



An analysis of the COM Top Level Domain

 The article was obtained at the following URL: <http://www.kfwebs.net/articles/article/37>
 The article might be distributed further as long as it is provided as it is, with the credits stated.
 The Article was written and first published by KF Webs, at <http://www.kfwebs.net>
 #####

KF Webs recently got access to the gTLD zone-files, of which we decided to have a closer look on the COM zone-file.
Added: 2006-04-25 00:52:40 - Modified: 2006-05-06 14:08:45 - Level: Intermediate

Introduction

KF Webs recently got access to the gTLD zone-files, of which we decided to have a closer look on the COM zone-file. KF Webs performs domain name services through the [interface at passive12.net](http://www.kfwebs.net/interface_at_passive12.net), a fully automated service for easy and quick maintenance and new registration of domain names.

All the work has been performed on a system purchased in 2001. It is a Dual PIII 1000MHz with a gigabyte of RAM and a terabyte of storage. Not the most powerful box in the world, but it gets the job done. We used MySQL 5.0.20 with MyISAM tables to perform the analysis. The zonefile was copied 2006-04-24

What is a zone-file

DNS is the abbreviated for of Domain Name System (DNS). When you visit a website such as KFWEB.S.NET your computer sends a request to translate the domain name, kfwebs.net., from a human readable form into a computer readable form, referred to as an IP-address. In the time of writing the IP address of kfwebs.net looks like 213.161.224.2.

The DNS is hierarchical, both in its form and its function. Considering a domain such as kfwebs.net , the Top Level Domain (TLD) would be NET, while the second level domain name would be KFWEB.S.

An interesting note is that the full domain name is really "KFWEB.S.NET.". Notice the final dot and feel free to try to enter it in your Web User Agent's address-bar. Although this is technically the full name, the final dot is usually omitted, but it shows the hierarchical structure of the domain.

A ZONE-file is used on DNS servers in order to delegate control and to store information about domains. In the COM zone-file there are the two records:

```
SECURE-MY-INTERNET NS NS3.KFWEB.S.NET.
SECURE-MY-INTERNET NS NS4.KFWEB.S.NET.
```

These lines tells where to get more information about the SECURE-MY-INTERNET second level domain, and hence delegates control to the mentioned nameservers.

NS3.KFWEBS.NET. and NS4.KFWEBS.NET. again contains the information required for the computer to turn the human-readable form of the name into a computer readable form using a so-called A record. This zone-file again looks like:

```
SECURE-MY-INTERNET.COM. A 72.29.83.156
```

For more in-dept information [this wikipedia entry](#) is recommended

Getting started

The COM zone file, known as `com.zone.gz` on the FTP server is 810 MiB of size. Uncompressed it is 3.7 GiB. There was a total of 117,379,441 records, as each domain name has multiple records due to the nature of the DNS system using multiple nameservers. A minimum of two is required, although it is recommended to use at minimum of three nameservers and not more than seven.

About 870 thousand records were A records for such nameservers, so those got deleted, we also ran the zone file through some shell scripts in order for it to become a list of unique domain names, resulting in a 685 MiB file.

Loading into the database and indexing it was going at a rate of about 560,000 domain names a minute. It was loaded into a total of 11 MyISAM database tables in MySQL, then combined using a 12th table of the MERGE storage engine. The MERGE storage engine, also known as MRG_MyISAM was introduced in MySQL 3.23.25. You can read more in the [documentations](#)

At the present time the COM zone-file holds 51,268,278 active domain names, meaning that on average a domain name has 2.2895 nameservers.

