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

Suriyan Laohaprapanon† Gaurav Sood‡

June 23, 2020

Abstract

In a broad set of domains, from protecting people from harmful 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 curated domain label lists carry at best a few million domains. We bridge the gap 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 identifying domains that carry harmful material or adult content is particularly important, we primarily focus on those categories. The 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 questions: 1. Do poor people, racial or ethnic 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?

1 Introduction

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 the available solutions—curated lists and commercial services—are problematic. Curated lists are

*Data and scripts behind the analysis presented here are posted at https://github.com/themains/pydomains and https://github.com/themains/domain_knowledge. The python package that implements the method discussed in the paper is available at https://github.com/themains/pydomains.

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

‡Gaurav can be reached at gsood07@gmail.com

limited, carrying, at best, a few million domains. And the quality of popular commercial services is low, with high rates of 'unknown' and incorrect labels (Deri et al. 2015).

Multiple solutions have been proposed to address the limitations of commercial services and curated lists. For instance, Shen et al. (2004) exploit text summarization to classify web pages. Zhang et al. (2010), on the other hand, exploits the topological structure for classification. More recently, Deri et al. (2015) used some of the HTML metadata from a small corpus (5,600 webpages) to classify the content and Wang et al. (2019) used similarity of binary file to classify malicious pages. All of these techniques, however, require significant computational resources—both memory and processing power. Given that there are nearly two billion hostnames on the web today (Netcraft 2018) and given the need to classify domains in a way that preserves privacy, we need computationally lighter methods, especially for classifying harmful and adult content. To address the issue, we extend some of the work that uses lexical features of the URL to classify harmful content (Aldwairi and Alsalman 2012; Jain and Gupta 2018).

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 identifying harmful and adult content, we primarily focus on those categories. Our models accurately predict domains that host pornographic content, with f1-scores of .9 or higher. We are less successful at predicting domains that carry malware, etc., with f1-scores of two of our best models around .8. We compare results from our preferred LSTM models to results from Random Forest and SVC and find that LSTM is consistently superior, with larger area under the curve.

To illustrate the utility of the models, we use them to answer two questions. 1. 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? 2. How does the consumption of pornography vary by education and age?

2 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 that predict the kind of content hosted by a domain based on the sequence of characters in the domain name.

For each dataset, we first extract the hostname from the domain name. For Phishtank and Malware Domains, we have no negative class labels. To build the negative class, we use the most popular domains from Alexa Top 1M. There are two benefits of using popular domains for the negative class. First, it helps build classifiers that are sensitive to the skew in Internet consumption, with most traffic going to a few popular domains. Second, harmful websites (Phishing and Malware websites) often try to dissemble as popular websites. For instance, over two hundred PhishTank URLs 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 the Malware Domains list, we do the same, except given that the list is short, we don't take a sample and use all the 15,238 domains on the list instead.

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. (We did this step on the entire dataset than on the training set alone which means that our final performance is likely worse.) For Shallalist, this leaves us with 29 categories (see Table SI 1.1). For the Toulouse data, it leaves us with eight categories (see Table SI 1.2). We consign the 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`. (We compare the results of our preferred LSTM models to Random Forest and SVC models built on the same set of tokens.) Next, we pad the sequences so that they are the same size. Finally, we use 128 as the window size.

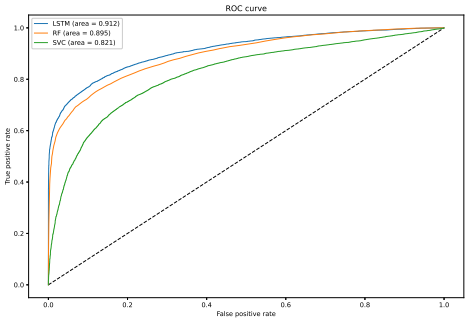
On this set of sequences, we train a LSTM model using Keras (Chollet et al. 2015) and TensorFlow (Abadi et al. 2016). Before estimating the LSTM model, we embed each of the words onto a 32 length real-valued vector. We then estimate a LSTM with a .2 dropout and .2 recurrent dropout for regularization (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 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 the Toulouse, we end after five epochs because we see no improvement after that.)

Table SI 1.3 presents metrics that shed light on how well we did with predicting Malware sites using the Malware Domains data. The weighted OOS precision is .84, recall is .85, and f1-score, the harmonic mean of precision and recall, is .84. Corresponding numbers from a random forest model are .83, .84, and .82. For the linear SVC model, the precision, recall, and f1-score are .82, .83, and .80, respectively. The ROC plot for the Malware models (see Figure 1) shows a more dramatic difference—the performance of the SVC model is a good 13% lower.

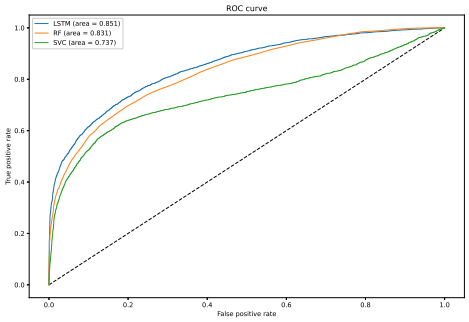
Moving to Phishtank 2017 data, as Table SI 1.4 shows, the weighted OOS precision, recall, and f1-score is .84, .83, and .83 respectively. Random forest and SVC do slightly worse, with weighted f1-scores of .81 and .73. As Figure 1 shows, the area under the curve for the SVC model is about 10% smaller while the Random Forest model's performance is as good as that of the LSTM model.

For Shallalist, precision, recall, and f1-score are .76, .76, and .76, respectively (see Table SI 1.5). There is a sizable variation in recall across categories. For instance, recall is .93 for pornography and just .35 for fortune-telling. Overall, however, Random Forest and SVC do much worse

Figure 1: Malware and PhishTank Model Performance—LSTM, Random Forest, and SVC



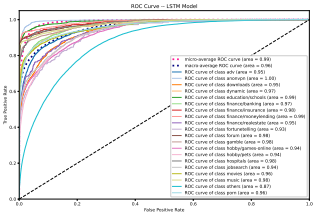
PhishTank



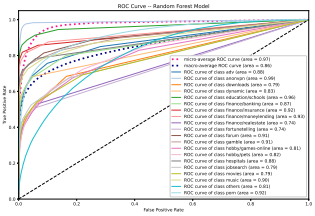
Malware

here with f1-scores of .62 and .04, respectively. (We had to truncate the SVC model after 500 iterations as even after running it for two days, the model never converged.) Figure 2 clearly shows the performance advantage of the LSTM model over the other two models.

