

# N-gram Character Sequence Analysis of Benign vs. Malicious Domains/URLs

Mike Geide Zscaler, Inc.

## Introduction:

A past colleague and I had discussed character based analysis of domains to provide indications of suspicious / malicious usage. For example, the Avalanche Group (formerly Rock Phish) has a domain registration and fast-flux infrastructure that is frequently used for hosting Zeus, phishing, and money-mule recruitment sites. Typically, this group bulk registers domains that often do not have any correlation to any particular word or alias, for example,

erasv	yxeedlrp	zah7kio
jkkk	yy1azsva	zkmasl
yuuikom	yyy1zsve	zuh7kio
yuvtzc	yyyazsvd	zyuojli
yuxlijz	zaaaasaa	zzztyreb

The theory is that by conducting character based analysis on malicious and benign domains, indicators would emerge that if certain patterns were present or absent it would be a domain worth investigating further. However, there may be a problem with this theory since there are a number of legitimate domains that have no apparent relationship to linguistic patterns and a number of malicious domains that do.

In addition to investigating character sequences of domains, the same theory may apply to full URL strings. This paper investigates the practicality of character-based analysis of domains and URLs to find patterns that may differentiate between malicious and benign sites, and provides some preliminary results.

## Methodology:

Zscaler, Inc. is a Software as a Service (SaaS) vendor and is in a unique position to observe a large number of web transactions across customers from around the world. In addition to content inspection, Zscaler also consumes a number of data-feeds to block against transactions to known malicious IPs, domains, or URLs.

Using a technique called N-gram analysis<sup>1</sup>, it is possible to very quickly extract sequences of characters from text and calculate the frequency of sequence occurrences within the text.

---

<sup>1</sup> <http://en.wikipedia.org/wiki/N-gram>

In order to have an adequate sample size for this analysis, I used a list of 250,000 malicious URLs and 250,000 benign URLs. I used the CPAN module `Text::Ngrams`<sup>2</sup> within Perl scripts to calculate the frequency of sequence occurrences.

## Results:

The overall top results for each character sequence was somewhat interesting, but for the most part the top results were similar across both malicious and benign results. This is likely due to common domain/URL character usage (TLDs, common HTML directories, common HTTP strings, etc.). For those interested, the top 25 results for sequences of 1 to 6 characters for malicious and benign domains and URLs can be seen in Appendix A.

What was more interesting were the character sequences that had the most significant frequency differences between malicious and benign domains/URLs.

### 1-Gram Results:

Results of ngram-1 analysis on the malicious and benign domains show that these characters are more popular within malicious domains than within benign domains:

-	18581
h	18386
r	14622
p	12684
j	7100
u	6834
z	4200
w	2660
y	928
q	452

The second column is the number of instances more that the character was seen in malicious domains than benign domains. While these exact scores will vary based upon the malicious and benign samples used, these characters are likely less popular characters in words or aliases, meaning that they stand out a bit within the malicious / slightly random domain character choices.

### 2-Gram Results:

The two-character sequence occurrence frequency between our malicious and benign samples did vary, though there is not a solid explanation for the top results:

te	14566
on	12649
os	12055
ch	11902
to	9529

---

<sup>2</sup> <http://search.cpan.org/~vlado/Text-Ngrams/Ngrams.pm>

po	8667
or	8396
wa	8064
it	7481
uo	7371

However, there were a few results within the top 25 that did have explanations related to TLD usage within malicious versus benign domains, for example,

nl	7067
fr	5261
ru	5247

### 3-Gram Results:

Similar to the results from the 1-Gram analysis, there were character sequences that appear to be less common in words or aliases that rise to the top within malicious domains. The below difference table indicates the top 10 three character sequences that only occur with malicious domains, and the top 10 frequency differences:

Not in Benign		Difference	
uon	7410	ine	10859
huo	7352	lin	9125
aev	3349	nli	8788
50m	2260	onl	8488
gnh	2025	tch	7501
faj	1064	chu	7489
joj	1061	atc	7390
bdj	1027	wat	6986
fds	1001	ost	6290
dfd	998	ste	5956

Some of the character patterns appear to be related to the QWERTY keyboard, for example, “fds” and “dfd.”

