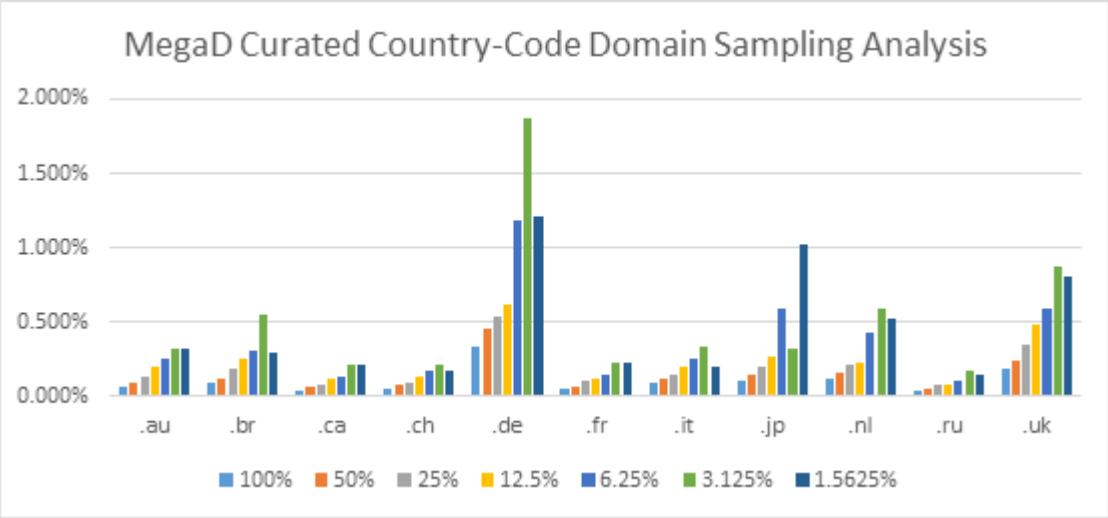


Figure 6.37: The percentage of addresses for each sampling subset of the MegaD list which end in each of the Country-Code domains listed along the x-axis. These domains are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original MegaD list's addresses each subset contains is shown in the legend.

It seems that the consistency found with the top-level domains is not present for the country-code domains. It is also clear that as the subsets get smaller, the percentage of addresses each country-code domain occurs in increases. Comparing the proportional difference for the domains of each subset against the original list shows substantial differences as seen in Figure 6.38. Every comparison exceeds 20% and from the 25% subset onwards they all exceed 50%. It is clear that the country-code domains are not a consistent identifier for MegaD.

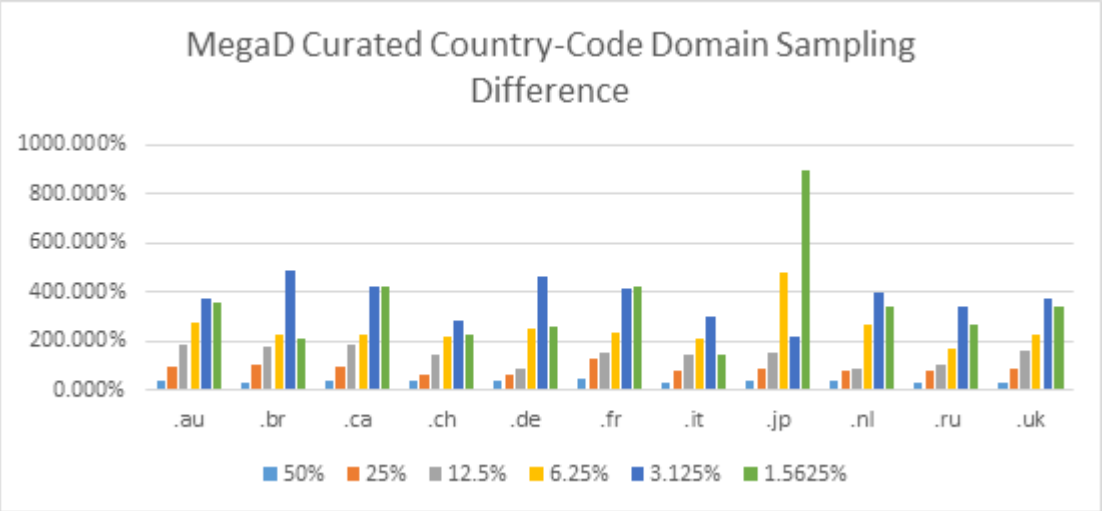


Figure 6.38: The percentage difference for MegaD's country-code domains compared against the distributions of the entire list. These domains along the x-axis are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original MegaD list's addresses each subset contains is shown in the legend.

The sampling distributions for the registered domains are shown in Figure 6.39. While it is clear that some variance is found in the subsets, the consistent patterns seen with the country-code domains are not present.

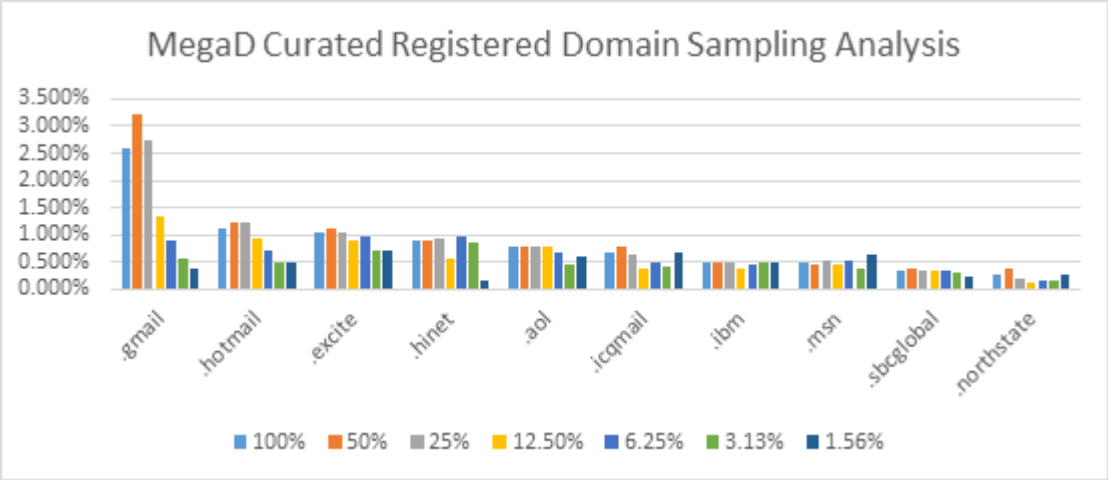


Figure 6.39: The percentage of addresses for each sampling subset of the MegaD list containing a registered domain listed along the x-axis. The domains are the top 10 most encountered registered domains for the MegaD list. The percentage of the original MegaD list's addresses each subset contains is shown in the legend.

The comparative differences found for the registered domains between the subsets and the original lists are seen in Figure 6.40. While many of the comparisons show similarities with the original list there are also several large distinctions.

For the 50% subset, seven of the ten comparisons are below a 10% difference with one domain remaining under a 20% difference and two exceeding 20%. Similarly, the 25% subset has 90% of its comparisons fall below 10%. Both these subsets show a strong ability to use the top-level domain as an identifier.

As the subsets get smaller, this identifier becomes weaker. At 12.5%, three domains have a difference between 10% and 20% and four exceed a 20% difference from the original list. Similarly, at 6.25% only half the domains fall below a 10% difference. And for the 3.125% and 1.5625% subsets only 20% and 30% of the comparisons fall below 10% respectively with 60% and 70% of the comparisons exceeding a 20% difference.

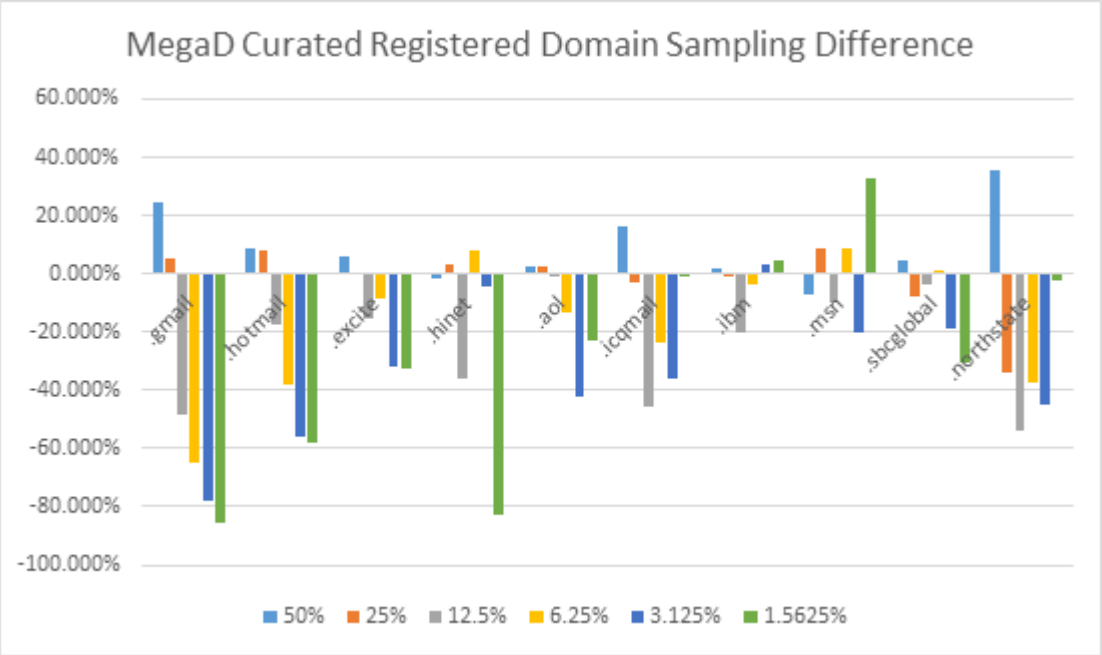


Figure 6.40: The percentage difference for MegaD's top 10 most encountered registered domains compared against the distributions of the entire list. The percentage of the original MegaD list's addresses each subset contains is shown in the legend.

This shows that the ability to use the registered domain distributions to identify MegaD as a source starts to break down with between 25% and 12.5% of the addresses in the original list. If the fluctuation point is assumed to be half way between these two percentages at 18.75%, that would require over 33 million addresses to form a distribution with a likelihood of positive identification based on the number of valid addresses in the original MegaD list.

6.2.3 Pushdo

The Pushdo lists top-level domain distributions for the sampling subsets (Figure 6.41) show the same behavior with the four major domains dominating the list for all the distributions. Like MegaD, while minor variances can be seen in each of these four domains, only the *.com* domain shows anything of any relative size.

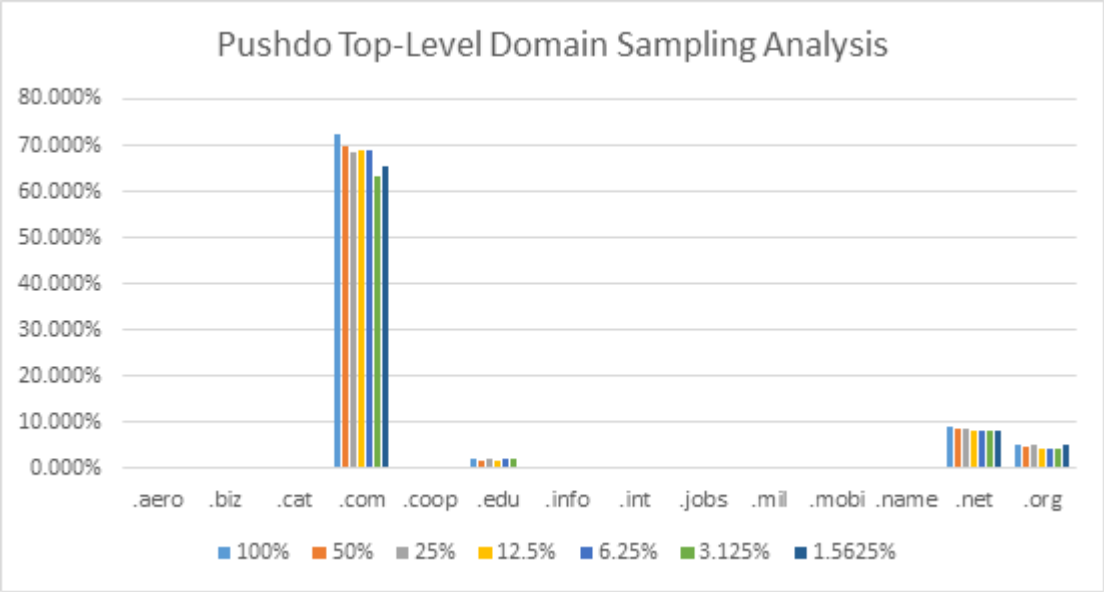


Figure 6.41: The percentage of addresses for each sampling subset of the Pushdo list which end in each of the Top-Level domains listed along the x-axis. The percentage of the original Pushdo list's addresses each subset contains is shown in the legend.

When reviewing the proportional differences for these four domains (see Figure 6.42), it becomes clear that there are very few differences for the subsets. Only one comparison exceeds a 20% difference and starting at the 12.5% subset only eight exceed 10% (but still fall below 20%). This shows that for the Pushdo list, the top-level domain distributions seem to have relatively consistent values, at least with regards to these subsets.

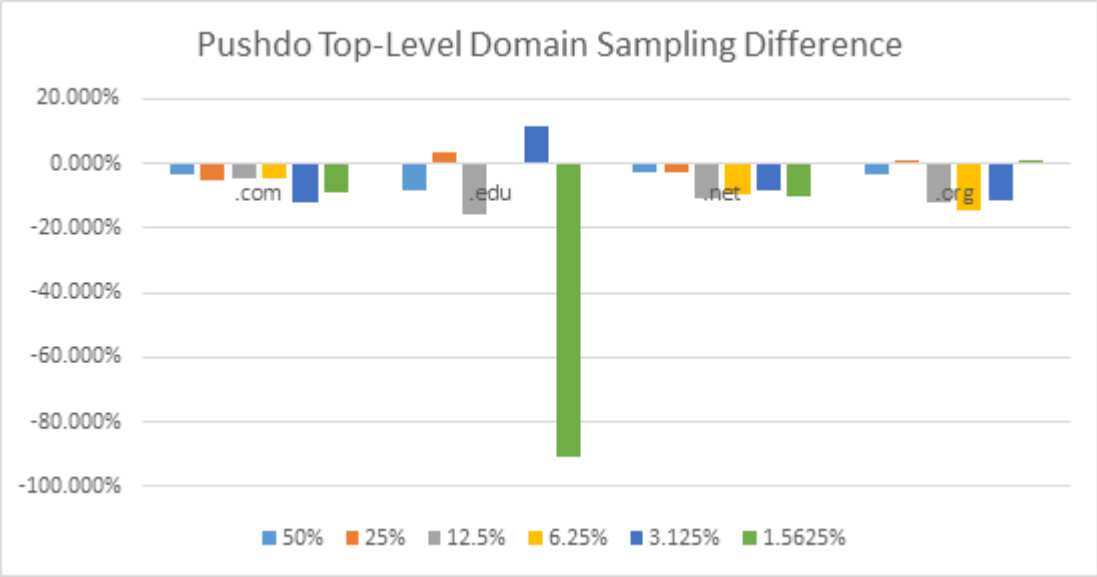


Figure 6.42: The percentage difference for Pushdo's top-level domains compared against the distributions of the entire list. Only the four dominant domains are listed along the x-axis. The percentage of the original Pushdo list's addresses each subset contains is shown in the legend.