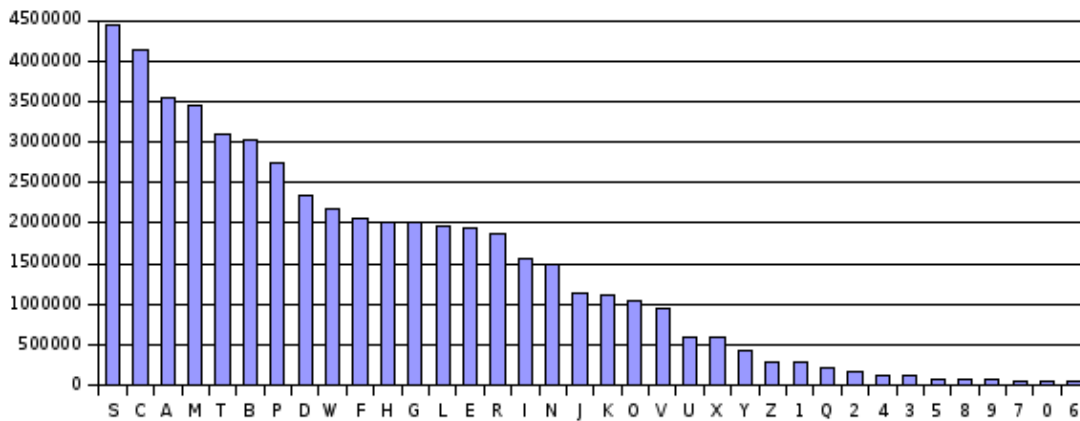


Distribution of names

Domain names by first character

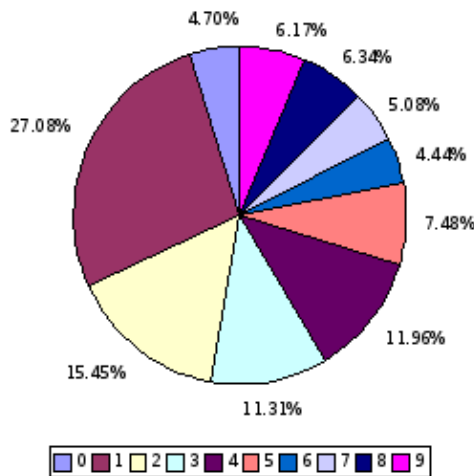
As shown by the graph, the characters most used as the first characters are S,C,A and M. Domain names starting with these characters account for 30.5 per cent of all COM domain names in total. Domain names starting with a digit accounts for less than two per cents.

Domain name by first character



Benford's law states that lists of numbers from many (but not all) real-life sources of data, the leading digit 1 occurs much more often than the others (namely about 30% of the time). Furthermore, the larger the digit, the less likely it is to occur as the leading digit of a number.

Domain names starting with a digit



To be precise Benford's law states that the leading digit n ($n \in \{1, \dots, b - 1\}$) in base b ($b \geq 2$) occurs with probability proportional to $\log_b(n + 1) - \log_b(n)$. In base 10, the leading digits have the following distribution by Benford's law compared to occurrence in COM domain names

Digit Benford's law Occurrence in COM

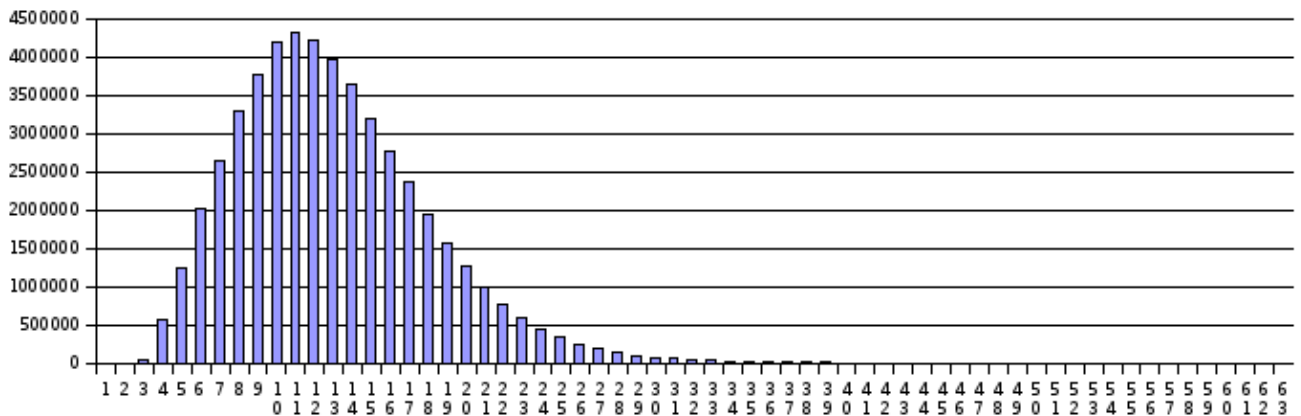
0	-	4.7%
1	30.1%	27.08%
2	17.6%	15.45%
3	12.5%	11.31%
4	9.7%	11.96%