As expected, some of the X-rated themed three character sequences rank higher with malicious domains:

sex	2372
xxx	481

There are also some interesting three-character sequences that rank much higher with legitimate URLs, such as “:443” and common web file extensions (CSS, GIF, PNG, JPG, etc.).

### 4-Gram Results:

A few word differences start to appear between malicious and benign domains when analyzing four character sequences.

	<b>Not in Benign</b>		<b>Difference</b>
huon	7346	line	8844
uonl	7317	nlin	8609
tchu	7316	onli	8582
chuo	7315	atch	7404
aeve	3346	watc	7095
naev	3338	ento	3544
gnae	3337	oste	3508
porn	3198	ntos	3478
quet	2321	even	3407
150m	2228	stec	3382

Specifically, the words:

- porn
- online
- watch

Some additional word sequences are more common in malicious than benign domains:

free	2773
host	2713
game	2033
scan	1581
anti	1371
evil	1037
viru	860
irus	860
best	816
adul	762
dult	761

Many of these words appear to be related to fake anti-virus / fake codec malware sites.

There are also a number of QWERTY sequences that are more prevalent within malicious domains:

dfds	982
fdsi	980
jdfd	980
kdsf	681
mkds	680
dsfn	680
fnsd	680
sfns	680

### **5- and 6-Gram Results:**

Many of the five and six character sequences build out the patterns that we began to see emerge within the 4-Gram results:

<b>Not in Benign</b>		<b>Difference</b>	
free-	652	online	8523
funpic	605	watch	7094
spywar	427	forum	3236
pyware	427	rchive	1223
-sell-	470	archiv	1212
		gamer	860
		virus	857
		phone	827
		adult	761
		ntivir	616
		antivi	615
		music	573
		hostin	476
		osting	475
		videos	475
		ultima	446
		movie	430
		securi	412
		scanne	408
		ecurit	408
		canner	404
		system	363

### **Conclusions:**

Many of the character sequences that were more popular with malicious domains/URLs where actual words, and many of these words seemed to correspond with fake anti-virus / codec sites. There were also several examples of QWERTY keyboard patterns used for domain strings.

N-gram character heuristics can be used on sufficiently large sample sizes to extract character sequences that are more likely to occur within malicious than benign domains/URLs. Understanding and recognizing these character patterns could assist in page risk indexing and/or prioritizing offline analysis.

### **Future Work:**

Take large samples of domains/URLs hosting specific threats (e.g., Zeus, Conficker, Koobface, etc.) to compare against benign domains/URLs. This may provide more specific character patterns for malicious sites.

Create a dictionary of regular expressions for extracting specific URLs out of a large sample size for offline analysis. Tweak the dictionary to limit the benign to malicious URL ratio.

Perform N-gram analysis of the actual content from the URL to detect threats through textual patterns, such as those present within exploit kits, command and control (C&C) configuration files, and other shared exploitation and obfuscation routines.

## Appendix: Top 25 Results

### Top 25 1-Grams

Benign Domains		Malicious Domains		Benign URLs		Malicious URLs	
Gram	%	Gram	%	Gram	%	Gram	%
o	10.6212	e	9.8961	e	4.9545	e	7.8673
e	9.1480	a	8.8525	a	3.9774	a	6.5640
c	8.6476	o	7.9437	t	3.9332	o	6.3726
m	7.4518	i	6.7860	o	3.5850	i	5.5937
a	6.5515	s	6.3064	s	3.4976	t	5.4885
i	5.7019	t	6.2721	/	3.4552	s	5.1077
t	5.4337	n	6.1222	=	3.4314	m	4.8274
s	5.0663	r	5.9377	i	3.4081	/	4.8117
l	5.0544	c	4.5674	1	3.4072	n	4.8052
n	5.0161	l	4.5273	0	3.2895	r	4.5064
g	4.0335	m	3.7763	c	3.2295	c	4.2674
d	3.8891	u	3.2981	m	2.8696	l	3.9699
r	3.5667	d	3.1263	2	2.7861	w	3.6526
b	2.0479	p	2.6999	n	2.7806	p	2.8922
u	2.0150	g	2.5534	r	2.6531	h	2.8689
k	1.6133	h	2.5340	l	2.6125	d	2.8633
p	1.4630	b	2.0085	3	2.5202	-	2.7248
1	1.3710	f	1.6985	d	2.3614	g	2.3392
h	1.2025	y	1.5152	&	2.1115	u	2.3180
f	1.1518	-	1.4566	6	2.0919	b	1.7544
v	1.1091	w	1.3918	p	1.9377	f	1.4070
y	0.9839	v	1.3737	g	1.8363	y	1.0565
w	0.8562	k	1.0280	8	1.8106	=	1.0260
2	0.8551	x	0.5832	4	1.8034	k	0.9606
0	0.7549	j	0.5636	5	1.7961	_	0.9081
Total	95.606	Total	96.819	Total	72.140	Total	90.953