With Pushdo's country-code domains, a similar pattern as encountered with MegaD's sampling distributions is seen. Generally, it seems that the smaller the subset, the higher the percentage rate each country domain occurs within the list. However, this pattern is not as strict as the one seen with MegaD and the progression is less linear in nature.

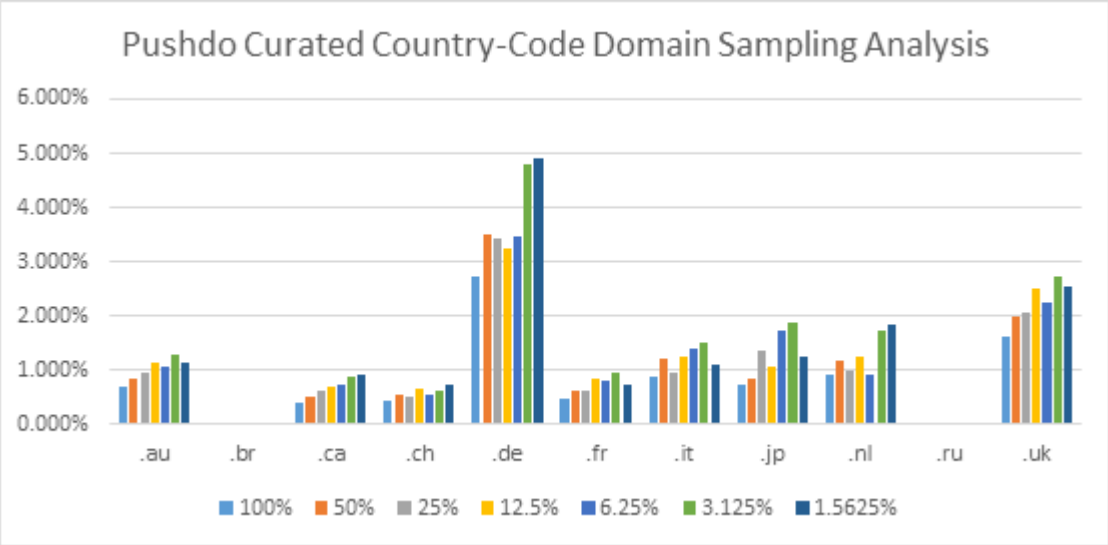


Figure 6.43: The percentage of addresses for each sampling subset of the Pushdo list which end in each of the Country-Code domains listed along the x-axis. These domains are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Pushdo list's addresses each subset contains is shown in the legend.

The percentage differences between the subsets and the original Pushdo list are shown in Figure 6.44. The majority of the comparisons are above a 20% difference showing very few similarities exist between the original list's distributions and the subsets' for the country-code domains. This shows that these domains do not make viable candidates for correctly identifying the original source.

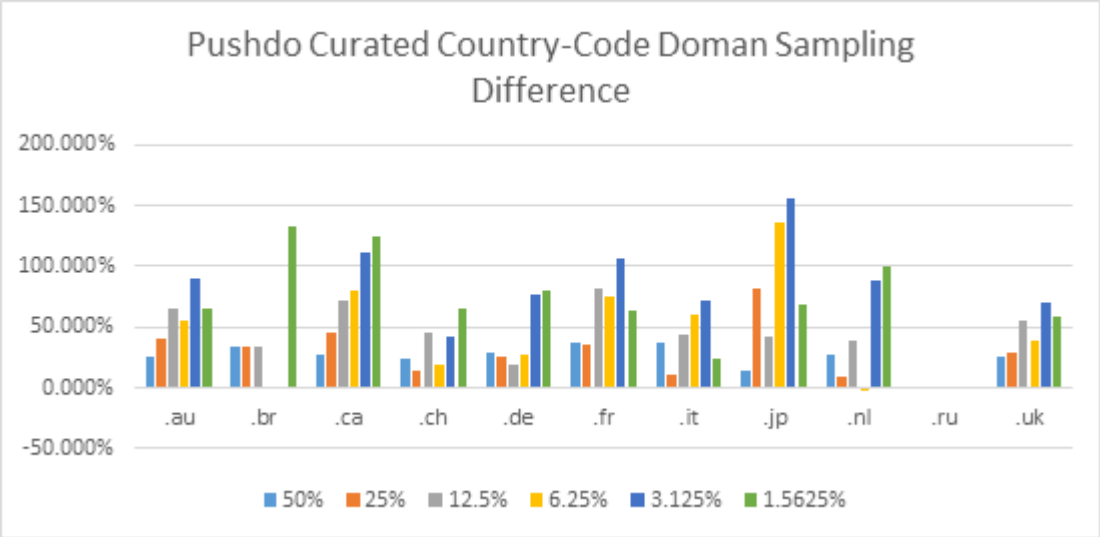


Figure 6.44: The percentage difference for Pushdo's country-code domains compared against the distributions of the entire list. These domains along the x-axis are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Pushdo list's addresses each subset contains is shown in the legend.

The distributions for the registered domains for the Pushdo list and its sampling subsets can be seen in Figure 6.45. The two largest domains (*.gmail* and *.jnj*) show a fair amount of fluctuation in their distributions while the other domains seem to be more stable across their subsets.

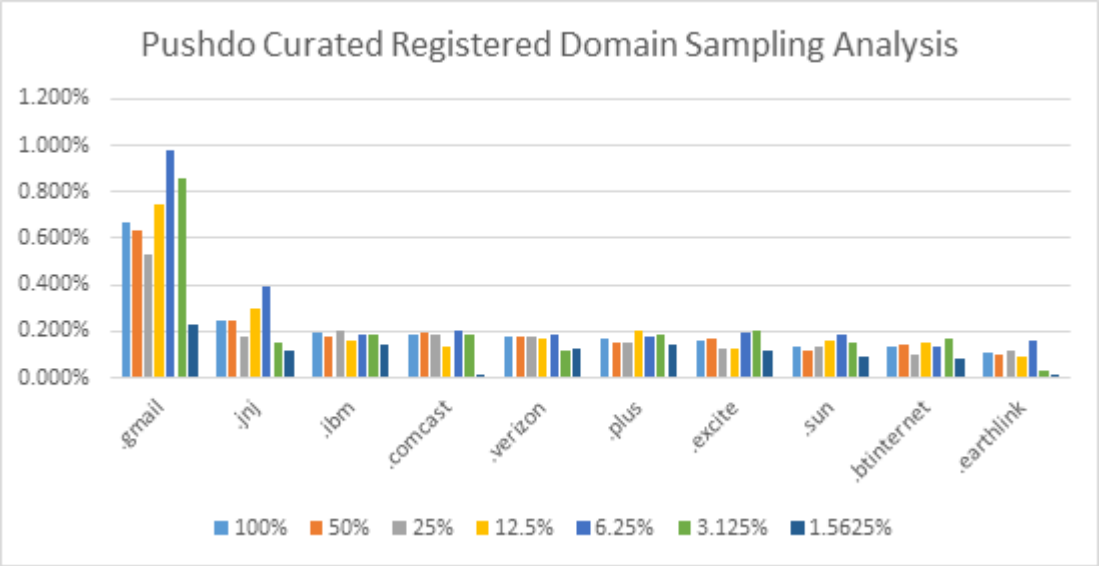


Figure 6.45: The percentage of addresses for each sampling subset of the Pushdo list containing a registered domain listed along the x-axis. The domains are the top 10 most encountered registered domains for the Pushdo list. The percentage of the original Pushdo list's addresses each subset contains is shown in the legend.

The sampling differences for the registered domain as seen in Figure 6.46 show several larger differences. The first subset (with 50% of the original list's addresses) is similar to the original list with eight of its ten domains having less than a 10% difference and the remaining two domains falling below 20%.

After the 50% subset, things become less clear. The 25% subset only has four domains with under a 10% difference. Three domains are under 20% while the final three exceed 20%. For the 12.5% subset, only one domain has below a 10% difference whereas seven domains have less than a 20% difference. This shows that while the distribution for few of the domains for the 12.5% subset are distinctly different from the original list, few are also clearly identifiable as being the similar to the original list.

From 6.25% and smaller it becomes clear that the subsets' distributions are very dissimilar from the original list's. From 6.25% to 1.5625% the subsets have 50%, 60%, and 90% of their domains respectively exceeding a 20% difference.

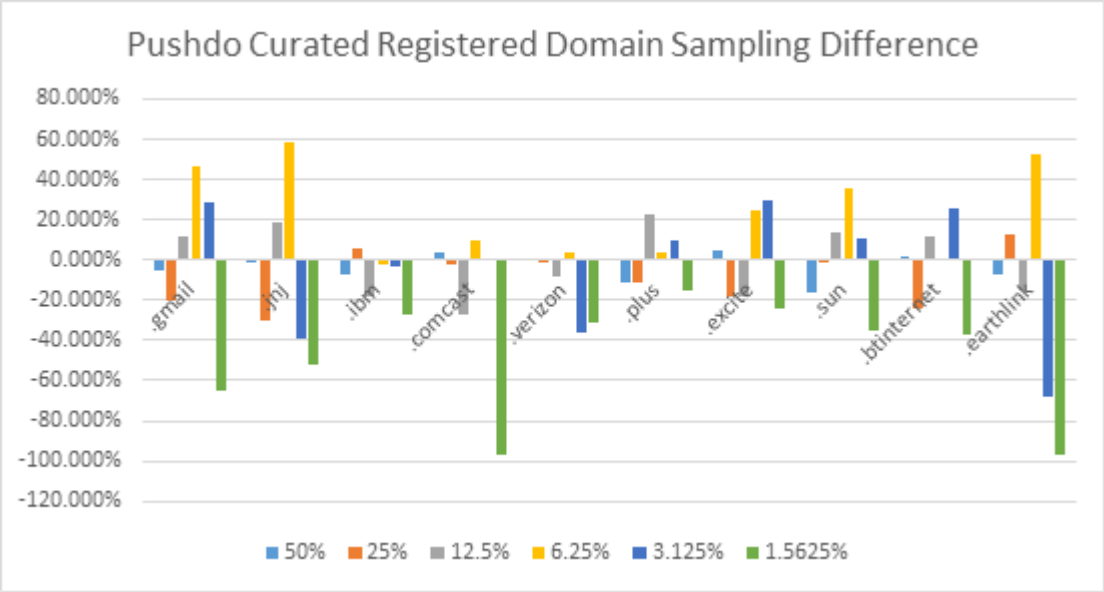


Figure 6.46: The percentage difference for Pushdo's top 10 most encountered registered domains compared against the distributions of the entire list. The percentage of the original Pushdo list's addresses each subset contains is shown in the legend.

Using the registered domain distribution to classify a set of addresses as sharing Pushdo as a source is viable with the 50% subset. The 25% subset is questionable with regards to its overall success across a variety of sources and the 12.5% subset is even more so. Based on the size of the original list, over 800,000 addresses are required for the questionable identification provided by the 12.5% subset and over 1.7 million are required to reach the confidence provided by the 25% subset.

6.2.4 Rustock

The Rustock distributions for the top-level domains show slightly more variance than the other lists reviewed already. As seen below in Figure 6.47, the four primary top-level domains still dominate the distribution. However, many of the other domains are just visible in this graph for the Rustock list. The *.net* domain also shows some of the divergences previously only seen in the *.com* domain across the samplings.

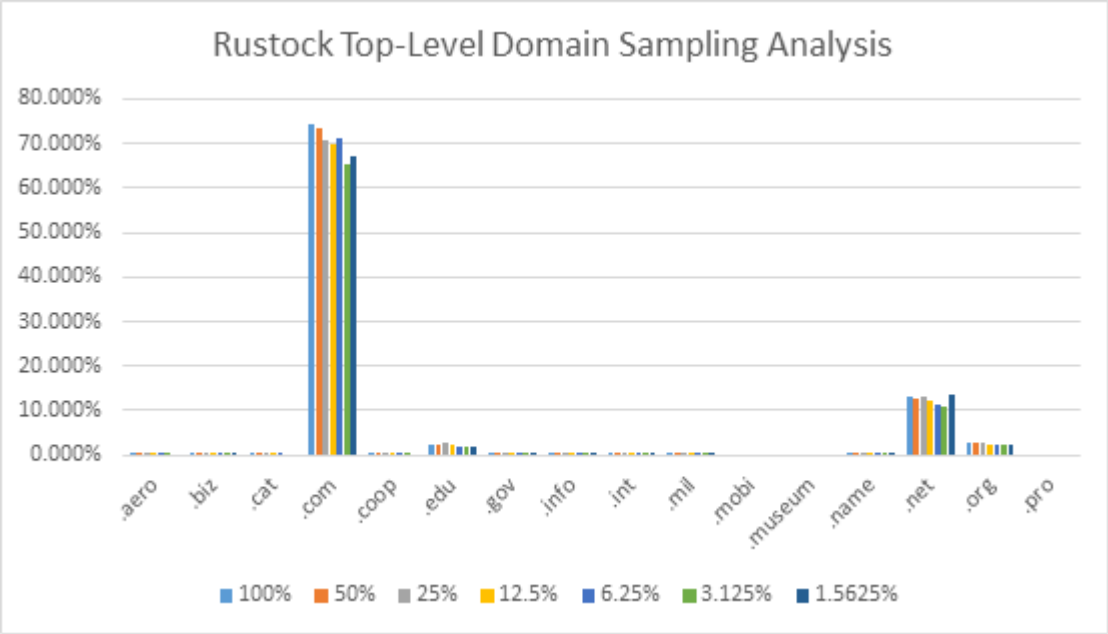


Figure 6.47: The percentage of addresses for each sampling subset of the Rustock list which end in each of the Top-Level domains listed along the x-axis. The percentage of the original Rustock list's addresses each subset contains is shown in the legend.

Figure 6.48 shows the proportional differences for the samplings of the top-level domains. Only a single comparison exceeds a 20% difference and over half the comparisons fall below a 10% difference. These distributions show that the top-level domains remain fairly consistent for these subsets.