5	7.9%	7.48%
6	6.7%	4.44%
7	5.8%	5.08%
8	5.1%	6.34%
9	4.6%	6.17%

Length of domain names

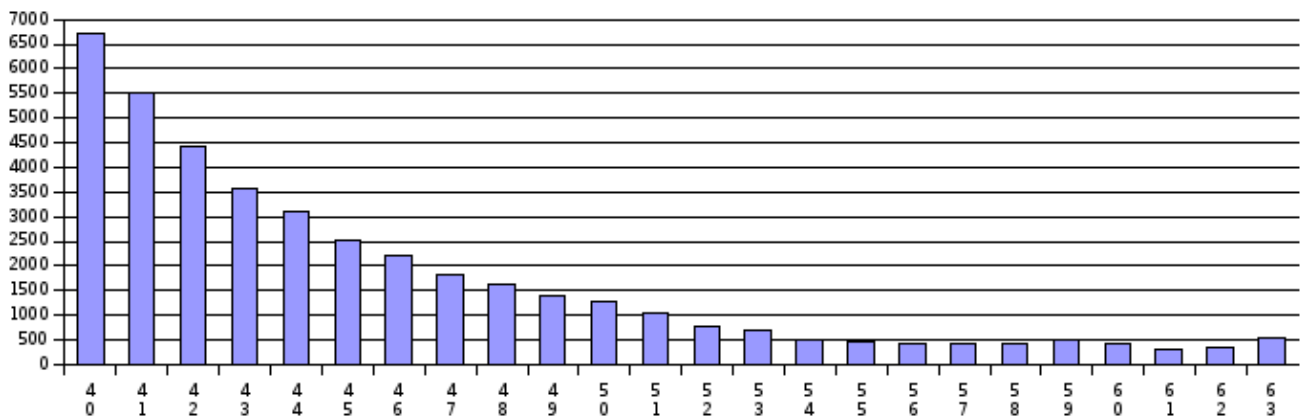
96.8 per cent of all domain names with 3 characters is taken, and the remainders probably contain a hyphen, or is considered without use for a domain name. The odds are somewhat better if looking at domain names with four characters where 31.8 per cent of the namespace is occupied. The length most common is 11 characters. [See if your name is available and/or register it at passive12.net](#)

Domain names by length of domain name



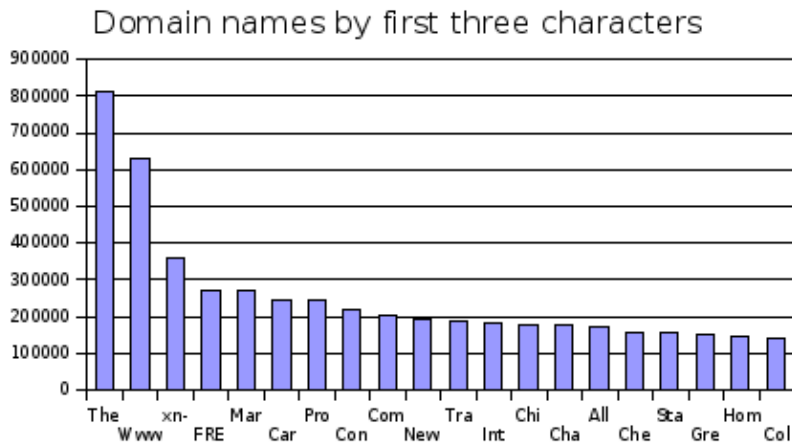
Although it looks like no domain name is registered with more than 40 characters, that isn't true, although the numbers are relatively small, and only 541 domain names are registered with the max length, 63 characters.

Domain names by length of domain name



Domains by first three characters

Looking at the top 20 groups of first three characters, "the" is the most common group with 811,220 active domain names.