### Top 25 2-Grams

Benign Domains		Malicious Domains		Benign URLs		Malicious URLs	
Gram	%	Gram	%	Gram	%	Gram	%
co	6.0739	in	1.7027	co	1.0199	ww	1.8810
om	5.6091	er	1.5747	om	0.9024	co	1.6382
le	2.2569	on	1.3839	1	0.8868	om	1.4735
ne	1.9738	te	1.2474	m/	0.6409	in	1.3894
et	1.9115	or	1.1899	ww	0.6245	es	1.1822
oo	1.5842	es	1.1832	0	0.5957	er	1.0724
ad	1.5374	ne	1.1733	in	0.5418	ht	1.0707
ic	1.4838	co	1.1568	er	0.5411	tm	1.0655
li	1.4098	st	1.1516	AA	0.5327	or	1.0527
og	1.3465	an	1.1459	le	0.5270	m/	1.0285

go	1.3229	ch	1.0796	12	0.5152	s/	1.0141
gl	1.2281	ar	1.0258	ad	0.5029	ml	0.9409
er	1.0989	li	0.9905	es	0.4936	ma	0.9345
in	1.0851	at	0.9661	re	0.4526	on	0.9102
es	0.9481	al	0.9592	s/	0.4358	ne	0.8688
ti	0.8709	re	0.9554	ne	0.4292	te	0.8077
ve	0.8688	om	0.9209	on	0.4215	de	0.7669
ec	0.8636	to	0.8457	en	0.4146	an	0.7163
at	0.8589	os	0.8179	li	0.4115	re	0.7054
cl	0.8433	ma	0.8078	10	0.4115	ag	0.6761
ck	0.8247	en	0.7998	tm	0.4096	ar	0.6699
ma	0.8011	se	0.7921	ut	0.4039	st	0.6630
im	0.7514	ra	0.7432	at	0.3939	ge	0.6522
an	0.7448	ho	0.6970	ar	0.3891	at	0.6271
al	0.7103	le	0.6887	et	0.3799	en	0.6013
Total	39.008	Total	25.999	Total	13.278	Total	24.409

### Top 25 3-Grams

Benign Domains		Malicious Domains		Benign URLs		Malicious URLs	
Gram	%	Gram	%	Gram	%	Gram	%
com	7.2653	ine	0.6585	com	0.8093	com	1.3970
net	1.8285	com	0.6553	om/	0.6476	htm	1.1152
goo	1.6004	lin	0.6135	AAA	0.4638	om/	1.0452
gle	1.5993	nli	0.5111	www	0.3204	www	1.0163
oog	1.5975	onl	0.5053	utm	0.2870	tml	1.0008
ogl	1.5967	tch	0.4108	age	0.2693	es/	0.5775
cli	1.0077	atc	0.4036	12	0.2271	age	0.5487
ick	0.9625	chu	0.3860	126	0.2176	mag	0.4508
ads	0.7966	ost	0.3843	ads	0.2141	ima	0.4115
lec	0.7417	ter	0.3819	0	0.1978	ges	0.3974
lic	0.7393	wat	0.3795	id=	0.1916	php	0.3851
ble	0.7205	uon	0.3753	268	0.1910	_im	0.3426
ubl	0.7126	huo	0.3724	&ut	0.1863	/_i	0.3362
dou	0.7062	por	0.3585	ent	0.1850	net	0.3188
oub	0.7037	ste	0.3522	686	0.1779	ine	0.2912
ecl	0.7023	ent	0.3492	ttp	0.1777	ion	0.2859
ive	0.6436	for	0.3123	htt	0.1769	ind	0.2793
dia	0.5595	nto	0.2797	=ht	0.1641	/in	0.2786
liv	0.5559	ing	0.2637	/ad	0.1614	id=	0.2761
ati	0.4977	ion	0.2595	863	0.1607	for	0.2753
tic	0.4743	tec	0.2455	net	0.1601	lin	0.2726
cdn	0.4553	sta	0.2432	cli	0.1578	et/	0.2667
nts	0.4345	log	0.2421	tp:	0.1552	nde	0.2595
ent	0.4318	ech	0.2386	p:/	0.1542	ing	0.2377
age	0.4220	nce	0.2374	gle	0.1532	web	0.2323
Total	27.755	Total	9.419	Total	6.207	Total	12.298