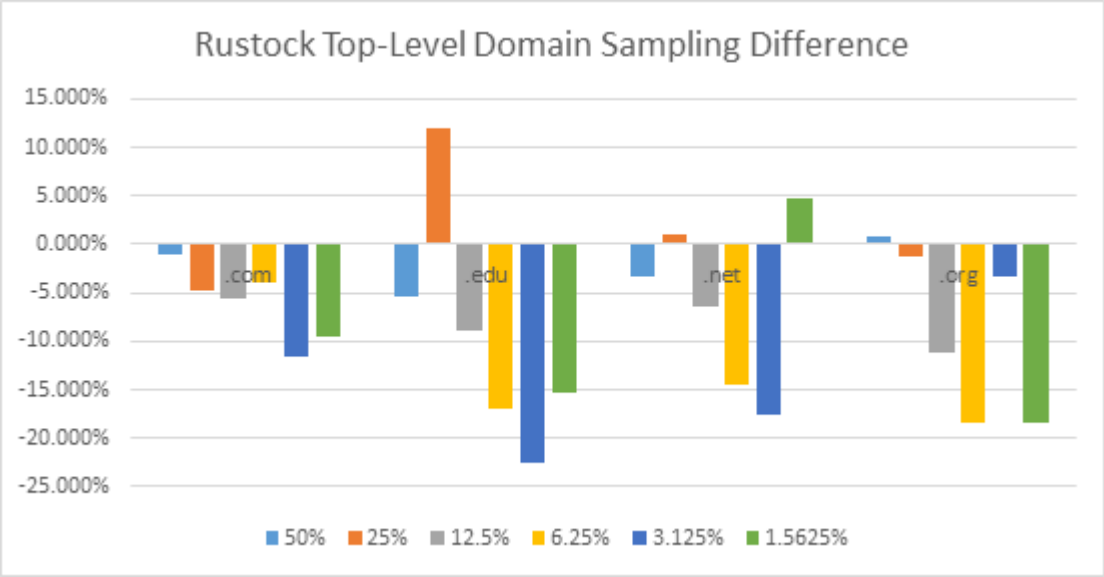


Figure 6.48: The percentage difference for Rustock's top-level domains compared against the distributions of the entire list. Only the four dominant domains are listed along the x-axis. The percentage of the original Rustock list's addresses each subset contains is shown in the legend.

The country-code domains again show an increasing trend for their percentage distribution as the sampling size decreases. This trend is seen in Figure 6.49.

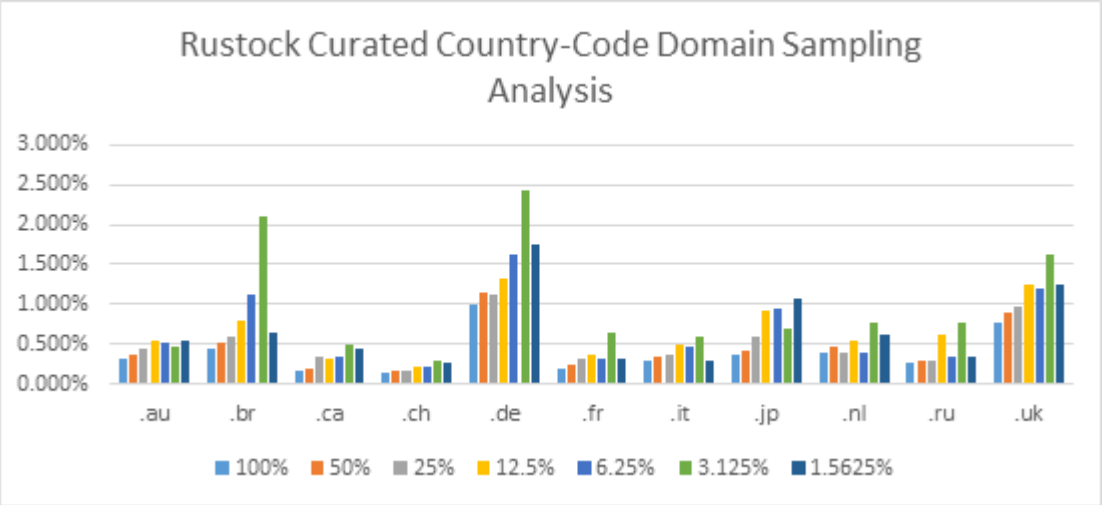


Figure 6.49: The percentage of addresses for each sampling subset of the Rustock list which end in each of the Country-Code domains listed along the x-axis. These domains are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Rustock list's addresses each subset contains is shown in the legend.

As expected with this increasing trend, the proportional comparisons shown in Figure 6.50 have several significant differences. Only three of the comparisons fall below a 10% difference and only an additional ten out of the 66 comparisons fall below 20%. Of these ten, all of them occur within the 50% and 25% subsets. It is clear that the country-code domains cannot be used to correctly identify the Rustock source.

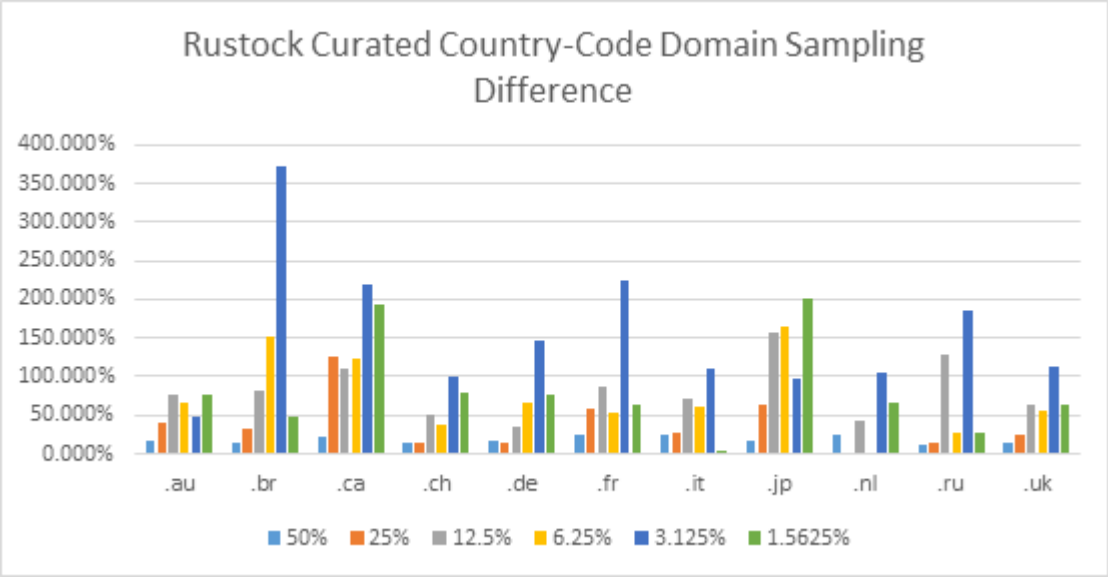


Figure 6.50: The percentage difference for Rustock's country-code domains compared against the distributions of the entire list. These domains along the x-axis are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Rustock list's addresses each subset contains is shown in the legend.

The final sampling analysis for Rustock is against the registered domains. The distributions for each sampling subset are shown in Figure 6.51. In this chart, it appears that the largest differences occur in the three most prominent domains (*.hotmail*, *.gmail*, and *.hinet*).

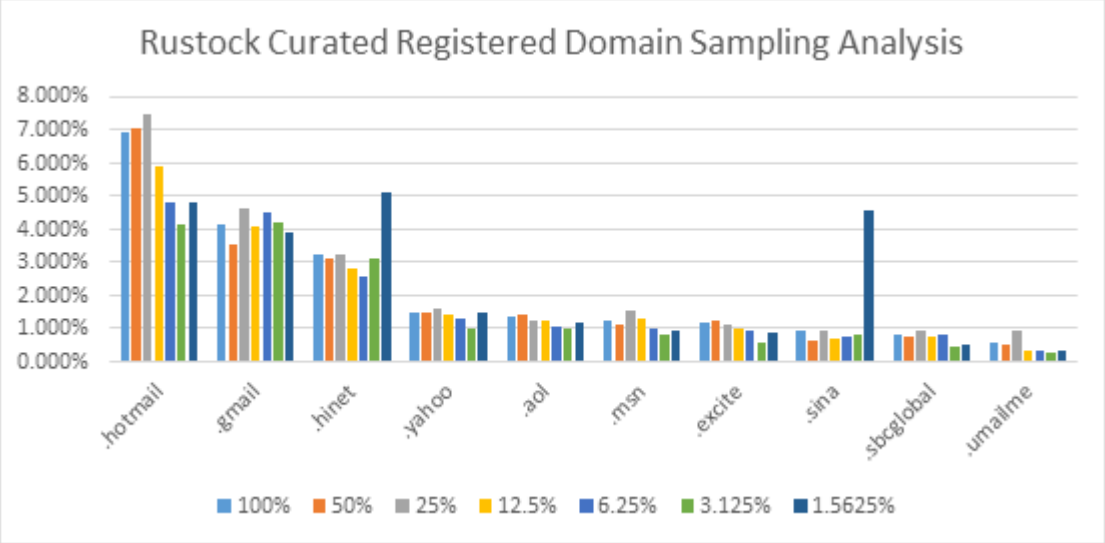


Figure 6.51: The percentage of addresses for each sampling subset of the Rustock list containing a registered domain listed along the x-axis. The domains are the top 10 most encountered registered domains for the Rustock list. The percentage of the original Rustock list's addresses each subset contains is shown in the legend.

The proportional differences for Rustock's registered domain samplings are shown in Figure 6.52. The first two subsets are similar with six of the ten comparisons resulting in under a 10% difference each. However, starting with the third subset, greater differences begin to appear. With the 12.5% sampling, four of the comparisons exceed a 10% difference and another two exceed 20%. From 6.25% and lower, only two domains for each subset have less than a 10% difference. For 6.25% only four of the ten domains exceed a 20% difference; whereas for the 3.125% and 1.5625% samplings seven of the domains exceed this threshold.

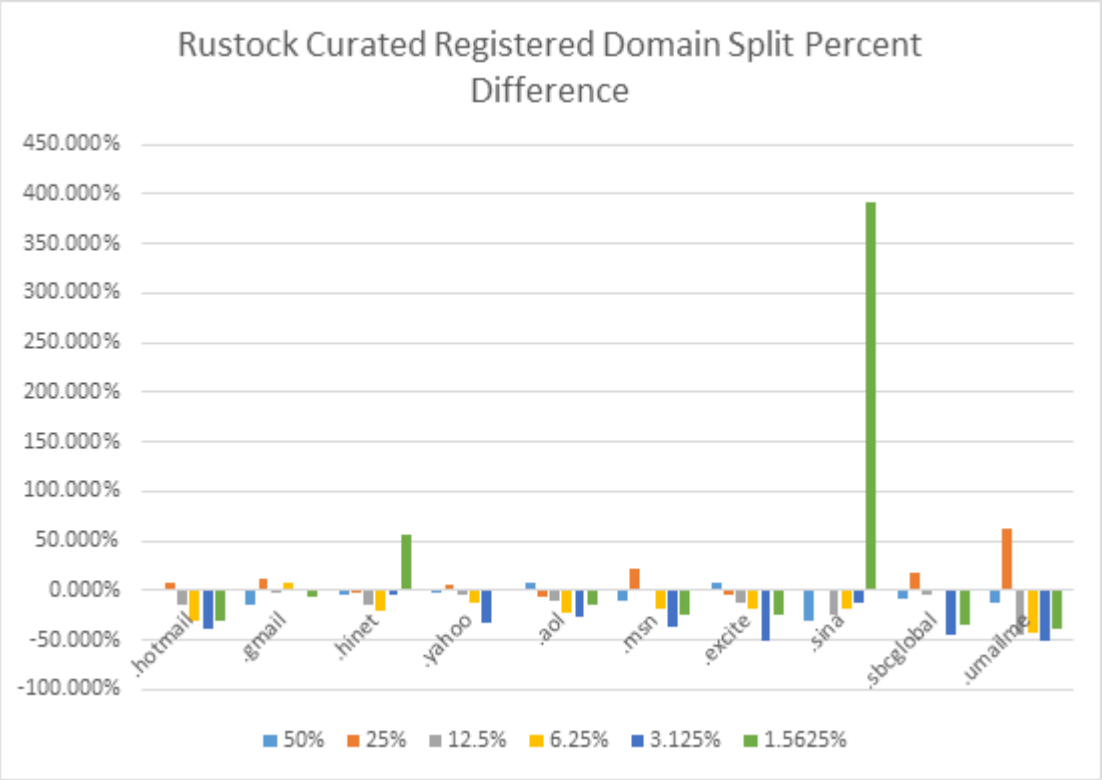


Figure 6.52: The percentage difference for Rustock's top 10 most encountered registered domains compared against the distributions of the entire list. The percentage of the original Rustock list's addresses each subset contains is shown in the legend.

The confidence in the ability to classify the Rustock source from a list's registered domain distributions is questionable at best from the 12.5% sampling and below. To have the confidence in the classification provided by the 25% subset, over 1.7 million addresses are required to generate the distribution. For the slightly less accurate results provided by the 12.5% subset, almost 900,000 addresses are needed.

6.2.5 Srizbi

The top-level domain distributions across the sampling subsets for Srizbi show similar behaviors to the lists previously analyzed. As shown in Figure 6.53, the top four domains still lead the distributions and the *.com* domain shows the most divergence across its subsets.

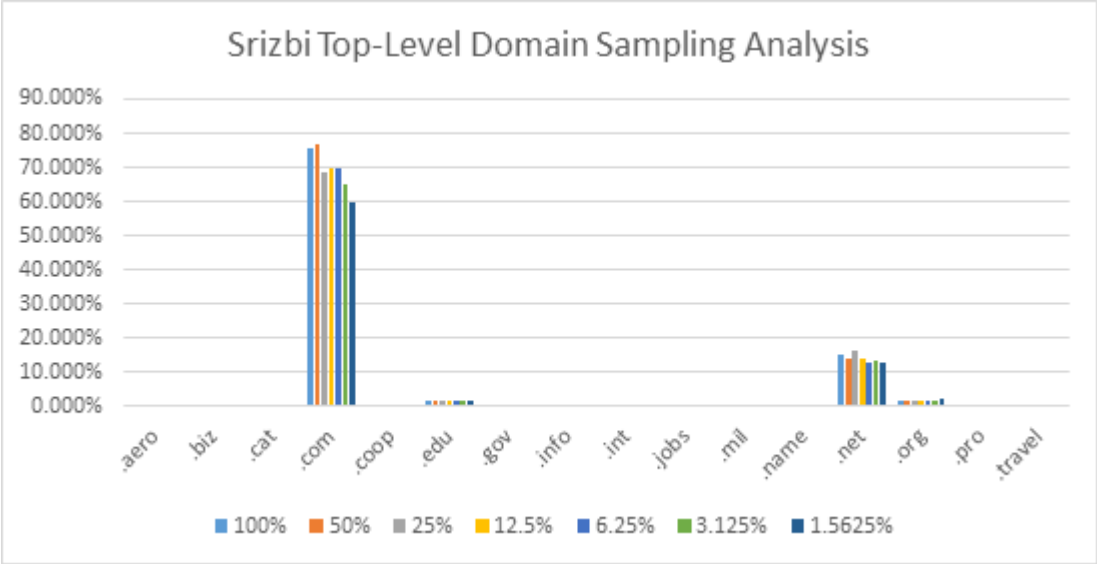


Figure 6.53: The percentage of addresses for each sampling subset of the Srizbi list which end in each of the Top-Level domains listed along the x-axis. The percentage of the original Srizbi list's addresses each subset contains is shown in the legend.

The proportional differences calculated for the top-level domain sampling subsets are shown in Figure 6.54. The first three subsets show few differences with only two comparisons exceeding a 10% difference by only a minor amount. The three smallest subsets do not fare as well with two of the four domains exceeding a 10% difference each for the 6.25% and 3.125% subsets (one of these larger differences for 6.25% barely exceeds a 20% difference). Only a single comparison falls below a 10% difference for the smallest subset, which also has one comparison exceed a 20% difference.