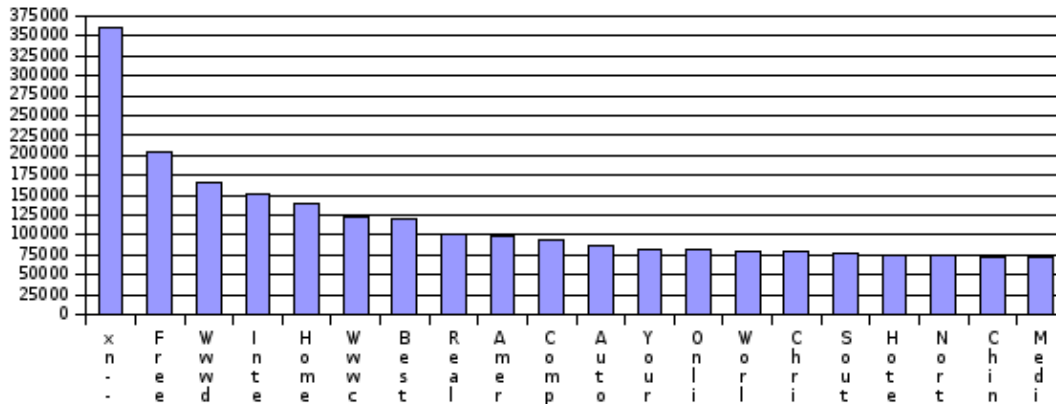
It is somewhat surprising to find "www" as the second largest group. One reason might be companies registering www-company.com in case someone mis-type www.company.com, another possible reason is that this is being done for phishing reasons.

Many domain name squatters register domains such as wwwcompanyname.com as well and try to sell it to the company or shows advertisement on the sites.



Domains by first four characters

Domain names by first four characters



Grouping by the first four characters, the grandest group is international domain names (xn--). This was followed by the phrase "free" which had roughly 205 thousand names in the group. Words like "home", "best", "real" and "your" was not unexpected to find here.

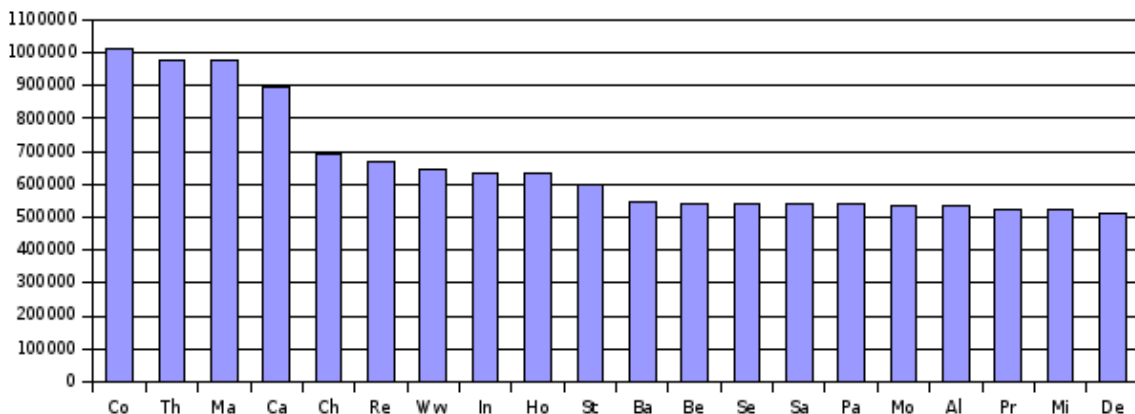
Based on the data from looking at the first four characters some additional lookups were performed. The following is a list of prefixes investigated

Prefix	Number
america*	82,410
online*	79,991
hotel*	73,699
computer*	25,826
christian*	21,565

Domains by first two characters

The distribution of names by the first two characters is the following

Domain names by groups of two characters



It was commented that it is surprising that "go" is not on the list, with occurrences such as GoDaddy and GoPedro. 354,548 domain names starts with the prefix "go". This is relatively high considering the number of words starting with go as compared to the other groups.

International domain names (xn--)

International domain names are stored as punycode. There are presently 361,420 domain names stored with international domain name support.

Punycode, defined in RFC 3492, is the self-proclaimed "bootstring encoding" of Unicode strings into the limited character set supported by the Domain Name System. The encoding is used as part of IDNA, which is a system enabling the use of internationalized domain names in all languages that are supported by Unicode, where the burden of translation lies entirely with the user application (a web browser for example).