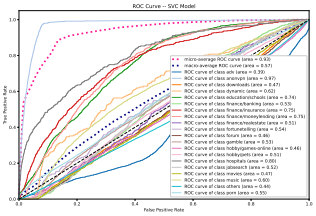
Figure 2: Shallalist Model Performance



LSTM



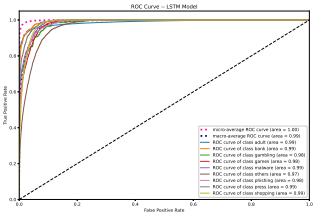
Random Forest



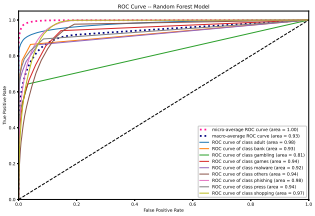
SVC

For the Toulouse data, with much fewer categories than Shallalist, things look vastly better (see Table SI 1.6). The average accuracy, recall, and f1-score are .95, .96, and .95, respectively. A closer inspection, however, suggests that the gains are largely driven by the largest category, which is adult content. In all, we can learn an excellent adult domain classifier. Once again, SVC and Random Forest perform much worse (see Figure 3).

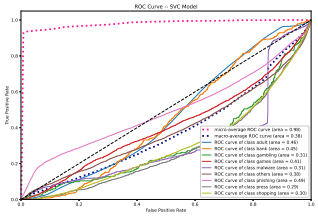
Figure 3: Toulouse Model Performance



LSTM



Random Forest



SVC

3 Application

We use the models developed in the paper to shed light on two important questions. First, are disadvantaged people are more at risk online? In particular, do racial minorities, the less well-educated, and poor visit websites that host harmful content more frequently than their respective counterparts—those who belong to the majority racial group, the better educated, the higher income. Second, there is concern that consumption of adult content negatively affects attitudes and behavior toward women (Hald, Malamuth and Yuen 2010). We shed light on the issue by measuring the extent to which people consume adult content, and how it varies by education and age.

We answer the 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 at the machine level along with some household-level characteristics to researchers. We capitalize on the data here.

Aside from Hawthorne effect, 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 varies by the kind of person. For instance, say we label `http://nytimes.com` as political news. For a political junkie, the measurement error for that class label may be zero. For a political teetotaler, it may be 100%. More generally, domain-level data mean that we count all visits to a domain the same. Some domains host

heterogeneous content. And people may self-select into 'off-label' content.

- **Machine level data**: these days people browse the Internet on multiple machines (Westcott et al. 2019). And increasingly, the primary machine they browse Internet on is a tablet or a phone. We do not have data on what proportion of browsing a person does on the machine from which the data is being collected and we do know whether the browsing behavior, e.g., what proportion of time is spent on different websites, etc., varies 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 the questions is roughly the same. We start with Comscore data for a year that is already aggregated at the machine-domain level and has four columns: machine id, domain name, the number of visits to the domain from people using the machine in a year, amount of time (in minutes, we think) spent on the domain by people using the machine. The data are in long-form—as many rows per device as the total number of unique domains they visit in a year.

We left-join this data using the domain name as key with pre-computed data about the kind of content hosted by a domain. We then aggregate these data by type 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 the key and investigate how the time spent and the number of visits varies by sociodemographic traits.

3.1 Browsing Data: Concerns and Solutions

Say you want to measure how often people visit pornographic domains over some period. To measure that, we build a model to predict whether or not a domain hosts pornography. 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 uncalibrated scores from such a model and ways to address the issues.

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 $p_1, ..., p_j, ..., p_k$.

With the formalization, we can illustrate one point clearly. 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 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, the larger 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 to bound inferences. Another way (and the one we take here) is to calibrate the probabilities so that they reflect the actual likelihood of a true positive. (Calibration means setting FP to FN.)

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 lot of weight on the features of too few domains. And that would likely degrade model performance. 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 saves us from the most costly errors. Doing so also means that we won't be introducing (adding to) measurement error for cases where we have little measurement error.

We can further reduce errors by downloading the top 1M domains from Alexa, taking 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. We adopt this technique when looking at pornographic domains.

Yet another way to make analysis robust to skew is to winsorize usage within users. Winsorization reduces the impact of miclassified heavily visited domains. But the downside is that it adds bias to the usage data.

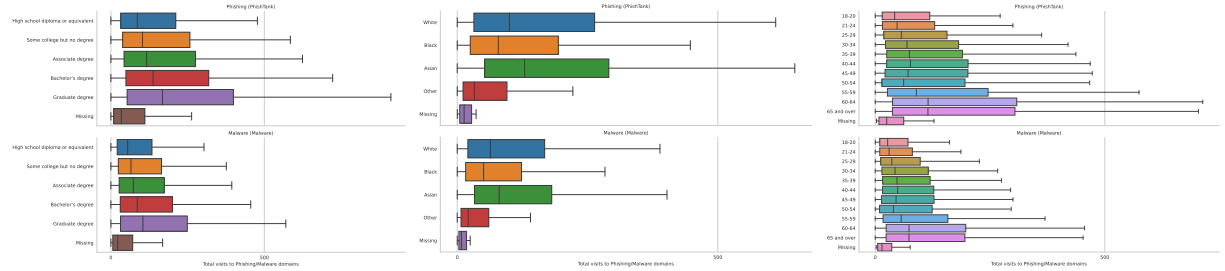
3.2 Bad Domains

Using Comscore data from 2016 and predictions from our LSTM Malware, 2017 PhishTank, and Toulouse models, we assessed whether African Americans, the less educated, and older people

are more at risk online.

Somewhat surprisingly, the most educated most frequently visit (spend most time on) phishing/malware websites (see Figure 4 and Figure SI 1.1). 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, if anything, slightly less than the less well educated (see SI 1.4).