While there are no dramatic differences for the top-level domain distributions. At 6.25% of the list and below, there are enough aberrations to lower the confidence of a classifier.

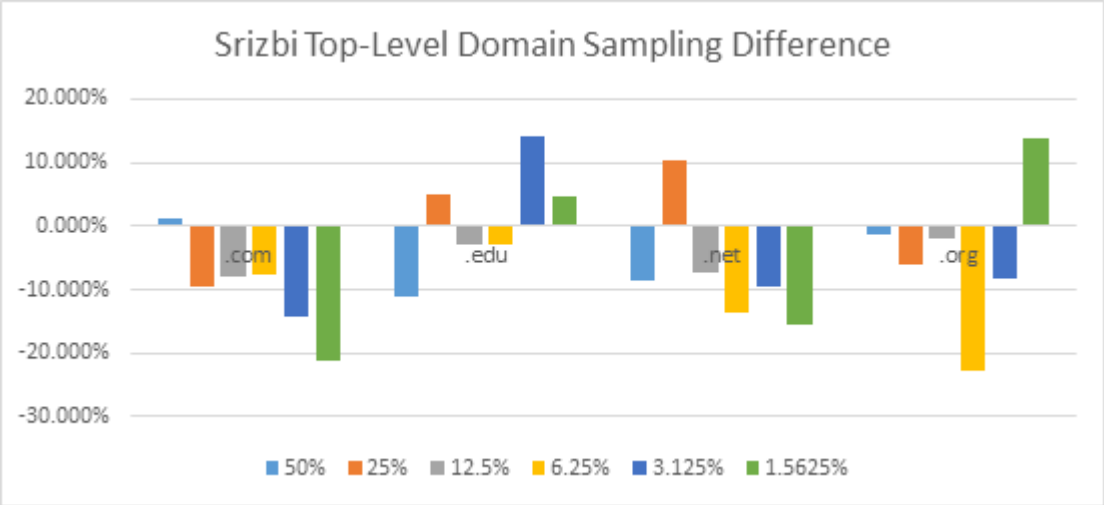


Figure 6.54: The percentage difference for Srizbi's top-level domains compared against the distributions of the entire list. Only the four dominant domains are listed along the x-axis. The percentage of the original Srizbi list's addresses each subset contains is shown in the legend.

The trend of increasing country-code domain distributions with the sampling process continues for the Srizbi list as shown in Figure 6.55. With large differences easily seen between the full list and its sampling subsets, specifically for the *.de* and *.uk* domains, it is already unlikely that the country domains will make viable classifiers.

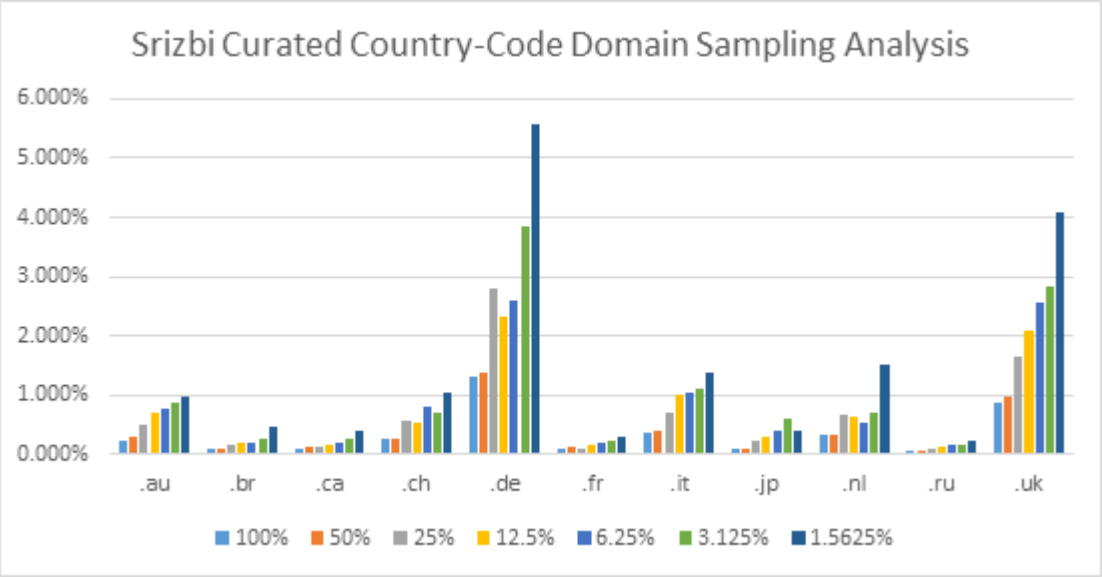


Figure 6.55: The percentage of addresses for each sampling subset of the Srizbi list which end in each of the Country-Code domains listed along the x-axis. These domains are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Srizbi list's addresses each subset contains is shown in the legend.

This observation is reinforced with the percentage differences between the subset distributions and the full list shown in Figure 6.56. Only the 50% subset has any comparisons with under a 20% difference (it has five below 10%, three below 20%, and three exceeding 20%). Most of the comparisons exceed a 100% difference and one exceeds 500% clearly indicating that the country-code domains are unsuitable for use in a classifier for Srizbi.

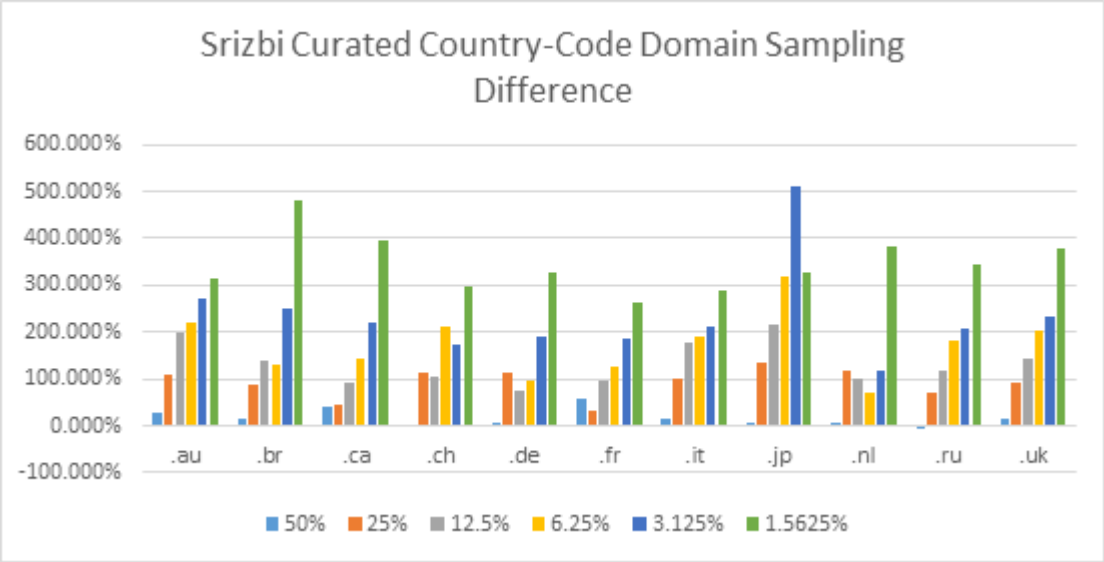


Figure 6.56: The percentage difference for Srizbi's country-code domains compared against the distributions of the entire list. These domains along the x-axis are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Srizbi list's addresses each subset contains is shown in the legend.

The registered domain sampling analysis for Srizbi begins with the subset distributions shown in Figure 6.57. The Srizbi distributions seem to show more variance than the other lists analyzed so far. This variance also does not seem to be limited to only the one or two top domains but instead affects each domain below.

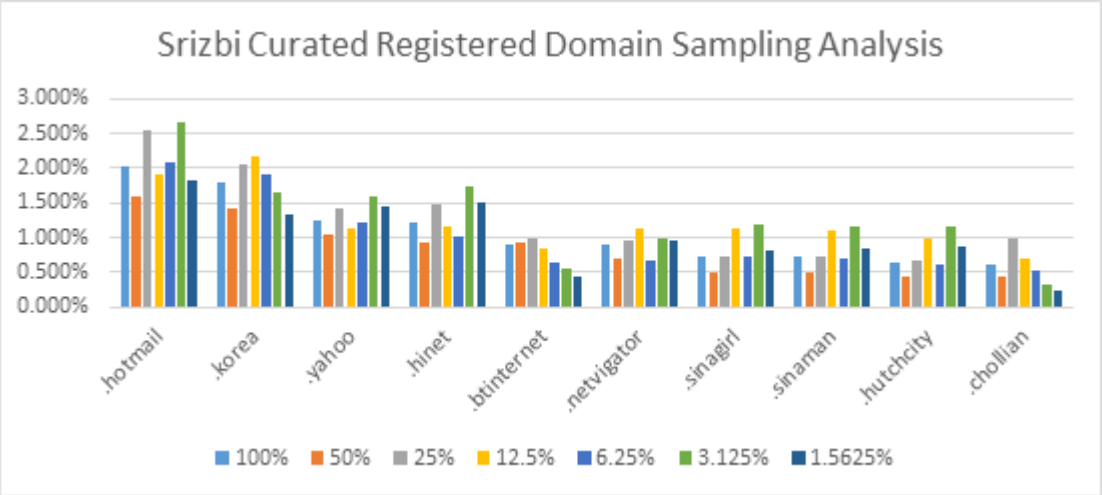


Figure 6.57: The percentage of addresses for each sampling subset of the Srizbi list containing a registered domain listed along the x-axis. The domains are the top 10 most encountered registered domains for the Srizbi list. The percentage of the original Srizbi list's addresses each subset contains is shown in the legend.

The proportional differences for Srizbi's registered domains are graphed in Figure 6.58. Unlike the other lists analyzed, the similarity of the distributions to the original list is not associated with the size of the subset. For instance, the largest subset only has one comparison result with less than a 10% difference and one under 20%. The other eight domains all exceed a 20% difference, although none surpass a 35% difference. Conversely, the next subset (25%) has five domains with under a 10% difference, two below a 20% difference, and only three which exceed 20%.

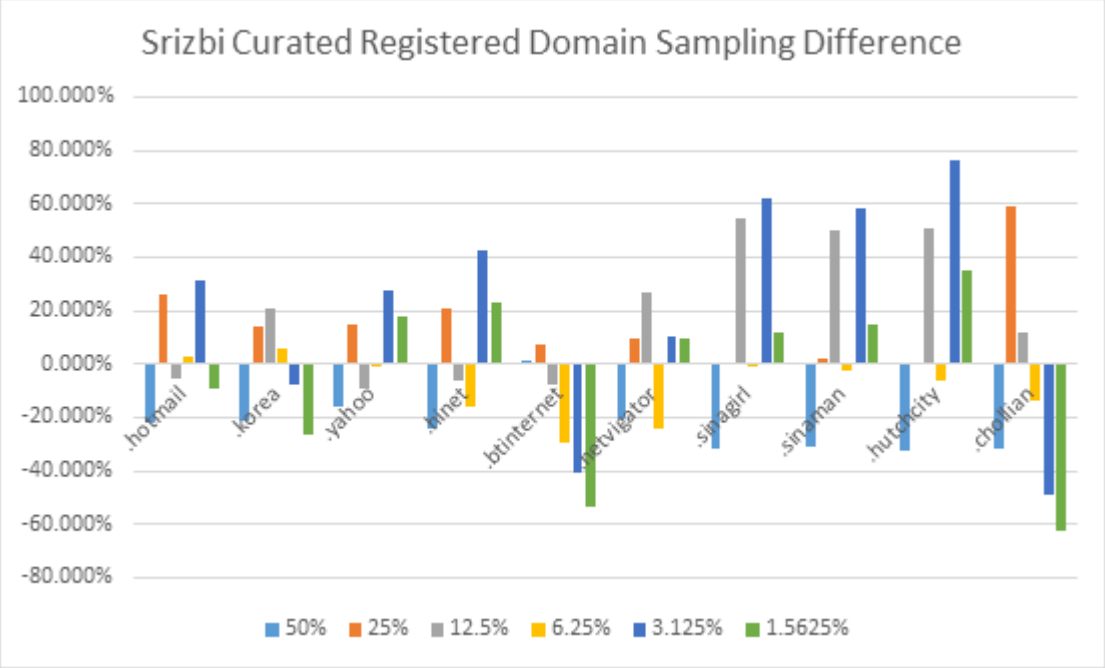


Figure 6.58: The percentage difference for Srizbis top 10 most encountered registered domains compared against the distributions of the entire list. The percentage of the original Srizbi list's addresses each subset contains is shown in the legend.

The unpredictable nature of the results for Srizbi may have been caused by the original raw list had a form of partial or periodic sorting. Nevertheless, this irregularity shows that the registered domain distributions will not correctly identify the Srizbi list. This is reinforced as the subset with half the original list's addresses shows consistent differences with most of its domains.

6.2.6 Storm

The first Storm list's sampling distributions for the top-level domains are shown in Figure 6.59. As already shown in Section 6.1.1, the Storm lists' top level domain distributions did not appear to provide strong candidates for a classifier. Reviewing the results in the figure below, the four major top-level domains are still dominant. Additionally, it appears the most divergences occur with the *.com* domain.

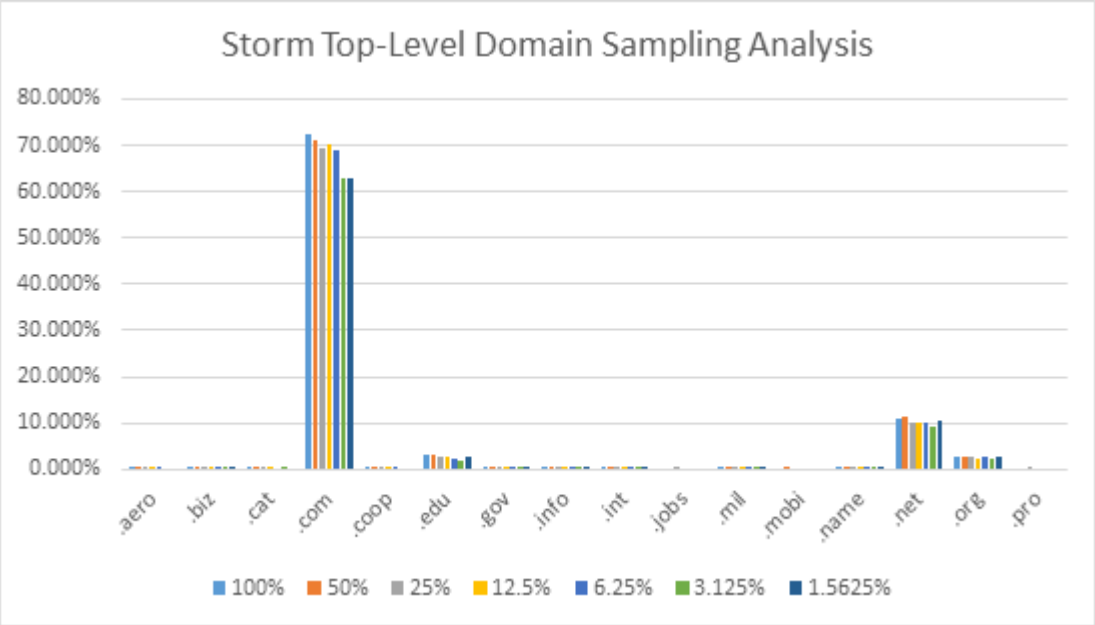


Figure 6.59: The percentage of addresses for each sampling subset of the Storm list which end in each of the Top-Level domains listed along the x-axis. The percentage of the original Storm list's addresses each subset contains is shown in the legend.