### Top 25 4-Grams

Benign Domains		Malicious Domains		Benign URLs		Malicious URLs	
Gram	%	Gram	%	Gram	%	Gram	%
goog	2.1916	line	0.5968	com/	0.6732	com/	1.1353
ogle	2.1916	nlin	0.5747	AAAA	0.4379	html	1.0880
oogl	2.1916	onli	0.5733	&utm	0.1938	ages	0.4240
lick	0.9925	atch	0.4591	1268	0.1911	imag	0.4238
clic	0.9816	watc	0.4327	http	0.1841	mage	0.4233
doub	0.9663	huon	0.4285	2686	0.1766	ges/	0.4006
uble	0.9658	uonl	0.4268	ttp:	0.1617	/_im	0.3634
oubl	0.9657	tchu	0.4267	tp:/	0.1606	_ima	0.3269
ecli	0.9631	chuo	0.4267	6863	0.1602	inde	0.2480
lecl	0.9567	tech	0.2235	=htt	0.1555	net/	0.2422
blec	0.9565	foru	0.2206	oogl	0.1508	php?	0.2414
live	0.7628	orum	0.2204	ogle	0.1508	ndex	0.2409
mliv	0.4794	ento	0.2170	p://	0.1500	line	0.2297
stat	0.4657	oste	0.2098	goog	0.1445	/ind	0.2264
serv	0.4570	ntos	0.2091	mage	0.1421	onli	0.2158
tati	0.4031	site	0.2086	imag	0.1338	nlin	0.2155
edia	0.3937	vent	0.2042	net/	0.1306	tion	0.1936
ient	0.3835	tion	0.2036	ages	0.1270	rama	0.1608
ents	0.3826	even	0.2028	126	0.1261	ebor	0.1565
lien	0.3822	post	0.2023	/www	0.1187	bora	0.1560
mail	0.3816	stec	0.1981	0000	0.1174	oram	0.1556
clie	0.3810	magn	0.1965	html	0.1094	webo	0.1553
medi	0.3771	agna	0.1958	/ads	0.1028	&web	0.1550
atic	0.3761	aeve	0.1952	/ima	0.0981	a=-1	0.1550
erve	0.3744	naev	0.1947	sear	0.0948	ama=	0.1550
Total	20.323	Total	7.647	Total	4.392	Total	7.888

### Top 25 5-Grams

Benign Domains		Malicious Domains		Benign URLs		Malicious URLs	
Gram	%	Gram	%	Gram	%	Gram	%
googl	2.7666	nline	0.6632	AAAAA	0.4115	image	0.4534
oogle	2.7666	onlin	0.6624	12686	0.1823	ages/	0.4293
click	1.2388	watch	0.5025	http:	0.1674	mages	0.4180
doubl	1.2191	uonli	0.4958	ttp:/	0.1663	_imag	0.3507
ouble	1.2191	chuon	0.4956	26863	0.1652	/_ima	0.3434
eclic	1.2081	atchu	0.4956	=http	0.1610	index	0.2549
lecli	1.2077	tchuo	0.4956	oogl	0.1562	/inde	0.2353
blecl	1.2075	huonl	0.4956	tp://	0.1552	nline	0.2302
ublec	1.2075	forum	0.2556	googl	0.1491	onlin	0.2298
mlive	0.6052	entos	0.2420	image	0.1383	ebora	0.1667
stati	0.5073	event	0.2304	1268	0.1215	orama	0.1666
clien	0.4810	stech	0.2298	mages	0.1196	webor	0.1663
lient	0.4810	poste	0.2288	/imag	0.1015	rama=	0.1663
media	0.4732	ostec	0.2278	ages/	0.0963	ma=-1	0.1663
ients	0.4732	vento	0.2278	68637	0.0951	boram	0.1663
serve	0.4718	aeven	0.2265	click	0.0909	ama=-	0.1663
tatic	0.4690	magna	0.2262	00000	0.0800	&webo	0.1663

fbcdn	0.4653	agnae	0.2261	//www	0.0782	forum	0.1509
india	0.4582	gnaev	0.2261	p://w	0.0776	a=-1&	0.1504
ents1	0.4279	naeve	0.2261	://ww	0.0773	s/_im	0.1178
icket	0.3953	esign	0.1912	com/i	0.0764	backu	0.1107
ricke	0.3894	desig	0.1909	earch	0.0706	ackup	0.1107
crick	0.3894	utler	0.1817	lient	0.0700	/back	0.1051
image	0.3848	butle	0.1816	clien	0.0679	watch	0.1038
ketne	0.3805	orumb	0.1800	searc	0.0614	atchu	0.0985
Total	21.293	Total	8.005	Total	3.136	Total	5.224

### Top 25 6-Grams

Benign Domains		Malicious Domains		Benign URLs		Malicious URLs	
Gram	%	Gram	%	Gram	%	Gram	%
google	3.6219	online	0.7807	AAAAAA	0.3900	images	0.4435
double	1.5960	uonlin	0.5857	http:/	0.1720	mages/	0.4419
eclick	1.5815	huonli	0.5855	126863	0.1708	_image	0.3721
leclic	1.5810	atchuo	0.5855	ttp://	0.1606	/_imag	0.3644
ublecl	1.5808	watchu	0.5855	=http:	0.1548	/index	0.2467
blecli	1.5808	tchuon	0.5855	google	0.1543	online	0.2435
oublec	1.5808	chuonl	0.5855	12686	0.1235	orama=	0.1765
client	0.6296	ostech	0.2691	images	0.1180	&webor	0.1765
lients	0.6195	postec	0.2685	/image	0.1050	rama=-	0.1765
static	0.6140	aevent	0.2674	268637	0.0984	ama=-1	0.1765
ients1	0.5601	naeven	0.2671	mages/	0.0955	webora	0.1765
cricke	0.5097	magnae	0.2671	tp://w	0.0803	borama	0.1765
ricket	0.5097	evento	0.2671	://www	0.0797	eboram	0.1765
etnext	0.4980	gnaeve	0.2671	p://ww	0.0788	ma=-1&	0.1596
icketn	0.4980	agnaev	0.2671	client	0.0702	backup	0.1175
cketne	0.4980	ventos	0.2671	search	0.0635	s/_ima	0.1097
ketnex	0.4980	design	0.2255	268636	0.0610	/backu	0.1092
alytic	0.4195	butler	0.2145	double	0.0601	uonlin	0.1045
lytics	0.4195	forumb	0.2127	eclick	0.0595	watchu	0.1045
analyt	0.4194	orumbu	0.2125	000000	0.0595	tchuon	0.1045
nalyti	0.4194	mbutle	0.2125	leclic	0.0594	huonli	0.1045
=-							
analy	0.4133	umbutl	0.2125	oublec	0.0594	atchuo	0.1045
ogle-a	0.4122	rumbut	0.2125	ublecl	0.0594	chuonl	0.1045
oogle-	0.4122	sitese	0.1703	blecli	0.0594	ackup/	0.1010
gle-an	0.4122	itesee	0.1697	com/im	0.0547	/onlin	0.0922
Total	21.885	Total	8.544	Total	2.648	Total	4.664