Figure 4: Number of Visits to Phishing/Malware Domains



Education Race Age

When we split the sample by race, we find that Asian and White households more frequently visit (spend more time on) malware/phishing websites than other racial groups. Again, it seems part of the reason is that Asians/Whites spend more time online (see Figure SI 1.7).

Splitting by age, we see that households with older people more frequently visit (spend the most time on) phishing/malware sites. Here there is some evidence that it is because they are choosing worse than younger people, with a slightly larger proportion of their visits going to malware or phishing sites (see Figure SI 1.11).

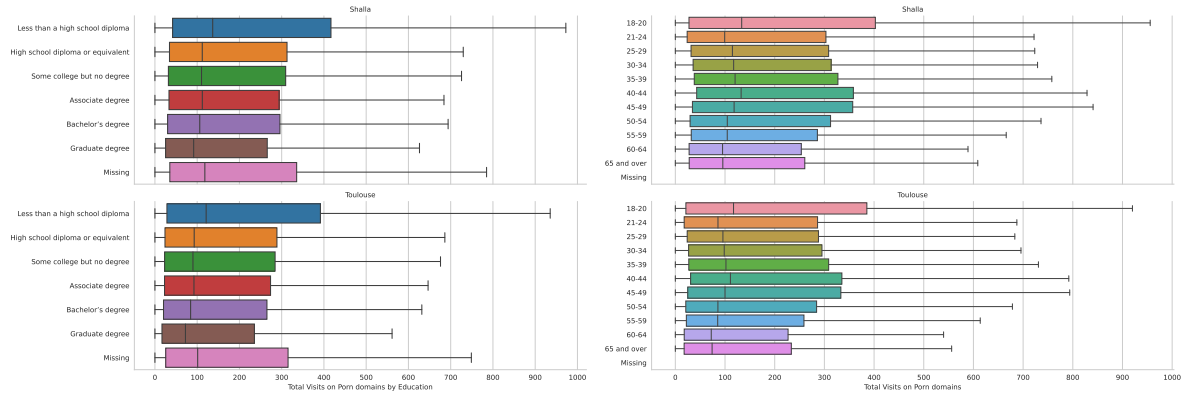
3.3 Consumption of Pornography

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 decline and then a modest upward trend peaking at 40–44 after which the average number of visits roughly monotonically decline (see Figure 5).

You see the same rough pattern in the average time spent as well (see Figure SI 1.9).

As education levels increase, number of visits to pornographic domains go down (see Figure 5). (As Figure SI 1.8 shows, time spent on such sites also decreases.) Households where the most educated person 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 an HS diploma.

Figure 5: Number of Visits to Pornographic Domains



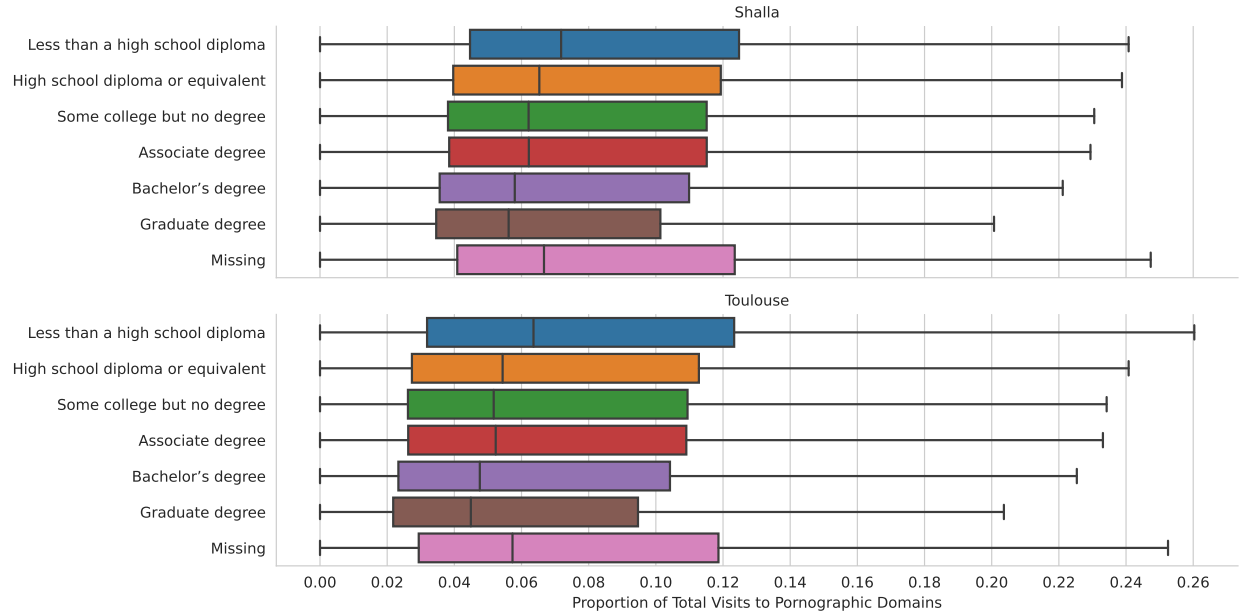
Education Age

Do we see the patterns because it just captures that certain people spend more time online? To check that, we look at the proportions (see Figure 6). The data are clear—as people get older, pornographic websites constitute a smaller share of visits with a roughly monotic decline after 40–44. Splitting by education also shows that the declining trend is a result of people in households where the education level is higher, spending less time on pornographic domains (see Figure 6).

4 Discussion

You are what you browse. Which sites people visit reveals a lot about them and the challenges they face. The sites that people visit also affect their attitudes and behaviors. And in some cases, sites people visit also affect material outcomes. For instance, visiting a website that phishes per-

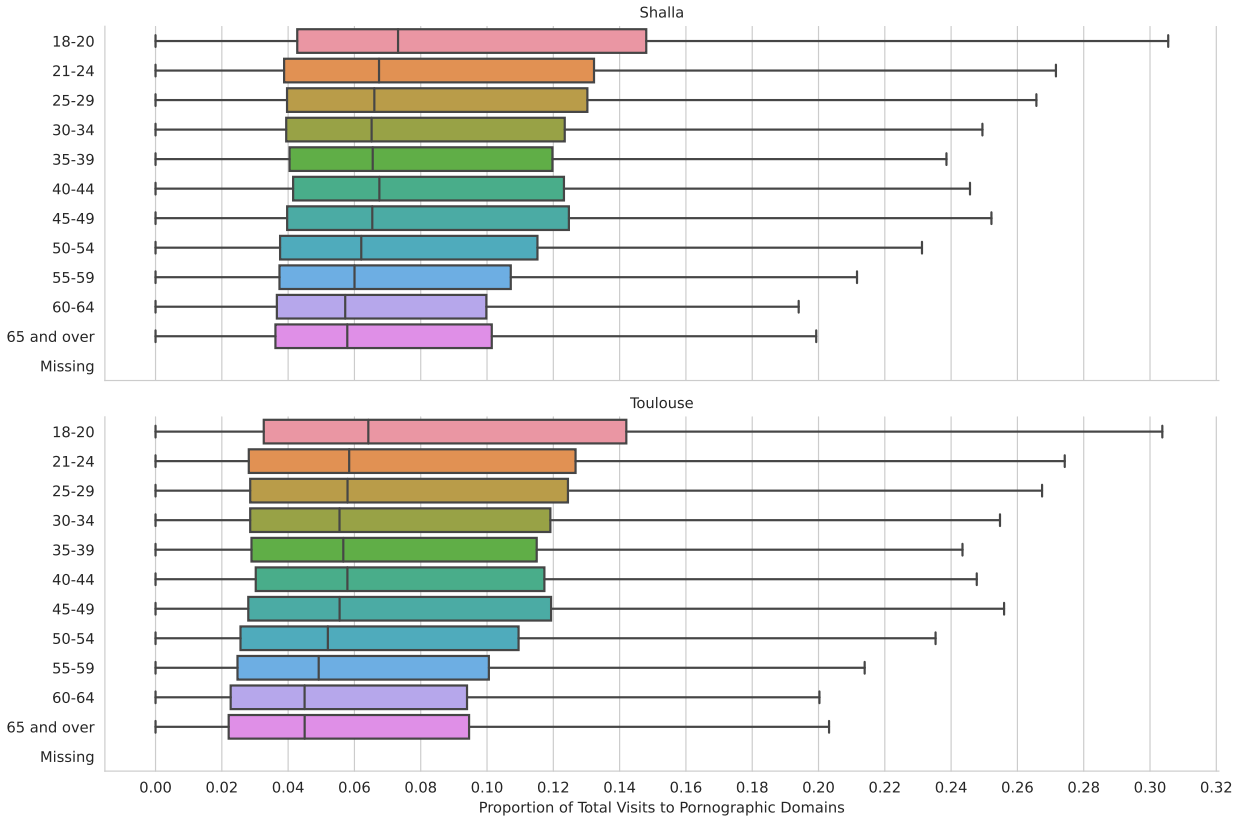
Figure 6: Proportion of Visits to Pornographic Domains by Education



sonal information or hosts malware can lead to loss of personal information and money. Given the rich information that browsing data can give us, it is imperative to find accurate, light-weight, and scalable ways to learn what kind of content people are browsing.

In this paper, we contribute a solution to the problem. We present a new way to learn the kind of content hosted by the domain using the information in the domain name alone. We use LSTM 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 find that we can learn good models for domains that host adult content with precision and recall around .9. Our models for classifying websites that carry harmful content do not work as well with f1-scores of about .8. 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 may still be too low.

Figure 7: Proportion of Visits to Pornographic Domains by Age



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.

Aldwairi, Monther and Rami Alsalman. 2012. "Malurls: A lightweight malicious website classification based on url features." *Journal of Emerging Technologies in Web Intelligence* 4(2):128–133.

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* .

Deri, Luca, Maurizio Martinelli, Daniele Sartiano, Michela Serrecchia, Loredana Sideri and Sonia Prignoli. 2015. Implementing Web Classification for TLDs. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. Vol. 1 IEEE pp. 85–88.

Gers, Felix A, JJürgen 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.

Hald, Gert Martin, Neil M Malamuth and Carlin Yuen. 2010. "Pornography and attitudes supporting violence against women: Revisiting the relationship in nonexperimental studies." *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 36(1):14–20.

Jain, Ankit Kumar and BB Gupta. 2018. PHISH-SAFE: URL features-based phishing detection system using machine learning. In *Cyber Security*. Springer pp. 467–474.

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.".

Shen, Dou, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu and Wei-Ying Ma. 2004. Web-page classification through summarization. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* pp. 242–249.

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.

Wang, Huan-huan, Long Yu, Sheng-wei Tian, Yong-fang Peng and Xin-jun Pei. 2019. "Bidirectional LSTM Malicious webpages detection algorithm based on convolutional neural network and independent recurrent neural network." *Applied Intelligence* 49(8):3016–3026.

Westcott, Kevin, Jeff Loucks, Dan Littmann, Phil Wilson, Shashank Srivastava and David Ciampa. 2019. "Connectivity and mobile trends survey." https://www2.deloitte.com/us/en/pages/technology-media-and-telecommunications/articles/global-mobileconsumer-survey-us-edition.html.

Zhang, Ji-bin, Zhi-ming Xu, Kun-li Xiu and Qi-shu Pan. 2010. "A Web Site Classification Approach Based On Its Topological Structure." *Int. J. of Asian Lang. Proc.* 20(2):75–86.

SI 1 Supporting Information

SI 1.1 Data

Table SI 1.1: 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

Table SI 1.2: 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

SI 1.2 Model Performance

Table SI 1.3: OOS Performance of the Malware LSTM Model

malware or not	precision	recall	f1-score	support
0	0.86	0.95	0.91	10,000
1	0.77	0.51	0.61	3,048
weighted avg	0.84	0.85	0.84	13,048

Table SI 1.4: OOS Performance of the PhishTank LSTM Model

phishing or not	precision	recall	f1-score	support
0	0.79	0.90	0.84	10,000
1	0.89	0.76	0.82	10,000
weighted avg	0.84	0.83	0.83	20,000

Table SI 1.5: OOS Performance of the Shalla LSTM Model

categories	precision	recall	f1-score	support
adv	0.80	0.41	0.54	2,542
anonvpn	0.80	0.68	0.74	1,396
downloads	0.54	0.38	0.45	835
dynamic	0.81	0.54	0.65	213
education/schools	0.89	0.78	0.83	2,014
finance/banking	0.79	0.53	0.64	998
finance/insurance	0.96	0.82	0.88	616
finance/moneylending	0.86	0.80	0.83	760
finance/realestate	0.69	0.39	0.50	276
fortunetelling	0.71	0.35	0.47	215
forum	0.73	0.80	0.77	1,612
gamble	0.83	0.75	0.79	2,765
hobby/games-online	0.68	0.48	0.56	2,772
hobby/pets	0.66	0.34	0.45	3,233
hospitals	0.83	0.69	0.75	327
jobsearch	0.83	0.46	0.59	859
movies	0.68	0.47	0.56	1,112
music	0.85	0.85	0.85	1,784
others	0.49	0.30	0.37	22,198
porn	0.85	0.93	0.89	165,489
radiotv	0.61	0.48	0.54	712
recreation/restaurants	0.79	0.28	0.41	282
recreation/sports	0.64	0.63	0.64	24,085
recreation/travel	0.71	0.65	0.68	27,789
redirector	0.84	0.66	0.74	5,873
religion	0.90	0.82	0.86	1,838
science/astronomy	0.69	0.87	0.77	207
sex/lingerie	0.41	0.45	0.42	211
shopping	0.54	0.62	0.58	33,453
webmail	0.78	0.56	0.65	705
webradio	0.48	0.47	0.48	451
weighted avg	0.76	0.76	0.76	307,622

Table SI 1.6: OOS Performance of the Toulouse/Squidguard LSTM Model

categories	precision	recall	f1-score	support
adult	0.97	0.99	0.98	374,149
bank	0.68	0.56	0.61	338
gambling	0.49	0.18	0.27	202
games	0.84	0.45	0.58	1,871
malware	0.97	0.48	0.64	893
others	0.60	0.19	0.28	4,288
phishing	0.72	0.60	0.65	12,543
press	0.78	0.53	0.63	882
shopping	0.67	0.44	0.53	7,266
weighted avg	0.95	0.96	0.95	402,432

SI 1.3 Supplementary Results

Figure SI 1.1: Time Spent on Phishing/Malware Domains by Education

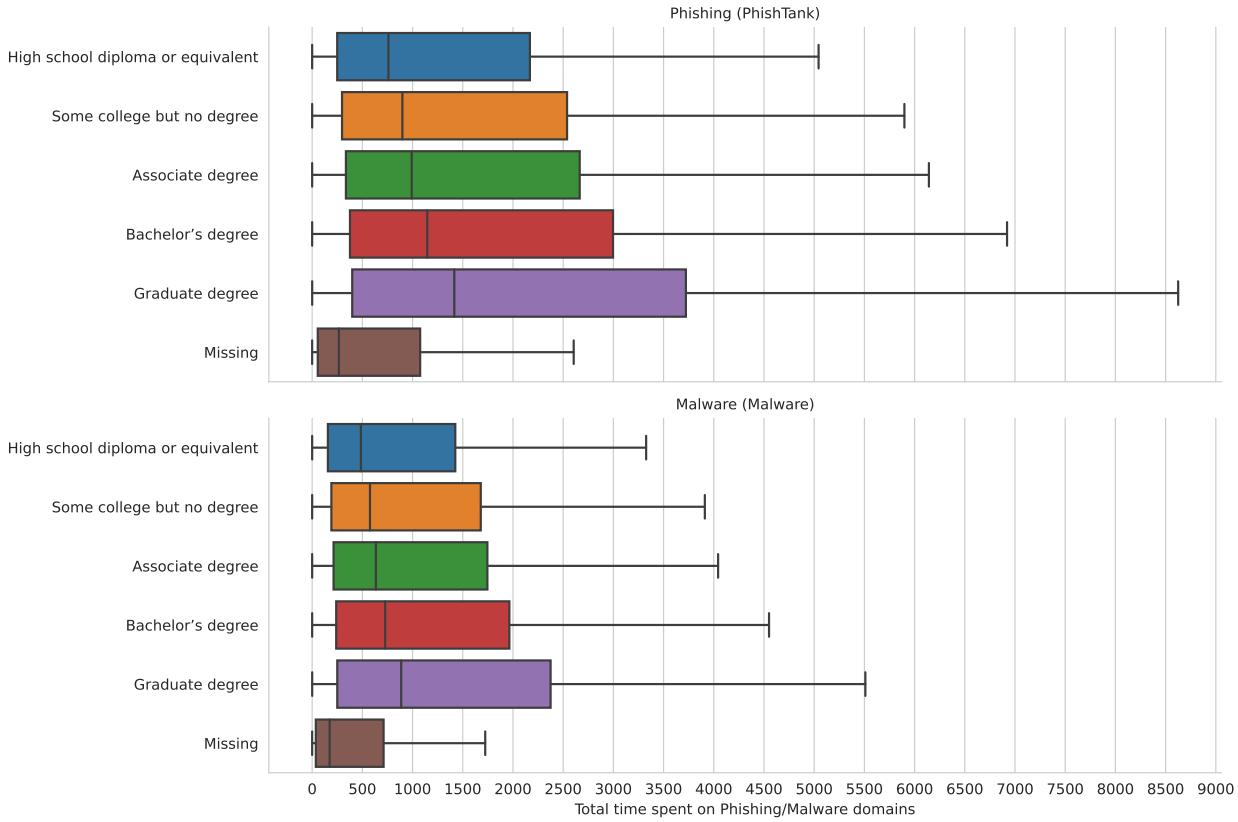


Figure SI 1.2: Time Spent on Phishing/Malware Domains by Race

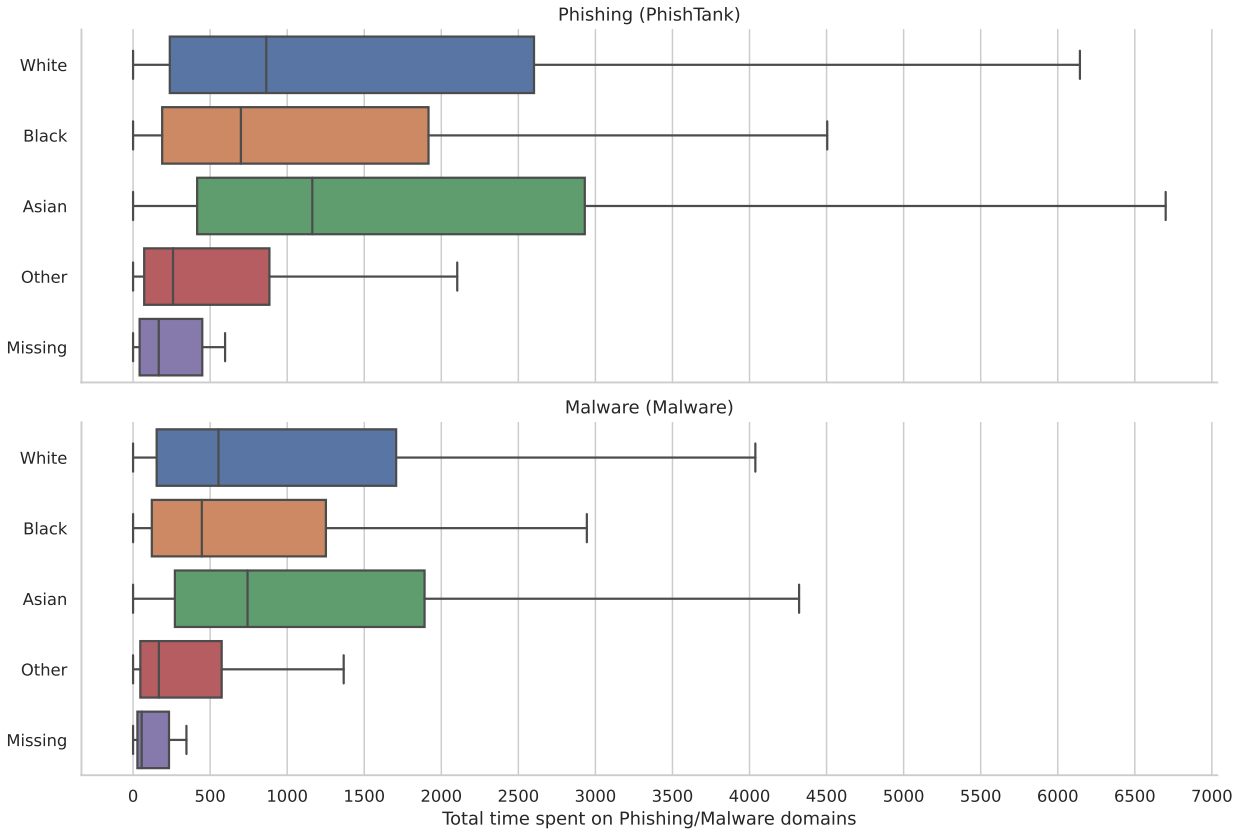


Figure SI 1.3: Time Spent on Phishing/Malware Domains by Age

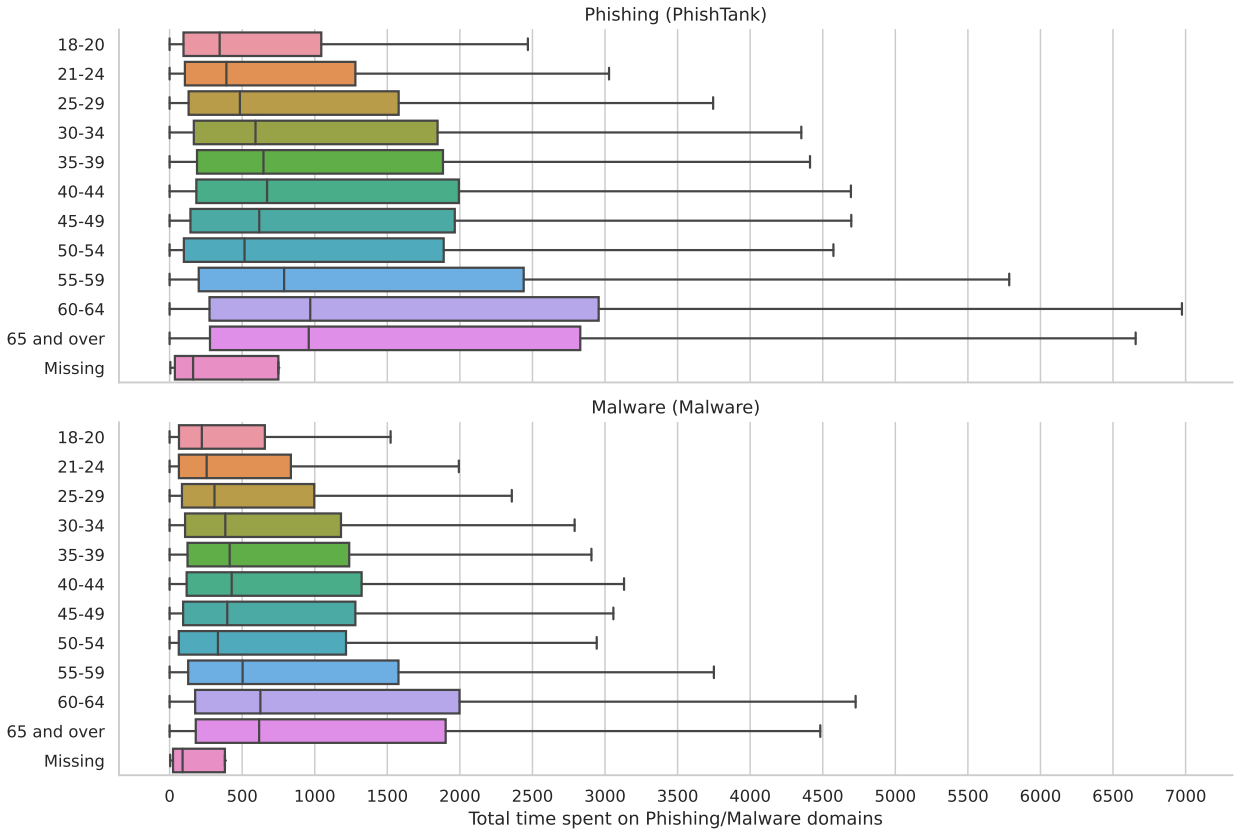


Figure SI 1.4: Proportion of Time Spent on Phishing/Malware Domains by Education

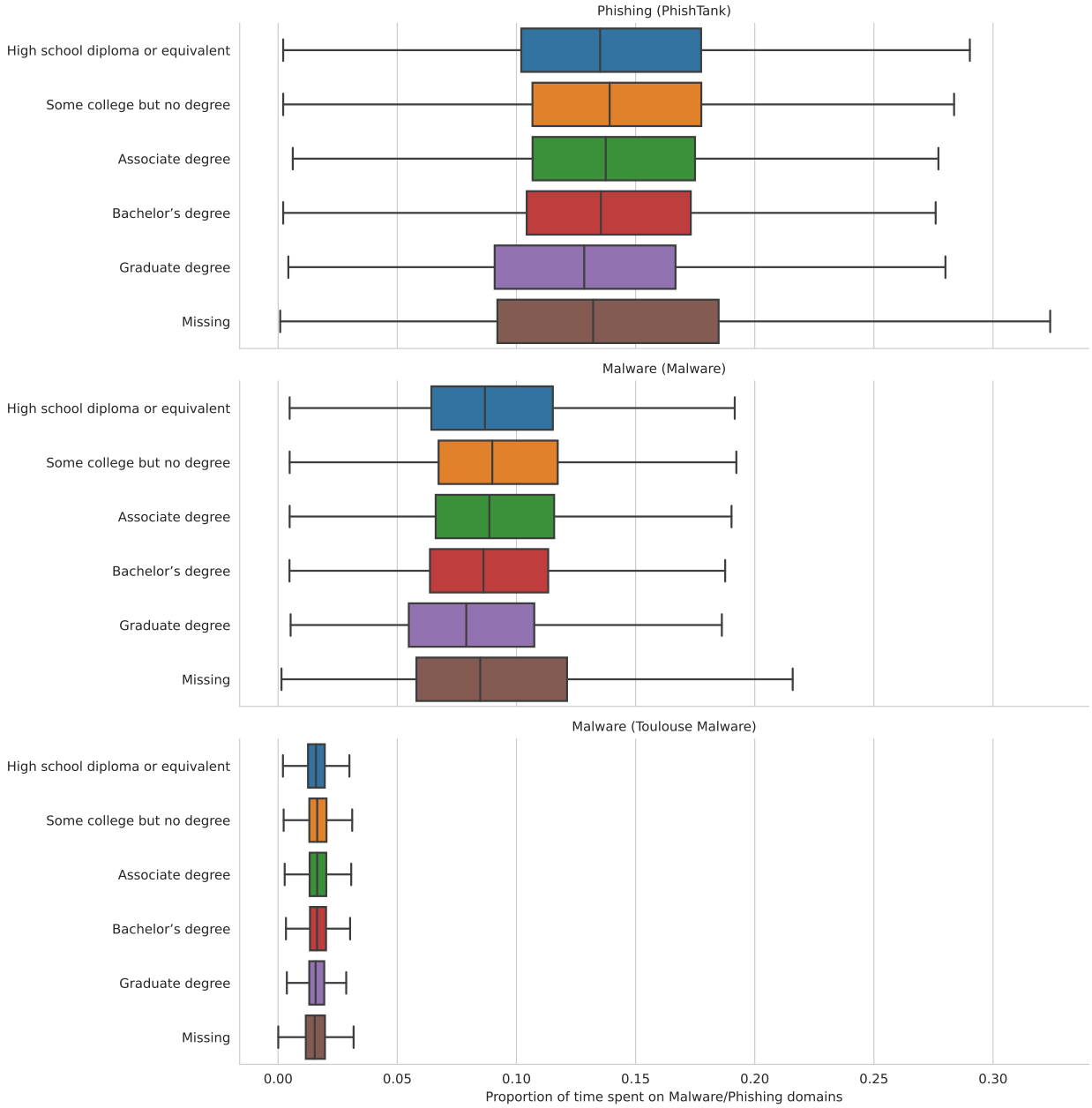


Figure SI 1.5: Proportion of Time Spent on Phishing/Malware Domains by Race