The proportional differences for the first Storm's top-level sampling differences are shown in Figure 6.60. The first three subsets' distributions are very similar to the original list. Only the *.org* domain for the smallest of these three subsets exceeds a 10% difference, and it only exceeds that threshold by less than a percentage point.

Larger differences begin to appear after these initial subsets. The *.edu* domain exceeds a 20% difference for both the 6.25% and 3.125% subsets. Additionally, for the 3.125% subset, the *.com*, *.net*, and *.org* domains all show a difference between 10% and 20%. Surprisingly, the smallest sampling subset only has a single comparison which exceeds a 10% difference.

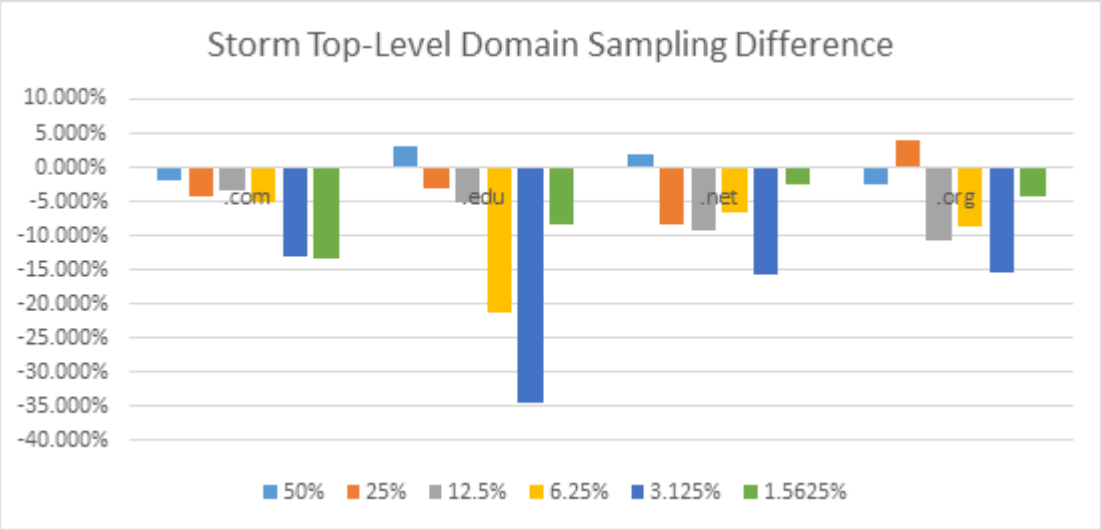


Figure 6.60: The percentage difference for Storm's top-level domains compared against the distributions of the entire list. Only the four dominant domains are listed along the x-axis. The percentage of the original Storm list's addresses each subset contains is shown in the legend.

The first Storm list shows that the top-level domains can be relatively similar, especially in the large subsets. However, the differences in the smaller subsets, shown specifically in the 3.125% group, show that these domains are not always a reliable indicator.

The Storm list's country code domain samplings continue the same increasing trend in Figure 6.61 as seen previously with the other lists. It is already clear there are several large differences in the subsets with the full list, particularly in the smaller samplings.

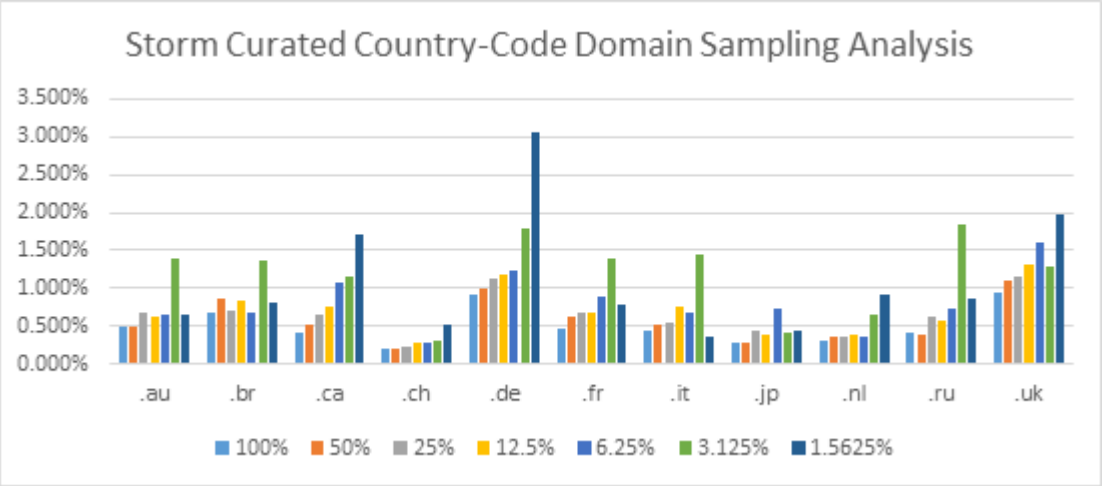


Figure 6.61: The percentage of addresses for each sampling subset of the Storm list which end in each of the Country-Code domains listed along the x-axis. These domains are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Storm list's addresses each subset contains is shown in the legend.

The percentage differences between the subsets and the original country-code domain distributions for the first Storm list are shown in Figure 6.62. The results should make it clear that the country-code domains diverge significantly from the original distribution in each of the subsets and will not be useful for any classifier. Only 50% of the domains in the largest subset have under a 10% difference while 30% of its domains approach a 30% difference.

In the following five subsets, only two out of the 55 comparisons have a result below 10% and only five fall between 10% and 20%. Starting with the 6.25% subset, several comparisons exceed a 100% difference culminating in a 352% difference for the *.ru* domain with the 3.125% sampling.

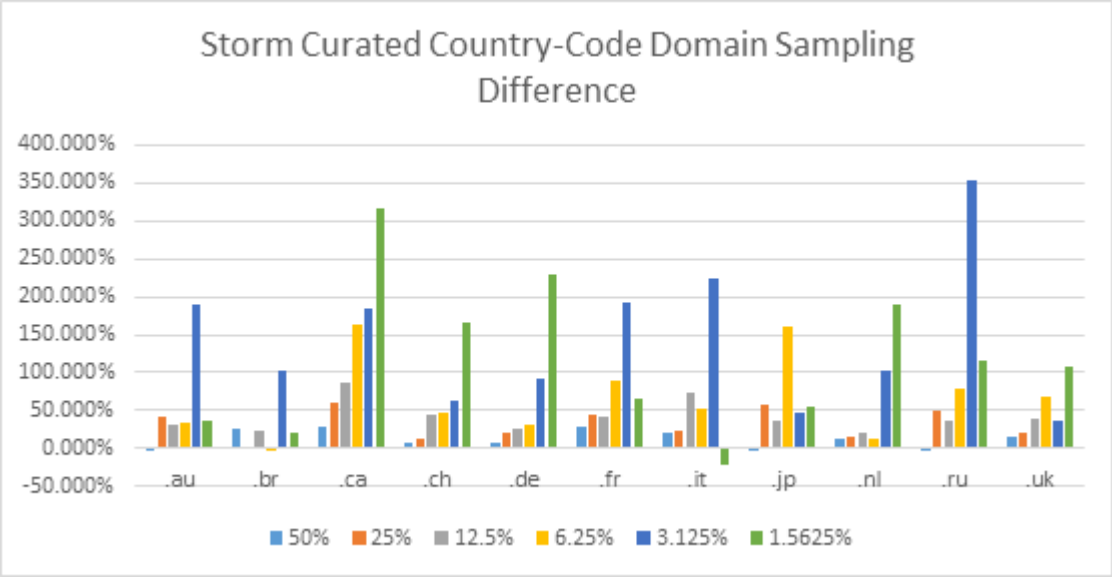


Figure 6.62: The percentage difference for Storm's country-code domains compared against the distributions of the entire list. These domains along the x-axis are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Storm list's addresses each subset contains is shown in the legend.

The first Storm list's registered domain sampling distributions are shown in Figure 6.63. The largest domain (*.hotmail*) initially seems to have the most differences. Differences are also readily apparent with the following three domains (*.yahoo*, *.gmail*, and *.aol*).

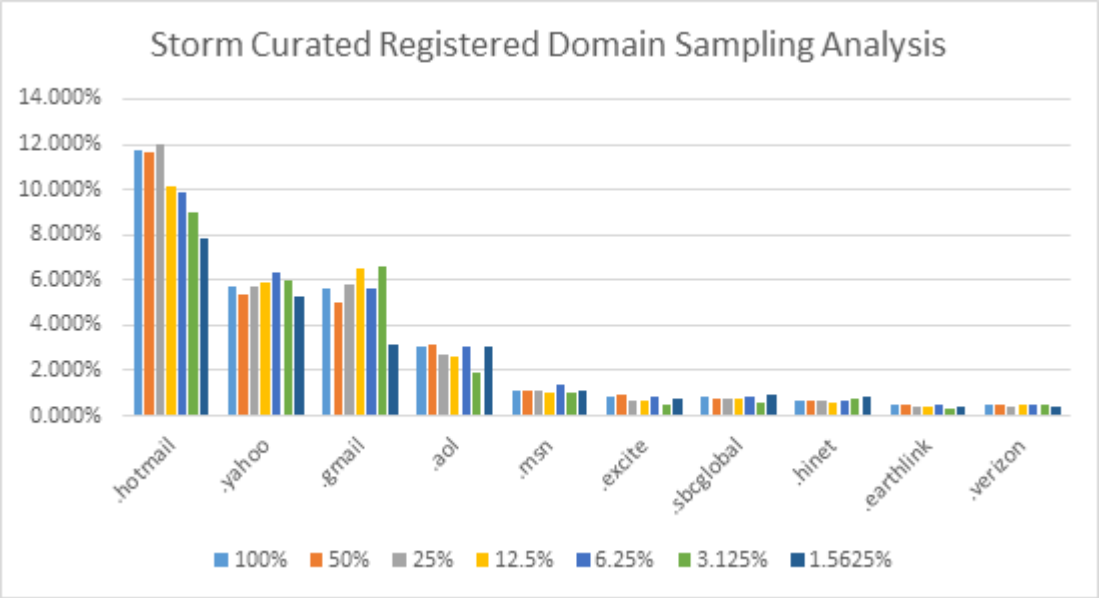


Figure 6.63: The percentage of addresses for each sampling subset of the Storm list containing a registered domain listed along the x-axis. The domains are the top 10 most encountered registered domains for the Storm list. The percentage of the original Storm list's addresses each subset contains is shown in the legend.

The registered domain distributions for this Storm list do not show the large variations seen in some of the previous lists. These results can be seen in Figure 6.64, where it is readily apparent that none of the comparisons exceed a 50% difference. For the 50% and 25% subsets eight and seven out of the ten comparisons respectively fall below a 10% difference and the other results remain below 20%.

This consistency wavers with the following samplings: the 12.5% distribution only has four domains under a 10% difference and another five below 20%. The final comparison just surpasses a 20% difference. However, the 6.25% sampling shows more similarities to the original distribution with six differences below 10% and the remaining four below 20%. The final two subsets both show stronger dissimilarities with 50% and 40% of their domain comparisons exceeding a 20% differential with the original distribution respectively.

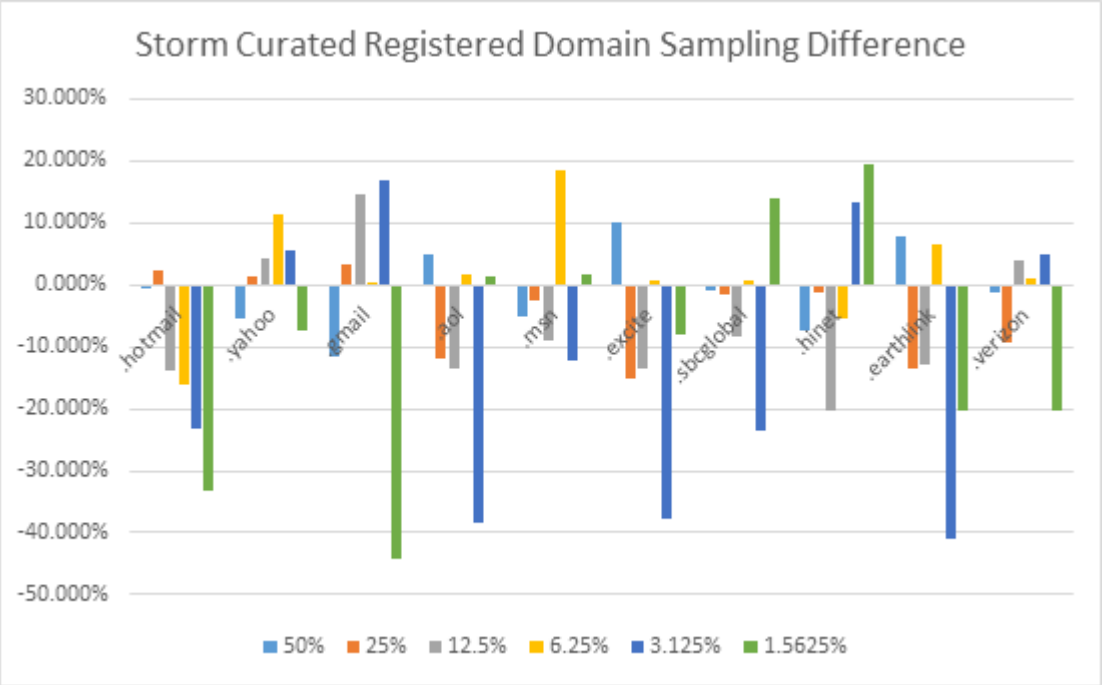


Figure 6.64: The percentage difference for Storm's top 10 most encountered registered domains compared against the distributions of the entire list. The percentage of the original Storm list's addresses each subset contains is shown in the legend.

The distribution sampling results for the registered domains seem to indicate that they may be useful in identifying the Storm botnet as a source. The two largest subsets show few disparities and require at least 236,000 addresses to achieve this level of confidence. Both the 12.5% and 6.25% showed several variances, but had few major outliers. They require over 118,000 and 59,000 addresses each respectively.

However, the previous cross-list analysis in Section 6.1.3 showed that the registered domain distributions for this Storm list did not correspond to the second Storm list. Although it seems plausible that the registered domain distributions may be useful as a classifier for Storm campaigns from this sampling analysis, it is unlikely these domains will successfully identify addresses from an entirely separate collection with the divergences seen previously comparing the two Storm lists.

6.2.7 Storm (C&C)

With the analysis of the second Storm list, ideally similar results as the first Storm list analyzed in the previous section will be seen. Yet the second list is a much larger source which may lead to different results. The top-level domain sampling distributions for the Storm (C&C) list are shown below in Figure 6.65. These distributions appear to be very similar to the first Storm lists shown in Section 6.2.6. The *.com* domain dominates the distributions while also showing the largest variance.

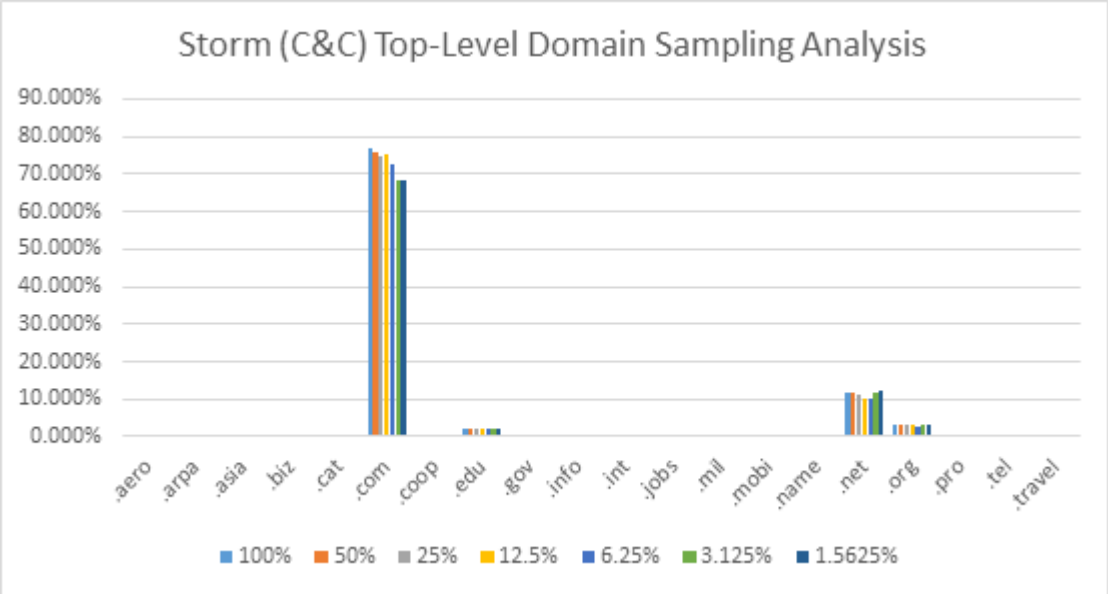


Figure 6.65: The percentage of addresses for each sampling subset of the Storm (C&C) list which end in each of the Top-Level domains listed along the x-axis. The percentage of the original Storm (C&C) list's addresses each subset contains is shown in the legend.

Unlike what was previously seen when comparing the top-level domain distributions of the two Storm lists, the sampling distributions for the second Storm list shows fewer differences as seen in Figure 6.66. The first two subsets shows that no comparison exceed a 20% difference. Additionally, the two smallest subsets only have one domain each with just over an 11% difference while the other three domains remain below 10%. The middle two samplings show more divergence, with the 12.5% subset having two

domains falling under a 10% difference and two which exceed a 10% difference. The 6.25% subset has one domain below a 10% difference, two above a 10% difference, and one just exceeding a 20% difference.

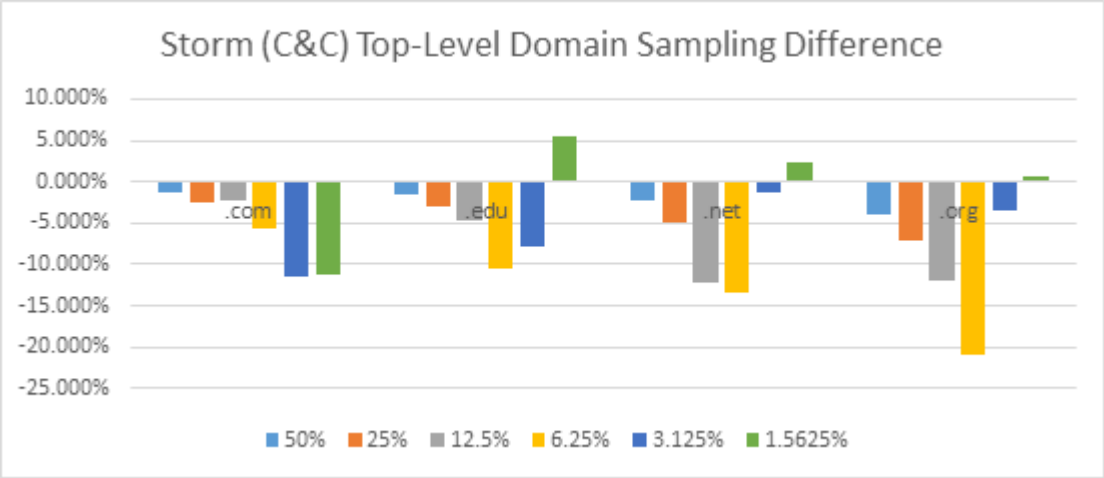


Figure 6.66: The percentage difference for Storm (C&C)s top-level domains compared against the distributions of the entire list. Only the four dominant domains are listed along the x-axis. The percentage of the original Storm (C&C) list's addresses each subset contains is shown in the legend.

In this situation, it seems like the top-level domain distributions have fewer differences than previously seen for the Storm lists. Nevertheless, with the small variety of viable top-level domains for analysis, and the minor variations present in some of the sampling distributions, it still appears unlikely that the top-level domain distributions would create a strong classifier for the Storm lists.

The country domain distributions for the second Storm list show the now common upward trend in Figure 6.67. However, the distributions for these domains do not show the sudden differences which appear in some of the samplings from the first Storm list.

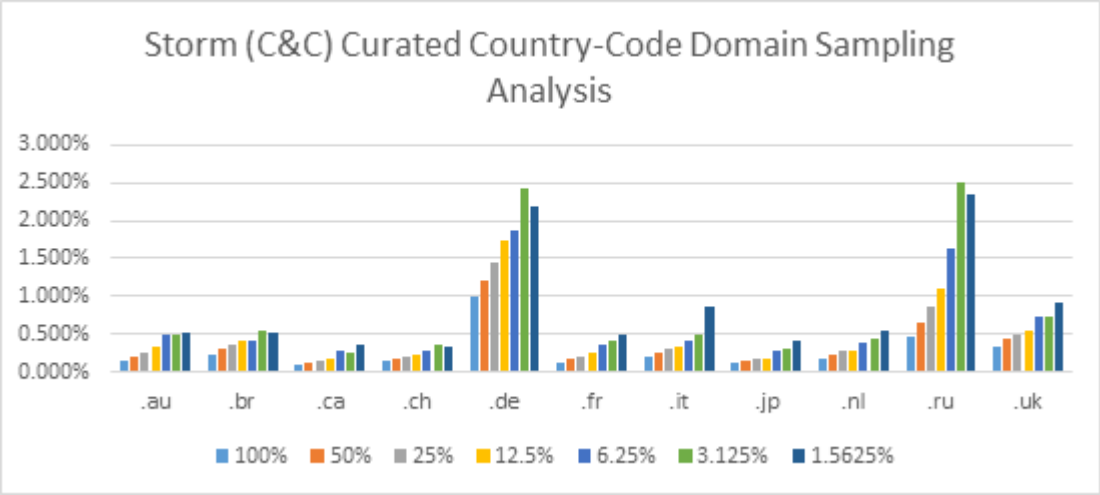


Figure 6.67: The percentage of addresses for each sampling subset of the Storm (C&C) list which end in each of the Country-Code domains listed along the x-axis. These domains are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Storm (C&C) list's addresses each subset contains is shown in the legend.

The percentage difference between each subset's distribution and the original list is shown below in Figure 6.68. Every comparison from these samplings exceeds a 20% difference. By the 12.5% subset, comparisons begin to exceed a 100% difference. At the 6.25% subset some exceed 200% and with the 3.125% subset they exceed 400%. It is clear the country-code domain distributions will not correctly identify a list from Storm.

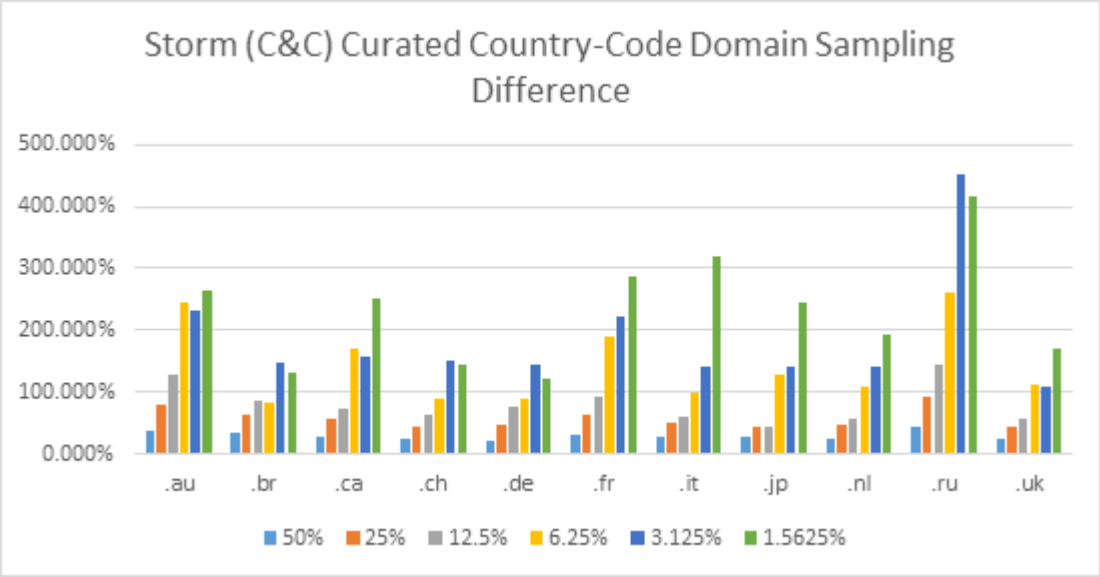


Figure 6.68: The percentage difference for Storm (C&C)s country-code domains compared against the distributions of the entire list. These domains along the x-axis are the same subset of the possible country-code domains as used previously in the cross-list domain distribution analysis. The percentage of the original Storm (C&C) list's addresses each subset contains is shown in the legend.

The registered domains sampling distributions for the second Storm list are detailed in Figure 6.69. These results show variances in several of the domains. It should also be noted that the top ten domains for this Storm list do not coincide with the ten domains from the first Storm list in Section 6.2.6 (four are different). This fact in addition to the registered domain analysis regarding the Storm lists in Section 6.1.3 together cast doubt on the ability to use the registered domains to classify Storm.

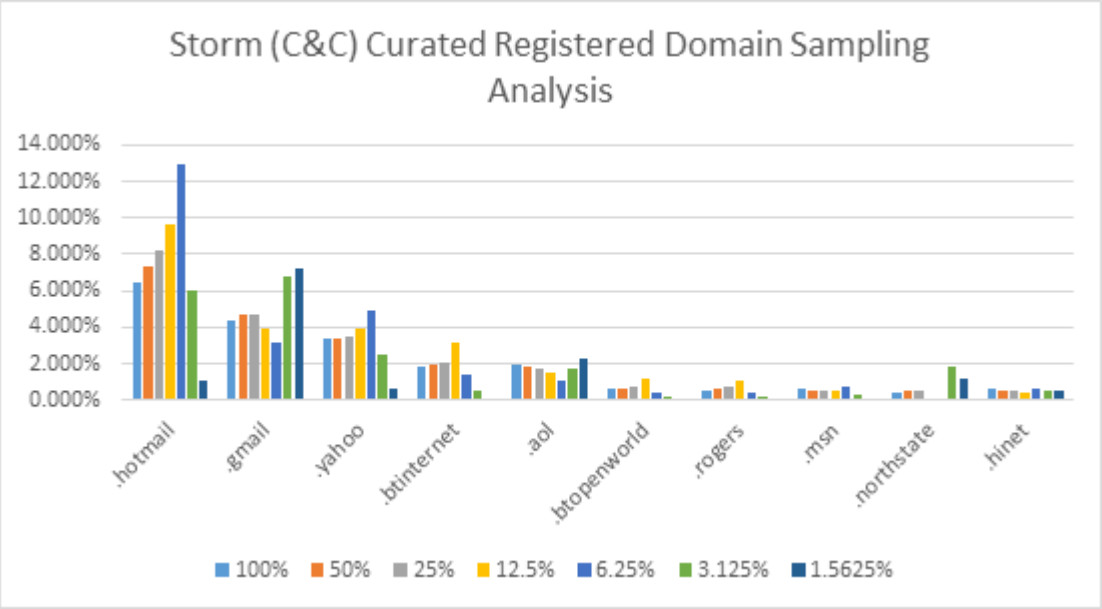


Figure 6.69: The percentage of addresses for each sampling subset of the Storm (C&C) list containing a registered domain listed along the x-axis. The domains are the top 10 most encountered registered domains for the Storm (C&C) list. The percentage of the original Storm (C&C) list's addresses each subset contains is shown in the legend.

The proportional differences for the Storm (C&C) registered domain sampling subsets are shown in Figure 6.70. The 50% subset has a few minor discrepancies from the original distribution with six domains below a 10% difference and the remaining four under 20%. While this is not as strong a correlation as seen in the other lists for the largest domain, it remains similar.

However, after this subset, the Storm (C&C) samplings become increasingly divergent. The 25% subset only has four domains with a resulting difference under 10%, while another three remain below 20% and its last three exceed a 20% difference. The remaining four subsets only have three comparisons between them with less than a 10% difference. Seventy-five percent of the comparisons in these four domains exceed a 20% difference. The comparisons almost reach a 100% difference starting with the 12.5% sampling and they peak at a 290% difference in the 3.125% subset.

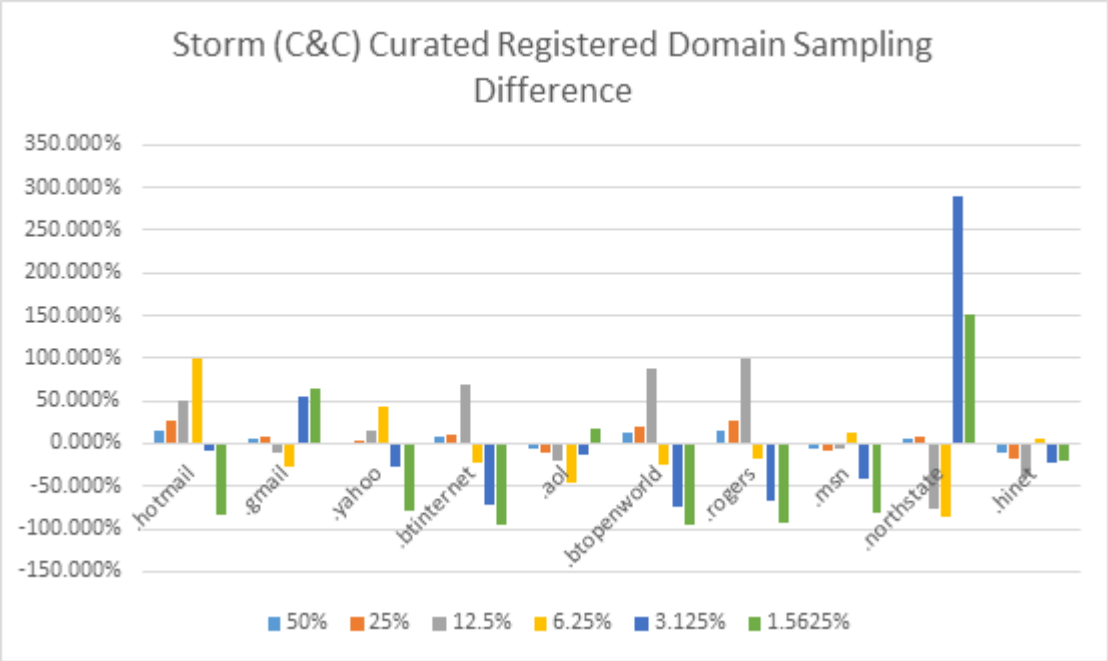


Figure 6.70: The percentage difference for Storm (C&C)s top 10 most encountered registered domains compared against the distributions of the entire list. The percentage of the original Storm (C&C) list's addresses each subset contains is shown in the legend.

It seems clear that the registered domain distributions cannot identify Storm. Disregarding the results affirming this from the cross-list analysis detailed in Section 6.1.3 and sampling analysis from the first Storm list in the previous section, these results also show consistent differences. To reach the relatively weak confidence shown by the 50% sampling subset above, approximately 37 million addresses are required.

Chapter 7

Conclusion

The focus of this project is to determine what can be learned from the addresses previously harvested by researchers from botnet campaigns. The first field of analysis concentrates on understanding what can be determined about the origins of the addresses and the general structure of these lists. The second set of analyses attempts to find defining characteristics unique to each botnet which can form a classifier to identify the botnet from a set of targeted addresses.

With regards to the first area of focus, there are several structural details of note for these lists. Determining the sources of the addresses, however, leads to many conjectures and few determined facts. From the results of the list sortings, it is clear that while they occur in every list except Pushdo, invalid addresses are uncommon within these lists. At most they account for 0.16% of the addresses in a list. These invalidities indicate a portion of the addresses in the botnets are gathered through at least one method with minor errors. It also indicates that a portion of the addresses are accumulated using an automated method.

Duplicated addresses are also uncommon in most lists and generally occur at a rate similar to that of the invalid addresses. There are two outliers in this case: the first is a minor one in the Grum lists where 0.87% of its addresses are duplicates; the second is much more extreme with the Storm (C&C) list where over 71% of its addresses are

duplicates. The duplication rate found in most lists indicate that the botnets put some effort into maintaining their lists. However, it is impossible to attribute the presence of the duplicates in these cases to an error within the botnets' address management systems or to an error or byproduct of the inspection and address collection methods researchers used to gather these lists. It is likely that a good portion of the duplicate addresses encountered are caused by the address normalization this sorting process employs. By removing registered sub-domains, a degree of distinctness which may exist for some addresses is eliminated. This likely accounts for the substantially lower rate of duplication found in the distinctness test run against 50,000 raw address samples.

Unlike the other lists, an extraordinarily high rate of duplication occurred in the Storm (C&C) list. The substantial proportion of duplicated addresses in the Storm (C&C) list are most likely caused by the collection method used by the researchers and from the address normalization already discussed. This assertion is reinforced by the low rate of duplication seen in the other Storm list in comparison to the unusually high rate of duplication in this list in addition to the much lower rate of duplication seen in the sampled distinctness test (approximately 4%).

The shared address analysis shows that a generous percentage of a list's addresses are found in other lists (as high as 15.95% of Pushdo's addresses are shared with MegaD). This initially indicates that an address source may be shared between the botnets or that one botnet sold a portion of its addresses to another. The address congruity search however shows no significant groupings of duplicated addresses across lists and the sampled distinct addresses test shows a much lower rate of shared addresses when working with addresses that have not been normalized. These observations indicate that it is unlikely the botnets share the exact same source or provide addresses to each other. It is possible that the shared entries occur from a shared or common implementation. An example of this would be two botnets each using automated processes which share a similar dictionary or generation routine for building addresses. Because users follow similar patterns and behaviors when selecting an email address (names and digits are obviously common) it is conceivable that the algorithms driving two separately created

automated processes might be similar.

However, it cannot be discounted that the results of the contiguity test may be misleading because it cannot be known if the address orderings were transformed before this analysis. This transformation might be an artifact of how the botnet operated or how the researchers gathered those addresses. For example, a list may also have been built through independent queries by researchers which result in the addresses having no relationship to each other or to the ordering they are stored internally within the botnet.

The second area of focus for this analysis is the plausibility of using the distributions of characteristics of the addresses within the lists to create a method of classifying the source of a spam campaign given a list of the addresses of a certain size that were targeted. This analysis concentrates on comparing the distributions of the different domain levels. These comparisons are made against other lists to validate differences can be distinguished and against themselves through sampling to ensure they can correctly identify the shared source.

With regards to the top-level domain distributions, only a limited number of domains are significant enough to be viable indices of the classifier. The *.com*, *.net*, *.org*, and *.edu* domains have significantly more addresses than any other top-level domain, but they remain generally consistent in their distribution across the lists. This is unsurprising as the Internets websites are significantly divided across these four domains. With few domains available, each comparison is important. The general consistency mentioned earlier shows in both the cross-list and the list sampling distribution analyses with only a few minor differences. This uniformity in the sampled distributions indicates the top-level domains are effective identifiers for the source of a list. However, this same consistency in the cross-list comparisons in addition to the lack of a variety of variables to compare with show the top-level domains are not capable of distinguishing between different sources.

The country code domains show the exact opposite trends. First, there are a variety of domains available to compare (although the domains are still bounded) which lessens the impact of a single outlying comparison. This remains true after the domains

examined are limited to only those found in at least 0.5% of the addresses in one or more of the lists. Second, where the top-level domains show reliable consistency, the ccTLDs show regular inconsistency. The variances in the cross-list comparisons show these domains can distinguish the different botnets. However, these same tests result in the two Storm lists presenting different distributions with significant discrepancies. This outcome is reinforced in the sample distribution analyses which also show consistent differences in the distributions. While the country code domains can differentiate lists from different sources, they also differentiate and do not correctly identify lists from the same source.

The final domains analyzed are the registered domains. Unlike the previous two domain types, the registered domains are not bounded to a known set. With hundreds of thousands of registered domains present across the lists, two subsets are used to limit the analysis to the most influential domains for this analysis. The first subset is comprised of the 10 most prevalent registered domains from each of the lists. The second contains each domain encountered in at least 1% of the addresses in one or more lists. In the cross-list analysis, results similar to the ccTLDs are seen. The registered domain distributions between lists show large differences, even between the two Storm lists. This shows that the registered domains are useful in distinguishing between lists, but cannot accurately identify two lists originating from the same botnet. This latter conclusion is confirmed in the sampling distribution analysis which generally reveals similarities with the large samplings progressing to greater differences as the sampling size decreased. To achieve a moderate level of confidence in a comparison often requires a sample size anywhere from 12.5% to 25% of the list. The Storm (C&C) list requires closer to 50% of its addresses in the sample to not result in large differences. The scale of these requirements is impractical in a real-world classifier scheme. Bearing that these samples are taken from the lists they are compared against, it is likely that a set of addresses gathered separately from the set a classifier is derived from would require a larger number of addresses to reach the same level of confidence.

The distributions of the domains in a set of addresses can successfully distin-

guish between two separate sets. However, the experiments run in this analysis show that they cannot identify a shared source between two separate lists from the same origin with any measure of trust. Without this second characteristic, a classifier cannot be designed using a distribution based signature to identify the source of a spam campaign given a set of targeted addresses.

Appendix A

Domain List

A.1 Top-Level Domain

The following is a list of all the top-level domains recognized by the list analysis components. This list was taken from ICANN and contains all the supported Top-Level Domains at the time the address lists were collected.

- .aero
- .arpa
- .asia
- .bitnet
- .biz
- .cat
- .com
- .coop
- .edu

- .gov
- .info
- .int
- .jobs
- .mil
- .mobi
- .museum
- .name
- .net
- .org
- .pro
- .tel
- .travel

A.2 Country-Code Domain

The table below contains a list of all the country-code domains recognized by the analysis components. The list contains the domain, the name of the country managing it, and optionally a list of all the country specific sub-domains the managing authority of the country domain supports (in addition to the Top-Level domains listed previously).

Table A.1: The Country-Code domains and their authorized sub-domains.

Domain	Country	Country-Specific Second-Level Domains
.ac	Acension Island	
.ad	Andorra	.nom; .internet; .web; .portal; .online; .wap; .clic; .telnet; .bbs; .tcp; .dns; .wais; .email; .www; .ftp; .smtp; .http; .mbone; .ietf; .rfc; .nom; .firm; .arts; .store; .shop; .home; .news
.ae	United Arab Emirates	.co; .con; .ent; .nae; .nat; .nea; .ner; .ac; .sch
.af	Afghanistan	
.ag	Antigua and Barbuda	.co; .nom; .enum; .example; .localhost; .ns; .ftp; .whois; .wpad; .brand; .tm; .ac; .bd
.ai	Anguila	.off
.al	Albania	.uniti; .tirana; .soros; .upt; .inima
.am	Armenia	
.an	Netherlands Antilles	
.ao	Angola	.ed; .gv; .og; .co; .pb; .it
.aq	Antarctica	
.ar	Argentina	.gob; .tur; .argentina; .educ; .gobiernoelectronico; .nic; .promocion; .retina; .uba
.as	American Samoa	
.at	Austria	.gv; .ac; .co; .or
.au	Australia	.csiro; .asn; .id; .act; .nsw; .nt; .qld; .sa; .tas; .vic; .wa; .archie; .conf; .gw; .otc; .oz; .telememo
.aw	Aruba	
.ax	Aland Islands	
.az	Azerbaijan	.pp; .co
.ba	Bosnia and Herzegovina	.unsa; .untz; .unmo; .unbi; .unze; .co; .rs
.bb	Barbados	.co; .store; .tv
.bd	Bangladesh	.ac
.be	Belgium	.ac; .vl; .vla; .vln; .fla; .dot
.bf	Burkina Faso	

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.bg	Bulgaria	.a; .b; .c; .d; .e; .f; .g; .h; .i; .j; .k; .l; .m; .n; .o; .p; .q; .r; .s; .t; .u; .v; .w; .x; .y; .z; .0; .1; .2; .3; .4; .5; .6; .7; .8; .9
.bh	Bahrain	.cc
.bi	Burundi	.co; .or
.bj	Benin	.gouv; .asso; .barreau
.bl	Saint Barthelemy	
.bm	Bermuda	
.bn	Brunei Darussalam	
.bo	Bolivia	.tv; .gob
.br	Brazil	.adm; .adv; .agr; .am; .arq; .art; .ato; .bio; .blog; .bmd; .cim; .cng; .cnt; .ecn; .eng; .esp; .etc; .eti; .far; .flog; .fm; .fnd; .fot; .fst; .g12; .ggf; .imb; .ind; .inf; .jor; .jus; .lel; .mat; .med; .mus; .nom; .not; .ntr; .odo; .ppg; .psc; .psi; .qsl; .rec; .slg; .srv; .tmp; .trd; .tur; .tv; .vet; .vlog; .wiki; .zlg; .nom; .sec3; .taxi; .teo; .far; .radio; .am; .b
.bs	Bahamas	
.bt	Bhutan	
.bv	Bouvet Island	
.bw	Botswana	.co
.by	Belarus	
.bz	Belize	
.ca	Canada	.ab; .bc; .mb; .nb; .nf; .nl; .ns; .nt; .nu; .on; .pe; .qc; .sk; .yk; .gc
.cc	Cocos Islands	.co
.cd	The Democratic Republic of the Congo	.fm; .am; .tv; .ws; .dj; .mu; .me
.cf	Central African Republic	
.cg	Congo	
.ch	Switzerland	
.ci	Cote DIvoire	
.ck	Cook Islands	.co; .gen
.cl	Chile	
.cm	Cameroon	

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.cn	China	.ac; .ah; .bj; .cq; .fj; .gd; .gs; .gz; .gx; .ha; .hb; .he; .hi; .hl; .hn; .jl; .js; .jx; .ln; .nm; .nx; .qh; .sc; .sd; .sh; .sn; .sx; .tj; .tw; .xj; .xz; .yn; .zj
.co	Columbia	.nom
.cr	Costa Rica	.ac; .co; .ed; .fi; .go; .or; .sa
.cu	Cuba	
.cv	Cape Verde	
.cx	Christmas Island	
.cy	Cyprus	.ac; .ekloges; .tm; .ltd; .press; .parliament
.cz	Czech Republic	
.de	Germany	
.dj	Djibouti	
.dk	Denmark	
.dm	Dominica	
.do	Dominican Republic	.gob; .sld; .web; .art
.dz	Algeria	.asso; .pol; .art
.ec	Ecuador	.fin; .med
.ee	Estonia	
.eg	Egypt	.eun; .sci; .eregistry
.eh	Western Sahara	
.er	Eritrea	
.es	Spain	.nom; .gob
.et	Ethiopia	
.eu	European Union	.eurid; .registry; .nic; .dns; .internic; .whois; .das; .coc; .eurethix; .euthics
.fi	Finland	
.fj	Fiji	.ac
.fk	Falkland Islands (Malvinas)	.co; .ac; .nom
.fm	Federated States of Micronesia	
.fo	Faroe Islands	
.fr	France	.tm; .asso; .nom; .prd; .presse; .gouv
.ga	Gabon	
.gb	United Kingdom	
.gd	Grenada	

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.ge	Georgia	.pvt
.gf	French Guiana	
.gg	Guernsey	.ac; .co; .sch
.gh	Ghana	
.gi	Gibraltar	.ltd; .mod
.gl	Greenland	.co
.gm	Gambia	
.gn	Guinea	.ac
.gp	Guadeloupe	.asso
.gq	Equatorial Guinea	
.gr	Greece	
.gs	South Georgia and South Sandwich Islands	
.gt	Guatemala	.gob; .ind
.gu	Guam	
.gw	Guinea-Bissau	
.gy	Guyana	.co
.hk	Hong Kong	.idv
.hm	Heard Island and McDonald Islands	
.hn	Honduras	
.hr	Croatia	.iz; .from
.ht	Haiti	
.hu	Hungary	
.id	Indonesia	.ac; .co; .or; .web; .sch; .go
.ie	Ireland	.irlgov
.il	Israel	.ac; .co; .k12; .muni; .idf
.im	Isle of Man	
.in	India	.co; .firm; .gen; .ind; .ac; .res; .nic
.io	British Indian Ocean Territory	
.iq	Iraq	
.ir	Islamic Republic of Iran	.ac; .co; .id; .nic; .ir; .sch
.is	Iceland	

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.it	Italy	
.je	Jersey	
.jo	Jordan	.co; .sch
.jp	Japan	.ac; .ad; .co; .ed; .geo; .go; .gr; .lg; .ne; .or
.ke	Kenya	.co; .or; .ne; .go; .ac; .sc
.kg	Kyrgyzstan	
.kh	Cambodia	.per
.ki	Kiribati	.de
.km	Comoros	.ass; .nom; .presse; .tm; .medecin; .notaire; .pharmacien; .verterinaire; .gouv; .prd; .journaliste; .aeroport; .avocat; .ccia; .port; .juriste; .ingenieur; .mandataireminier
.kn	Saint Kitts and Nevis	
.kp	Democratic Peoples Republic of Korea	
.kr	Republic of Korea	.co; .ne; .or; .re; .pe; .go; .ac; .hs; .ms; .es; .sc; .kg; .seoul; .busan; .daegu; .incheon; .gwangju; .daejeon; .ulsan; .gyeonggi; .gangwon; .chungbuk; .chungnam; .jeonbuk; .jeonnam; .gyeongbuk; .gyeongnam; .jeju
.kw	Kuwait	
.ky	Cayman Islands	
.kz	Kazakhstan	
.la	Lao Peoples Democratic Republic	
.lb	Lebanon	
.lc	Saint Lucia	.co; .l; .p
.li	Liechtenstein	
.lk	Sri Lanka	.sch; .ngo; .soc; .web; .ltd; .assn; .grp; .hotel
.lr	Liberia	.co; .or
.ls	Lesotho	
.lt	Lithuania	
.lu	Luxembourg	
.lv	Latvia	.id; .asn; .conf
.ly	Libya	.plc; .sch; .med; .id
.ma	Morocco	.ac; .press; .co

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.mc	Monaco	.tm; .asso
.md	Republic of Moldova	
.me	Montenegro	.co; .ac; .its; .priv
.mf	Saint Martin	
.mg	Madagascar	.nom; .prd; .tm
.mh	Marshall Islands	
.mk	The Former Yugoslav Republic of Macedonia	.inf
.ml	Mali	.presse
.mm	Myanmar	
.mn	Mongolia	
.mo	Macao	
.mp	Northern Mariana Islands	
.mq	Martinique	
.mr	Mauritania	
.ms	Montserrat	
.mt	Malta	
.mu	Mauritius	.ac; .co; .or
.mv	Maldives	
.mw	Malawi	.ac; .co
.mx	Mexico	.gob
.my	Malaysia	
.mz	Mozambique	.co
.na	Namibia	.co
.nc	New Caledonia	.asso
.ne	Niger	
.nf	Norfolk Island	.per; .rec; .web; .arts; .firm; .other; .store
.ng	Nigeria	
.ni	Nicaragua	.co; .gob; .ac; .nom; .web; .in
.nl	Netherlands	
.no	Norway	.fhs; .vgs; .gs; .fylkesbibl; .folkebibl; .idrett; .priv
.np	Nepal	
.nr	Nauru	

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.nu	Niue	
.nz	New Zealand	.ac; .co; .geek; .gen; .maori; .school; .cri; .govt; .iwi; .parliament; .archie
.om	Oman	.co; .ac; .sch; .med
.pa	Panama	.ac; .sld; .gob; .abo; .ing; .med; .nom
.pe	Peru	.gob; .nom; .sld
.pf	French Polynesia	
.pg	Papua New Guinea	.ac
.ph	Philippines	.ngo; .i
.pk	Pakistan	.fam; .web; .gob; .gok; .gon; .gop; .gos

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.pl	Poland	.art; .waw; .wroc; .krakow; .katowice; .poznan; .lodz; .gda; .gdansk; .slupsk; .radom; .szczecin; .lublin; .bialystok; .olsztyn; .torun; .zgora; .aid; .agro; .atm; .auto; .gmina; .gsm; .mail; .miasta; .media; .nieruchomosci; .nom; .pc; .powiat; .priv; .realestate; .rel; .sex; .shop; .sklep; .sos; .szkola; .targi; .tm; .tourism; .turystyka; .augustow; .babia-gora; .bedzin; .beskidy; .bialowieza; .bielawa; .bieszczady; .boleslawiec; .bydgoszcz; .bytom; .cieszyn; .czeladz; .czest; .dlugoleka; .elblag; .elk; .glogow; .gniezno; .gorlice; .grajewo; .ilawa; .jaworzno; .jelenia-gora; .jgora; .kalisz; .kazimierz-dolny; .karpacz; .kartuzy; .kaszuby; .kepno; .ketrzyn; .klodzko; .kobierzyce; .kolobrzeg; .konin; .konskowola; .kutno; .lapy; .lebork; .legnica; .lezajsk; .limanowa; .lomza; .lowicz; .lubin; .lukow; .malbork; .malopolska; .mazowsze; .mazury; .mielec; .mielno; .mragowo; .naklo; .nowaruda; .nysa; .olawa; .olecko; .olkusz; .opoczno; .opole; .ostroda; .ostroleka; .ostrowiec; .ostrowwlkp; .pila; .pisz; .podhale; .podlasie; .polkowice; .pormorze; .pormorskie; .prochowice; .pruszkow; .przeworsk; .pulawy; .rawa-maz; .rybnik; .rzeszow; .sanok; .sejny; .slask; .sosnowiec; .stalowa-wola; .skoczow; .starachowice; .stargard; .suwalki; .swidnica; .swiebodzin; .swinoujscie; .szczytno; .tarnobrzeg; .tgory; .turek; .tychy; .ustka; .walbrzych; .warmia; .warszawa; .wegrow; .wielun; .wlocl; .wloclawek; .wodzislaw; .wolomin; .wroclaw; .zachpomor; .zagan; .zarow; .zgorzelec
.pm	Saint Pierre and Miquelon	
.pn	Pitcairn	.co

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.pr	Puerto Rico	.isla; .est; .prof; .ac
.ps	Occupied Palestinian Territory	.sch; .mun
.pt	Portugal	.nome; .publ
.pw	Palau	.belau
.py	Paraguay	.una
.qa	Qatar	
.re	Reunion	.asso; .nom
.ro	Romania	.tm; .nt; .nom; .rec; .arts; .firm; .store; .www
.rs	Serbia	.co; .ac; .in
.ru	Russian Federation	.pp; .ac
.rw	Rwanda	.ac; .co; .gouv
.sa	Saudi Arabia	.sch; .med; .pub
.sb	Solomon Islands	
.sc	Seychelles	
.sd	Sudan	.med; .tv
.se	Sweden	.a; .b; .ac; .bd; .c; .d; .e; .f; .g; .h; .i; .k; .l; .m; .n; .o; .p; .r; .s; .t; .u; .w; .x; .y; .z; .pp; .tm; .parti; .press
.sg	Singapore	.per; .idn
.sh	Saint Helena, Ascension and Tristan da Cunha	.co; .nom
.si	Slovenia	
.sj	Svalbard and Jan Mayen	
.sk	Slovakia	
.sl	Sierra Leone	
.sm	San Marino	
.sn	Senegal	
.so	Somalia	
.sr	Suriname	
.st	Sao Tome and Principe	.saotome; .principe; .consulado; .embaixada; .store; .co
.su	USSR	
.sv	El Salvador	.gob; .red

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.sy	Syrian Arab Republic	.news
.sz	Swaziland	.co; .ac
.tc	Turks and Caicos Islands	
.td	Chad	
.tf	French Southern Territories	.eu; .us
.tg	Togo	
.th	Thailand	.ac; .co; .in; .go; .mi; .or
.tj	Tajikistan	.ac; .co; .dyn; .go; .my; .per; .web
.tk	Tokelau	
.tl	Timor-Leste	
.tm	Turkmenistan	
.tn	Tunisia	.ens; .fin; .ind; .intl; .nat; .perso; .tourism; .edunet; .rnrt; .rns; .rnu; .mincom; .agrinet; .defense; .turen
.to	Tonga	
.tp	East Timor	
.tr	Turkey	.gen; .av; .dr; .pol; .bel; .tsk; .bbs; .k12; .web; .tv
.tt	Trinidad and Tobago	.co
.tv	Tuvalu	
.tw	Taiwan, Province of China	.idv; .game; .ebiz; .club
.tz	United Republic of Tanzania	
.ua	Ukraine	.in; .ck; .cherkassy; .cn; .chernigov; .cv; .chernovtsy; .crimea; .dnepropetrovsk; .dp; .dn; .donetsk; .if; .ivano-frankivsk; .kh; .kharkov; .kherson; .ks; .khmelnitskiy; .km; .kiev; .kv; .kirovograd; .kr; .lugansk; .lg; .lutsk; .lviv; .nikolaev; .mk; .odessa; .od; .poltava; .pl; .rovno; .rv; .sebastopol; .sumy; .ternopil; .te; .uzhgorod; .vinnica; .vn; .zaporizhzhe; .zp; .zhitomir; .zt
.ug	Uganda	.co; .ac; .sc; .go; .ne; .or

Continued on next page

Table A.1 – continued from previous page

Domain	Country	Country-Specific Second-Level Domains
.uk	United Kingdom	.ac; .co; .ltd; .me; .mod; .nic; .nhs; .plc; .police; .sch; .govt; .orgn; .lea; .parliament
.um	United States Minor Outlying Islands	
.us	United States	.ak; .al; .ar; .az; .ca; .co; .ct; .de; .fl; .ga; .hi; .ia; .id; .il; .in; .ks; .ky; .la; .ma; .md; .me; .mi; .mn; .mo; .ms; .mt; .nc; .nd; .ne; .nh; .nj; .nm; .nv; .ny; .oh; .ok; .or; .pa; .ri; .sc; .sd; .tn; .tx; .ut; .va; .vt; .wa; .wi; .wv; .wy; .as; .dc; .gu; .pr; .vi; .dni; .fed; .isa; .kids; .nsn
.uy	Uruguay	.gub
.uz	Uzbekistan	.co
.va	Holy See (Vatican City-State)	
.vc	Saint Vincent and the Grenadines	
.ve	Bolivarian Republic of Venezuela	.gob; .co; .web; .tec
.vg	British Virgin Islands	
.vi	U.S. Virgin Islands	.co; .k12
.vn	Viet Nam	.ac; .health
.vu	Vanuatu	
.wf	Wallis and Futuna	
.ws	Samoa	
.ye	Yemen	.co; .ltd; .me; .plc
.yt	Mayotte	
.yu	Yugoslavia	
.za	South Africa	.ac; .city; .co; .law; .nom; .school; .alt; .ngo; .tm; .web; .bourse
.zm	Zambia	.ac; .co
.zw	Zimbabwe	.ac; .co

References

[1] Ironport, ”2008 Internet Security Trends,” 2008.

[2] MessageLabs, Symantec Hosted Services, ”Fall Harvests: Spooky Spam, Frightening Phish and Creepy CAPTCHA Cracking,” MessageLabs Intelligence, October 2008.

[3] Federal Trade Commission, ”Email Address Harvesting: How Spammers Reap What You Sow,” FTC Consumer Alert, 2002.

[4] Center for Democracy and Technology, ”Why Am I Getting All This Spam? Unsolicited Commercial E-mail Research Six Month Report,” Washington D.C., 2003.

[5] G. Schryen, ”The impact that placing email addresses on the Internet has on the receipt of spam: An empirical analysis,” *Computers & Security*, vol. 26, no. 5, pp. 361-372, 2007.

[6] M. B. Prince, B. Dahl, L. Holloway, A. M. Keller and E. Langheinrich, ”Understanding How Spammers Steal Your E-Mail Address: An Analysis of the First Six Months of Data from Project Honey Pot,” in *Second Conference on Email and Anti-Spam*, Stanford, CA, 2005.

[7] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson and S. Savage, ”On the Spam Campaign Trail,” in *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, San Francisco, 2008.

[8] J. P. John, A. Moshchuk, S. D. Gribble and A. Krishnamurthy, ”Studying Spamming Botnets Using Botlab,” in *Proceedings of the 6th USENIX Symposium on Networks Systems Design and Implementation*, Boston, 2009.

[9] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxon and

S. Savage, ”Spamalytics: An Empirical Analysis of Spam Marketing Conversion,” in *Proceedings of the 15th ACM Conference on Computer and Communications Security*, New York, 2008.

[10] A. Mushtaq, ”FireEye Blog: Killing the Beast - Part 5,” FireEye, 9 July 2012. [Online]. Available: http://blog.fireeye.com/research/2012/07/killing-the-beast-part-5.html.

[11] MessageLabs, Symantec Hosted Services, ”New Botnets Boost Spam Growth to Near Pre-McColo Levels,” MessageLabs Intelligence, January 2009.

[12] MessageLabs, Symantec Hosted Services, ”Cutwail’s Bounce-Back; Instant Messages can Lead to Instant Malware,” MessageLabs Intelligence, June 2009.

[13] MessageLabs, Symantec Hosted Services, ”Latest Investigation of Spam from Botnets Reveals Rapid Growth; Rustock’s Heartbeat, Maazben Gambles to Dominate and Grum Becomes Worst Offender,” MessageLabs Intelligence, September 2009.

[14] MessageLabs, Symantec Hosted Services, ”Survival of the Fittest: Selfish Botnets Dominate the Spam Landscape as Rustock Becomes the Largest Botnet; Linux Takes a Share of Spam from Windows,” MessageLabs Intelligence, April 2010.

[15] MessageLabs, Symantec Hosted Services, ”In the Battle of the Botnets Rustock Remains Dominant,” MessageLabs Intelligence, August 2010.

[16] MessageLabs, Symantec Hosted Services, ”The ”Search” for Recession Spam, Trojans Leave Fake Vapour Trails and Cutwail Plays Cupid,” MessageLabs Intelligence, February 2009.

[17] MessageLabs, Symantec Hosted Services, ”2009 Annual Security Report,” MessageLabs Intelligence, 2009.

[18] MessageLabs, Symantec Hosted Services, ””Behind the Scenes” of Spam URLs; New Internet in Africa Attracts Spam Botnets; Soccer World Cup Themed Malware,” MessageLabs Intelligence, May 2010.

[19] MessageLabs, Symantec Hosted Services, ”2010 Annual Security Report,” MessageLabs Intelligence, December 2010.

[20] MesssageLabs, Symantec Hosted Services, ”2008 Annual Security Report,” MessageLabs Intelligence, 2008.

[21] MessageLabs, Symantec Hosted Services, ”Reputable Sources are Cyber Criminals Favored Resources; Spammers Work by US Clocks,” MessageLabs Intelligence,

May 2009.

[22] MessageLabs, ”2007 Annual Security Report: A Year of Storms, Spam and Socializing,” MessageLabs Intelligence, 2007.

[23] MessageLabs, ”Storm Significantly Weakens & Web-Based Malware on the Rise,” MessageLabs Intelligence, April 2008.

[24] MessageLabs, ”Spam Levels Amplified Through Innovations in Exploitation,” MessageLabs Intelligence, May 2008.

[25] MessageLabs, ”Storm Botnet Serves-Up a Diet of Fast-Flux Spam,” MessageLabs Intelligence, August 2007.

[26] MessageLabs, ”Phishing Attacks Overtake Viruses,” MessageLabs Intelligence, January 2007.

[27] MessageLabs, ”One Fifth of All Spam Springs from Storm Botnet,” MessageLabs Intelligence, March 2008.

[28] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson and S. Savage, ”Spamcraft: An Inside Look at Spam Campaign Orchestration,” in *Proceedings of the 2nd USENIX Conference on Large-Scale Exploits and Emergent Threats: Botnets*, Spyware, Worms, and More, Boston, 2009.