The encoding is applied separately to each component of a domain name which is not represented solely within the ASCII character set, and a reserved prefix 'xn--' is added to the translated Punycode string. For example, bÃ¼cher becomes bcher-kva in Punycode, and therefore the domain name bÃ¼cher.ch would be represented as xn--bcher-kva.ch in IDNA.

Domain suffixes

There are currently 8879 active COM domain names beginning with "ebay". Many of these are probably used in phishing schemes trying to lure usernames and passwords from ebay users. Some examples might be "EBAY--PAYMENT" and "EBAY-ACCOUNT-REACTIVATE". Others again are perfectly legal and talks about ebay or the other services mentioned.

PayPal, known for payment processing, has been very exposed to hijacked accounts lately due to phishing schemes where the users themselves give their username and passwords to script kiddies. These again go on and purchase domain names, hosting and other objects and intangible items, often increasing the workload for those having to fix up after it.

1624 domain names begins with paypal. While some are perfectly legal names (not used in phishing), like paypal-alternative.com, too many can not have any possible use except trying to lure users.

Services differ when it comes to popularity, and seeing COMPANYNAME-haters.com and COMPANY-fans.com isn't uncommon, the following is a list of domain names that suffixes different companies.

Company Number

Ebay*	8879
google*	8378
yahoo*	7388
chase*	5229

microsoft* 3236
Paypal* 1624
ibm* 1586
egold* 766
Citibank* 519



Common names

Before proceeding, a little about some [DNS status messages](#). The statuses for the domains listed are per 2006-05-01.

REDEMPTIONPERIOD:

The registry sets this status when a registrar requests that the domain name be deleted from the registry and the domain has been registered for more than 5 calendar days (if the delete request is received within 5 days of initial domain registration it will instead be deleted immediately). The domain will not be included in the zone. The domain can not be modified or purged; it can only be restored. Any other registrar requests to modify or otherwise update the domain will be rejected. The domain will be held in this status for a maximum of 30 calendar days.

PENDINGDELETE:

The registry sets this status after a domain has been set in REDEMPTIONPERIOD status and the domain has not been restored by the registrar. The domain will not be included in the zone. Once in this status all registrar requests to modify or otherwise update the domain will be rejected. The domain will be purged from the registry database after being in this status for 5 calendar days.



The US Census Bureau has the following three lists of names, one for [male](#) and one for [females](#) and one for [last names](#)

First names

The collection from Census contains 1219 male names, two of which are in a NO NAMESERVER STATE, but none that are not registered. Looking at a list of norwegian names, currently holding 427 names. The following list shows as available at the present time.

Name	DNS Status
------	------------

TORKIL AVAILABLE

TOR-ERIK AVAILABLE

The collection from Census contains 4275 female names. Of which the following might be available.

Name	Frequency in percents	Rank	DNS Status
EARLEAN	0.002	2542	REDEMPTIONPERIOD
YULANDA	0.001	3188	REDEMPTIONPERIOD
DELORSE	0.001	3659	AVAILABLE
JOHNSIE	0.001	3715	REDEMPTIONPERIOD
LAKEESHA	0.001	3798	PENDINGDELETE

Looking at some norwegian naming list, with 515 names listed gives the following names:

Name	DNS Status
INGERID	AVAILABLE
MARTHINE	AVAILABLE
YNGVILD	AVAILABLE
ALVILDE	AVAILABLE

Other data

270,335 COM domain names contains the word "sex", of which it is the beginning phrase for 88,001 domain names.

16,653 domain names is prefixed with "ilove" and 742 is prefixed "i-love".

Finding your own domain name

A lot of domain names are obviously unavailable already, but this doesn't mean there aren't great names out there still. That said, quoting the famous balcony scene from Shakespeare's Romeo and Juliet:

What's in a name? That which we call a rose
By any other word would smell as sweet.

The domain name isn't the primary method of getting traffic today due to the large amount of total domain names. Direct links or direct references, such as business cards is more common.

With the exception of situations where you depend on a prospect to remember the domain name, such as a TV advertisement, you can easily go for a slightly longer domain name.

You can see if your domain name is available, and possibly register it at passive12.net



Related articles: