Domain Knowledge: Predicting the Kind of Content Hosted by a Domain*

Suriyan Laohaprapanon† Gaurav Sood‡

November 10, 2018

NB: Preliminary draft. Please do not cite without permission.

Abstract

In a broad set of domains, from protecting people from harmful or illegal content to segmenting online customers, we need to know the kind of content hosted by a web domain. But there are nearly 2 billion unique hostnames today. And it is a challenge to know the type of material hosted by each domain. Curated public lists carry at best a few million domains. And commercial APIs are opaque. We address the problem by exploiting labeled data from multiple large curated lists—`Shallalist`, `PhishTank`, `Malware Domains`, and `Squidguard`—to build models that predict the kind of content hosted by a domain using the sequence of characters in the domain name. Given learning which domains carry harmful material and adult content is often particularly useful, we primarily focus on those categories. Our models do very well at predicting domains that host pornographic content, with f1-scores of about .9 or higher. We are less successful at predicting domains that carry harmful content with f1-scores of two of our best models around .8. To illustrate the utility of our models, we use them to answer two compelling questions: 1. Do poor people, minorities, and the less well-educated visit malware sites more often than their respective complementary groups, and 2. Does the consumption of pornography vary by age and education?

*Data and scripts behind the analysis presented here can be downloaded from http://github.com/themains/pydomains and http://github.com/themains/domain_knowledge. The python package that implements the method discussed in the paper is available at: http://github.com/themains/pydomains.

†Suriyan can be reached at: suriyant@gmail.com

‡Gaurav can be reached at gsood07@gmail.com

In a broad set of domains, from cybersecurity to keeping adult content out of kids' reach to segmenting online customers, we need to know the kind of content hosted by a domain. But learning about the kind of content hosted by a domain is not straightforward. For one, there are nearly 2 billion hostnames on the web today (Netcraft 2018). And curated lists tend to have at best a few million domains. Relying on commercial services also seems unwise given that they are opaque, and the quality of their outputs is unknown. In this paper, we introduce a new way of inferring the type of content hosted by a domain and use it to answer two compelling questions.

We exploit labeled data from multiple large curated lists—`Shallalist`, `PhishTank`, `Malware Domains`, and `Squidguard`—to build models that predict the kind of content hosted by a domain using the sequence of characters in the domain name. Given the importance of correctly classifying domains that carry harmful material and adult content, we primarily focus on those categories. Our models do very well at predicting domains that host pornographic content, with f1-scores of about .9 or higher. We are less successful at predicting domains that carry harmful content with f1-scores of two of our best models around .8. To illustrate the utility of our models, we use them to answer two interesting questions. First, do poor people, minorities, and the less-well-educated visit sites that distribute malware or engage in phishing more frequently than their respective complementary groups—the better-off, the racial majority, the better educated? Second, how does the consumption of pornography vary by education and age?

Data and Model

We exploit data from Shallalist (KG 2017), PhishTank (OpenDNS 2017), Toulouse/SquidGuard (Prigent 2017), Malware Domains (RiskAnalytics 2017), and Alexa Top 1M Domains (Amazon 2017) to build models to predict the kind of content hosted by a domain based on the sequence

of characters in the domain name.

For all lists, we start by getting the hostname from the domain name. We then pre-process each list slightly differently based on the data in each list. For Phishtank and Malware Domains list for which we have no negative class labels, we use the list of popular domains from Alexa Top 1M to build the negative class. One virtue of using popular domain list for building the negative class is that it builds classifiers that are sensitive to the skew in Internet consumption. The other reason why we think using popular URLs is a particularly good idea is that harmful websites (Phishing and Malware websites) often try to dissemble as popular websites. For instance, there are over a couple of hundred PhishTank URLs that have the word 'paypal' in them. In particular, we use 50,000 unique domains from PhishTank for 2016–2017 and pair it with the top 50,000 most visited domains from the 1M Alexa domain list. For Malware Domains, we do the same except we don't take a sample from Malware Domains data as there are just 15,238 domains on the list.

For Shallalist and Toulouse/Squidguard, we filter out domains that are assigned multiple categories. We also filter out categories with fewer than 1,000 domains. We fit a model to these data (details below) and based on the model remove categories where the recall is less than about .3—suggesting categories in which there is little systematic pattern to the domain names based on the kinds of patterns our model can detect. For Shallalist, this leaves us with 29 categories (see Table 1). For Toulouse/Squidguard, it leaves us with 8 categories (see Table 2). We consign rest of the domains to the 'others' category.

To learn the association between the sequence of characters in domain names and the kind of content they host, we use LSTM (Graves and Schmidhuber 2005; Gers, Schmidhuber and Cummins 1999). For our models, we follow the same basic workflow. We split the strings (domain name) into two character chunks (bi-chars). For instance, yahoo.com becomes `ya, ah, ho, oo, o., .c, co, om`. Next, we pad the sequences so that they are the same size. Finally, we use 128 as the window size.

Table 1: Number of unique domains by category in the Shallalist Dataset

category	n
adv	12,712
anonvpn	6,981
downloads	4,177
dynamic	1,066
education/schools	10,068
finance/banking	4,989
finance/insurance	3,081
finance/moneylending	3,802
finance/realestate	1,379
fortunetelling	1,077
forum	8,058
gamble	13,827
hobby/games-online	13,861
hobby/pets	16,164
hospitals	1,637
jobsearch	4,294
movies	5,558
music	8,918
others	110,991
porn	827,444
radiotv	3,560
recreation/restaurants	1,408
recreation/sports	120,426
recreation/travel	138,943
redirector	29,366
religion	9,189
science/astronomy	1,035
sex/lingerie	1,056
shopping	167,262
webmail	3,525
webradio	2,254

On this set of sequences, we train a LSTM model using Keras (Chollet et al. 2015) and TensorFlow (Abadi et al. 2016). Before estimating the LSTM model, we embed each of the words onto a 32 length real-valued vector. We then estimate a LSTM with a .2 dropout and .2 recurrent dropout for regularization (Srivastava et al. 2014). The last layer is a dense layer with a softmax activation. Because it is a classification problem, we use log loss as the loss

Table 2: Number of unique domains by category in the Toulouse Dataset

category	n
adult	187,0741
bank	1,689
gambling	1,012
games	9,357
malware	4,463
others	21,441
phishing	62,712
press	4,410
shopping	36,331

function. And we use ADAM for optimization (Kingma and Ba 2014). We fit the models for 15 epochs with a batch size of 32. (For Toulouse/Squidguard we end after 5 epochs because we see no improvement after that.)

Table 3 presents some metrics that shed light on how well we did with predicting Malware sites using the Malware Domains data. The OOS precision is .84, recall is .85, and f1-score, the harmonic mean of precision and recall, is .84.

Table 3: OOS Performance of the Malware LSTM Model

malware or not	precision	recall	f1-score	support
0	0.87	0.95	0.91	10,000
1	0.75	0.52	0.62	3,048
avg / total	0.84	0.85	0.84	13,048

Moving to Phishtank 2017 data, all the metrics are about 5% worse. As Table 4 shows, the OOS precision, recall, and f1-score is .80 each.

Table 4: OOS Performance of the PhishTank LSTM Model

phishing or not	precision	recall	f1-score	support
0	0.78	0.84	0.81	10,000
1	0.83	0.76	0.79	10,000
avg / total	0.80	0.80	0.80	20,000

For Shallalist, the commensurate metrics are .76, .77, and .76 respectively (see Table 5).

There is sizable variation in recall across different categories. For instance, recall is .93 for pornography and just .36 for fortune telling. For Toulouse/Squidguard, with much fewer categories than Shallalist, things look vastly better (see Table 6). The average accuracy, recall, and f1-score is .95, .96, and .95 respectively. A closer-inspection, however, suggests that the gains are largely driven by the largest category of adult content. In all, we are able to learn a very good adult domain classifier.

Table 5: OOS Performance of the Shalla LSTM Model

categories	precision	recall	f1-score	support
adv	0.83	0.41	0.55	2,542
anonvpn	0.79	0.70	0.75	1,396
downloads	0.56	0.42	0.48	835
dynamic	0.80	0.54	0.64	213
education/schools	0.84	0.79	0.82	2,014
finance/banking	0.77	0.54	0.63	998
finance/insurance	0.93	0.82	0.87	616
finance/moneylending	0.92	0.79	0.85	760
finance/realestate	0.61	0.44	0.51	276
fortunetelling	0.68	0.36	0.47	215
forum	0.76	0.77	0.77	1,612
gamble	0.79	0.76	0.78	2,765
hobby/games-online	0.66	0.49	0.56	2,772
hobby/pets	0.62	0.39	0.48	3,233
hospitals	0.83	0.72	0.77	327
jobsearch	0.85	0.45	0.59	859
movies	0.72	0.40	0.51	1,112
music	0.84	0.85	0.85	1,784
others	0.48	0.32	0.39	22,198
porn	0.86	0.93	0.89	165,489
radiotv	0.62	0.45	0.52	712
recreation/restaurants	0.75	0.29	0.42	282
recreation/sports	0.64	0.64	0.64	24,085
recreation/travel	0.73	0.64	0.68	27,789
redirector	0.82	0.67	0.74	5,873
religion	0.87	0.83	0.85	1,838
science/astronomy	0.70	0.84	0.76	207
sex/lingerie	0.57	0.22	0.32	211
shopping	0.53	0.63	0.58	33,453
webmail	0.77	0.56	0.65	705
webradio	0.51	0.39	0.44	451
avg / total	0.76	0.77	0.76	307,622

Table 6: OOS Performance of the Toulouse/Squidguard LSTM Model

categories	precision	recall	f1-score	support
bank	0.66	0.57	0.61	338
gambling	0.44	0.28	0.34	202
games	0.80	0.46	0.59	1,871
malware	0.98	0.47	0.63	893
others	0.51	0.26	0.35	4,288
phishing	0.70	0.64	0.67	12,543
press	0.71	0.61	0.66	882
shopping	0.60	0.55	0.57	7,266
avg / total	0.95	0.96	0.95	402,432

Application

You are what you browse. More or less. We jest, just a bit but which sites a person goes to reveals a lot about them. Browsing data can also be used to answer socially consequential questions. For instance, there is widespread concern about the digital literacy gap. There is widespread concern that racial minorities, the less well-educated, and poor are especially likely to visit websites that host harmful information. Using our models, we set out to investigate the question of whether poor people, minorities, and the less-well-educated visit sites that distribute malware or engage in phishing more frequently than their respective complementary groups—the better-off, the racial majority, the better educated.

Separately, there is concern about what consumption of adult content does to attitudes toward women but also more generally about sex. We shed light on how to measure the antecedent variable in the equation: the extent to which people consume adult content, and how that varies by education and age.

We answer these questions using data from comScore. comScore maintains an online panel of approximately 100,000 users. It collects anonymous browsing data on a machine in exchange for small perks. comScore distributes anonymized domain level data (more granular data are withheld because of privacy concerns) at the machine level along with some

household-level characteristics to researchers. We capitalize on these data here.

As is clear, there are at least three problems with the data:

- **Domain level data**: Going to http://nytimes.com is not the same as reading political news. And measurement error may vary by the kind of person. For instance, say we label http://nytimes.com as political news. For the political junkie, the measurement error may be close to zero. For teetolaters, it may be close to 100%. More generally, domain level data mean that we must count all visits to a domain the same. Some domains host heterogeneous content. And people may self-select within domains to 'off-label' content. (We don't think this is a serious concern for pornographic or malware domains.)

- **Machine level data**: these days people browse the Internet on multiple machines. And increasingly the primary machine they browse on is a tablet or a phone. We do not have a lot to say about what proportion of browsing does a person do per machine, and we cannot say whether the browsing behavior (what proportion of time is spent on different websites, etc.) is similar across machines.

- **Household level demographic data**: Where you have more than one person in a household, we cannot cleanly attribute characteristics to a person. For age and education, for instance, comScore gives the age (education) of the oldest (most educated) person in the household. For income, comScore gives the household income.

Our strategy for answering all three questions is roughly the same. We start with comScore data for a year that is already aggregated at the domain level and has four columns: machine id, domain name, number of visits to a domain from people using the machine in a year, amount of time (in minutes, we think) spent on the domain by people on the machine. The data are in long form—as many rows per machine as the total number of unique domains they visit in a year.

We next left join this data using the domain name as key with pre-computed pydomains data about the kind of content hosted by a domain. We then aggregate these data by kind of content (inferred by a particular method) and we sum the visits and time spent by kind of content. We then left join each of these datasets with demographics data using machine id as key. We then do a groupby describe over sociodemographic traits to see how the total time spent and the total number of visits vary by sociodemographic traits.

Browsing Data: Concerns and Solutions

Say you want to measure the how often people visit pornographic domains over some period. To measure that, you build a model to predict whether or not a domain hosts pornography. And let's assume that for the chosen classification threshold, the False Positive rate (FP) is 10% and the False Negative rate (FN) is 7%. Here below, we discuss some of the concerns with using scores from such a model and discuss ways to address the issues.

Let's get some notation out of the way. Let's say that we have n users and that we can iterate over them using i. Let's denote the total number of unique domains—domains visited by any of the n users at least once during the observation window—by k. And let's use j to iterate over the domains. Let's denote the number of visits to domain j by user i by $c_{ij} = 0, 1, 2, ...$ And let's denote the total number of unique domains a person visits $(\sum(c_{ij} == 1))$ using t_i. Lastly, let's denote predicted labels about whether or not each domain hosts pornography by p, so we have $p_1, ..., p_j, ..., p_k$.

Let's start with a simple point. Say there are 5 domains with p: $1_1, 1_2, 1_3, 1_4, 1_5$. Let's say user one visits the first three sites once and let's say that user two visits all five sites once. Given 10% of the predictions are false positives, the total measurement error in user one's score $= 3 * .10$ and the total measurement error in user two's score $= 5 * .10$. The general point is that total false positives increase as a function of predicted $1s$. And the total number

of false negative increase as the number of predicted $0s$. More generally, the total error for user i is (in expectation):

$$\sum_{1}^{k} c_{ij} * (p_j == 1) * (FP) - c_{ij} * (p_j == 0) * (FN) \tag{1}$$

Formalizing clarifies three simple things. First, the net error is a function of $FP - FN$. Second, even when the share of visits to pornographic domains is the same, the larger the number of domains (t_i) a person visits, greater the error in their score (total number of visits to pornographic domains). Third, when c_{ij} are right-skewed, e.g., browsing data, errors in the right tail can be very costly. Concretely, misclassifying domains that people visit a lot can be super expensive—it may even change inferences wholesale.

One way to speak to the first two issues is to use different probability cutoffs for classification. Different probability cutoffs generate different FN and FP rates and allow us a way to provide bounds for the inferences.

The third point cuts deeper. To address the issue, one could tweak the cost function of the domain level model such that the cost of each error is proportional to usage. But given the skew, it would put a metric ton of weight on the features of too few domains. And that may mean that the performance of the model is pretty bad. A better, simpler solution may be to use the labels from the dataset used in training. Labeled datasets like the Shallalist cover a vast majority of the heavily visited domains. And using labels from the training set means that we are saved from the most costly errors. Doing so also means that we aren't doing the silly thing of introducing (adding to) measurement error for cases where we have little measurement error.

If still concerned about errors, one could download the top 1M domains from Alexa, take the difference from the original labeled dataset, and for the remaining domains that are

also in the universe of domains you are analyzing, use some reputable web service to get the category of content hosted by the domain.

Bad Domains

Somewhat surprisingly, the most educated most frequently visit (spend most time on) phishing/malware websites (see here). This is consistent with Cor and Sood (2018), who find that the educated are hacked more often. Part of the reason why the more educated visit harmful websites more is because they are online more often. When we look at the proportion of time spent on harmful websites, the most educated spend slightly less than the less well educated (see Table 7).

Table 7: Proportion of Visits to (Time Spent on) Phishing/Malware Domains by Education

	Phishing Visits	Phishing Time	Malware Visits	Malware Time	T1 malware Visits	T1 malware time
Less than HS	0.016	0.018	0.011	0.013	0.004	0.005
Some college	0.017	0.017	0.010	0.012	0.003	0.003
Associate degree	0.016	0.014	0.009	0.010	0.003	0.003
Bachelor's degree	0.015	0.013	0.009	0.009	0.003	0.003
Graduate degree	0.015	0.010	0.009	0.008	0.003	0.003
Missing	0.019	0.021	0.012	0.016	0.005	0.007

When we split the entire sample by race, Asians and Whites more frequently visit (spend more time on) malware/phishing websites than other racial groups (see here). Again, it seems part of the reason is that Asians/Whites spend more time online (see Table 9).

Table 8: Proportion of Visits to (Time Spent on) Phishing/Malware Domains by Race

	Phishing Visits	Phishing Time	Malware Visits	Malware Time	T1 malware Visits	T1 malware time
Asian	0.015	0.014	0.01	0.01	0.003	0.003
Black	0.018	0.02	0.011	0.015	0.005	0.006
White	0.017	0.017	0.01	0.012	0.003	0.004
Other	0.017	0.018	0.01	0.014	0.004	0.006
Missing	0.031	0.052	0.026	0.052	0	0

When we split by age, we see that the older people more frequently visit (spend most time on) phishing/malware sites (see here). Here there is some evidence that it is because they are choosing worse than younger people (see Table 9).

Table 9: Proportion of Visits to (Time Spent on) Phishing/Malware Domains by Age

	Phishing Visits	Phishing Time	Malware Visits	Malware Time	T1 malware Visits	T1 malware time
18-20	0.015	0.015	0.009	0.01	0.004	0.005
21-24	0.015	0.015	0.01	0.011	0.004	0.005
25-29	0.016	0.016	0.01	0.012	0.004	0.004
30-34	0.016	0.016	0.009	0.011	0.004	0.005
35-39	0.017	0.017	0.01	0.012	0.003	0.004
40-44	0.017	0.018	0.01	0.013	0.003	0.004
45-49	0.017	0.018	0.01	0.013	0.004	0.005
50-54	0.018	0.019	0.011	0.014	0.005	0.006
55-59	0.018	0.018	0.011	0.014	0.004	0.004
60-64	0.017	0.016	0.01	0.012	0.003	0.004
65 and over	0.018	0.017	0.011	0.013	0.003	0.004
Missing	0.006	0.005	0.011	0.002	0.003	0.005

Consumption of Pornographic Content

A consistent pattern emerges across all four versions of our measure: 18–20 visit the pornographic domains the most often but after that, there is a sharp decline and then a modest upward trend peaking at 40–44 after which the average number of visits roughly monotonically decline (see Table 10). You see the same rough pattern in the average time spent as well.

Table 10: Distribution of Number of Visits to Pornographic Sites by Age

	Shallalist (Reduce FN)				Shallalist (Reduce FP)				Shallalist (Reduce FP and FN)				Toulouse (Reduce FP and FN)			
	mean	25%	50%	75%	mean	25%	50%	75%	mean	25%	50%	75%	mean	25%	50%	75%
18-20	397	14	93	341	345	7	55	256	356	8	61	268	379	10	76	317
21-24	309	11	65	245	266	4	34	172	274	5	40	182	294	8	52	227
25-29	331	15	68	248	280	5	30	160	288	6	36	175	315	9	54	220
30-34	324	17	72	247	270	5	30	152	278	7	37	163	306	11	55	219
35-39	356	17	73	256	294	5	31	151	304	7	37	168	335	11	57	232
40-44	345	19	83	283	282	6	38	176	292	8	44	192	324	13	67	255
45-49	341	15	74	285	282	5	32	177	291	6	38	189	321	10	56	254
50-54	352	13	62	242	293	4	24	140	302	5	30	155	331	7	47	209
55-59	314	12	57	207	253	3	19	109	261	5	24	122	292	7	40	178
60-64	246	10	47	184	187	2	14	80	195	3	19	90	224	5	32	151
65 and over	244	10	47	183	183	2	13	80	191	3	17	91	221	6	33	149

Perhaps yet more importantly, it seems the average number of visits are pretty low. We concur. And that means that the absolute size of the differences is pretty small too even though the relative size may look big. The more serious concern is about the underlying browsing data. We don't have a lot to say about it.

As education levels increase, the average number of visits go down (see Table 11). Households where the most educated person in the household has a graduate degree visit pornographic sites less often and spent less time on them than households where the most educated person has less than a HS diploma.

Table 11: Distribution of Number of Visits to Pornographic Sites by Education

	Shallalist (Reduce FN)				Shallalist (Reduce FP)				Shallalist (Reduce FP and FN)				Toulouse (Reduce FP and FN)			
	mean	25%	50%	75%	mean	25%	50%	75%	mean	25%	50%	75%	mean	25%	50%	75%
Less than HS	355	19	92	346	294	7	41	213	305	9	49	233	334	14	74	307
HS or equiv.	316	15	67	245	256	5	28	144	266	6	34	157	295	9	52	213
Some college	327	13	64	238	268	4	25	133	277	5	30	147	306	8	48	204
Associate degree	318	14	64	232	261	4	24	135	269	5	30	146	299	8	48	202
Bachelor's degree	331	11	55	213	274	3	19	117	282	4	24	129	309	6	40	187
Graduate degree	288	9	47	194	234	2	15	93	241	3	19	102	267	4	32	162
Missing	331	16	76	272	271	5	33	167	281	7	39	182	311	10	59	242

Do we see the patterns because it just captures that certain people spend more time online? To check that we look at proportions.

The data are clear—as people get older, they generally spend a smaller proportion of time on pornographic websites with perceptible drop-offs after 50–54. Splitting by education also shows that the declining trend is a result of people in households where education level is higher spending less time on pornographic domains (see Table 12 and Table 13).

Table 12: Proportion of Visits to Pornographic Domains by Education

	Shallalist (Reduce FN)	Shallalist (Reduce FP)	Shallalist (Reduce FP and FN)	Toulouse (Reduce FP and FN)
18-20	0.105	0.089	0.092	0.100
21-24	0.099	0.083	0.086	0.094
25-29	0.097	0.078	0.081	0.091
30-34	0.091	0.072	0.075	0.085
35-39	0.089	0.069	0.071	0.083
40-44	0.090	0.070	0.073	0.084
45-49	0.091	0.071	0.074	0.085
50-54	0.087	0.068	0.070	0.081
55-59	0.079	0.059	0.062	0.073
60-64	0.073	0.052	0.054	0.066
65 and over	0.068	0.048	0.050	0.062

Table 13: Proportion of Visits to Pornographic Domains by Education

	Shallalist (Reduce FN)	Shallalist (Reduce FP)	Shallalist (Reduce FP and FN)	Toulouse (Reduce FP and FN)
Less than HS	0.099	0.080	0.083	0.093
HS or equiv.	0.087	0.068	0.070	0.081
Some college	0.085	0.066	0.068	0.079
Associate degree	0.085	0.066	0.068	0.079
Bachelor's degree	0.082	0.063	0.065	0.075
Graduate degree	0.076	0.057	0.059	0.069
Missing	0.091	0.071	0.073	0.084

Discussion

In this paper, we introduced a new way to learn a model between the sequence of characters in a domain name and the kind of content it carries using multiple curated lists and the Alexa top 1M domains data. We are able to learn very good models for domains that carry adult

content with precision and recall around .9. Our models for classifying websites that carry harmful content do not work as well with f1-score of about .8 but the models are still useful. We illustrate the utility of the models by using them to answer two interesting questions. We also provide a Python package that exposes the models: https://github.com/themains/pydomains/.

The limitations of our models are that for gatekeeping tasks, the accuracy and recall is still too low given the volume of Internet traffic. Ideally, we want to estimate models that combine some content from the domains with the domain name to predict well.

References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard et al. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI*. Vol. 16 pp. 265–283.

Amazon. 2017. "Alexa Top 1M Domains.".

Chollet, François et al. 2015. "Keras.".

Cor, Ken and Gaurav Sood. 2018. "Pwned: How Often Are Americans' Online Accounts Breached?" *arXiv preprint arXiv:1808.01883* .

Gers, Felix A, Jürgen Schmidhuber and Fred Cummins. 1999. "Learning to forget: Continual prediction with LSTM.".

Graves, Alex and Jürgen Schmidhuber. 2005. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." *Neural Networks* 18(5-6):602–610.

KG, Shalla Secure Services. 2017. "Shalla's Blacklists.".

Kingma, Diederik P and Jimmy Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* .

Netcraft. 2018. "January 2018 Web Server Survey.". [Online; accessed 11-November-2018].

OpenDNS, LLC. 2017. "PhishTank: An anti-phishing site.".

Prigent, Fabrice. 2017. "Toulouse/Squidguard Blacklist.".

RiskAnalytics. 2017. "Malware Domains.".

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. 2014. "Dropout: A simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research* 15(1):1929–1958.