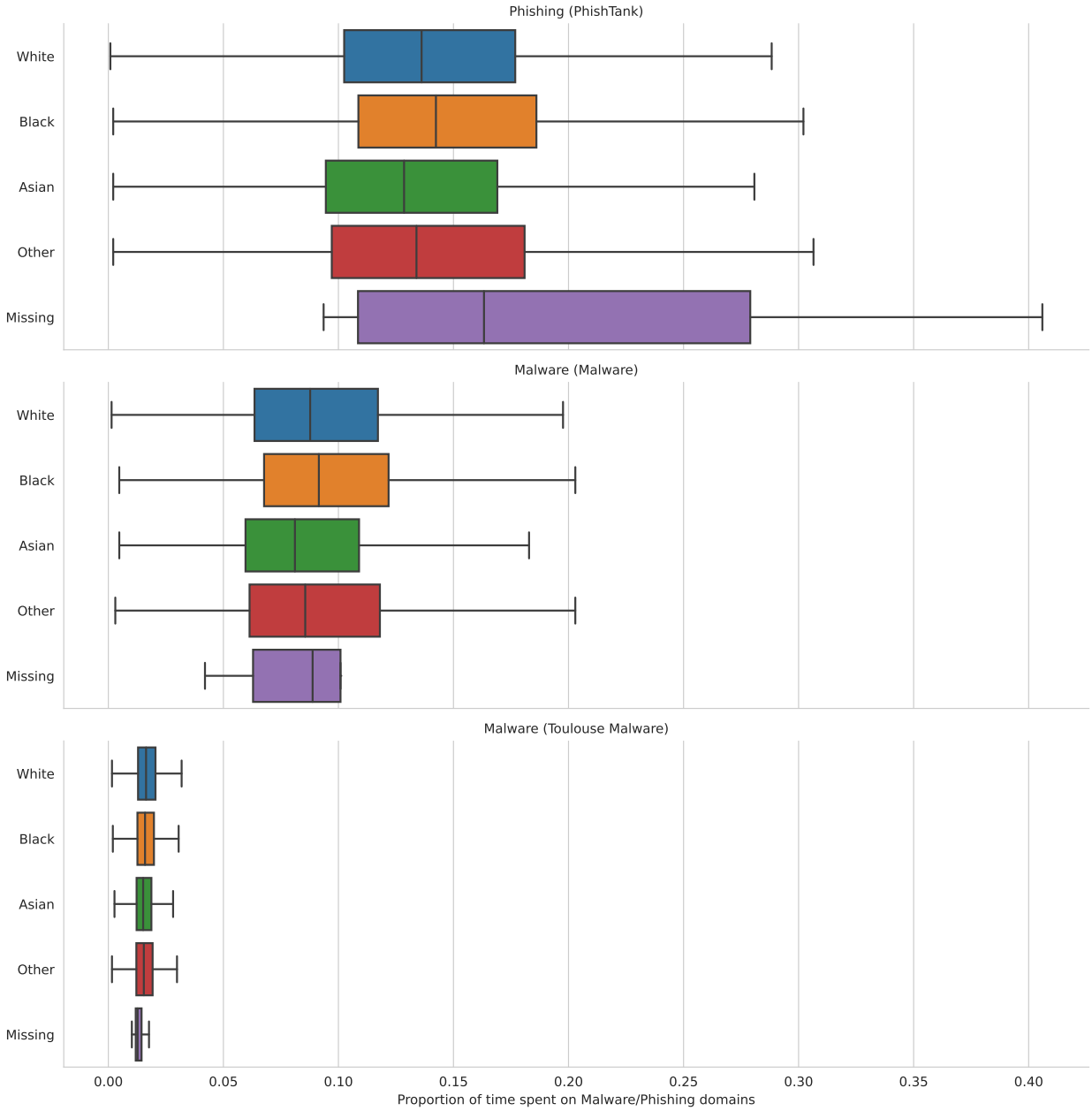


Figure SI 1.6: Proportion of Time Spent on Phishing/Malware Domains by Age

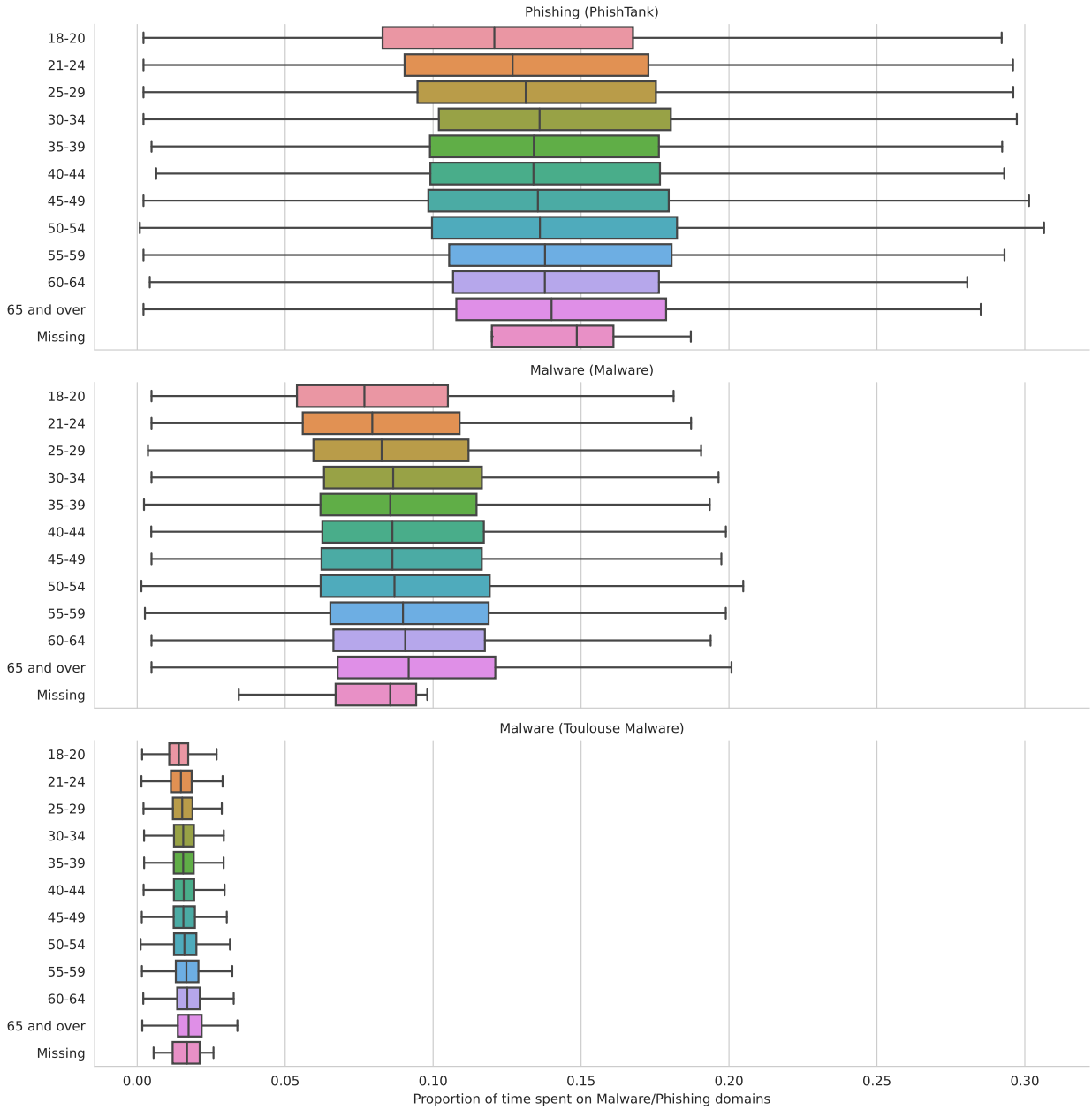


Figure SI 1.7: Proportion of Visits to Phishing/Malware Domains by Education

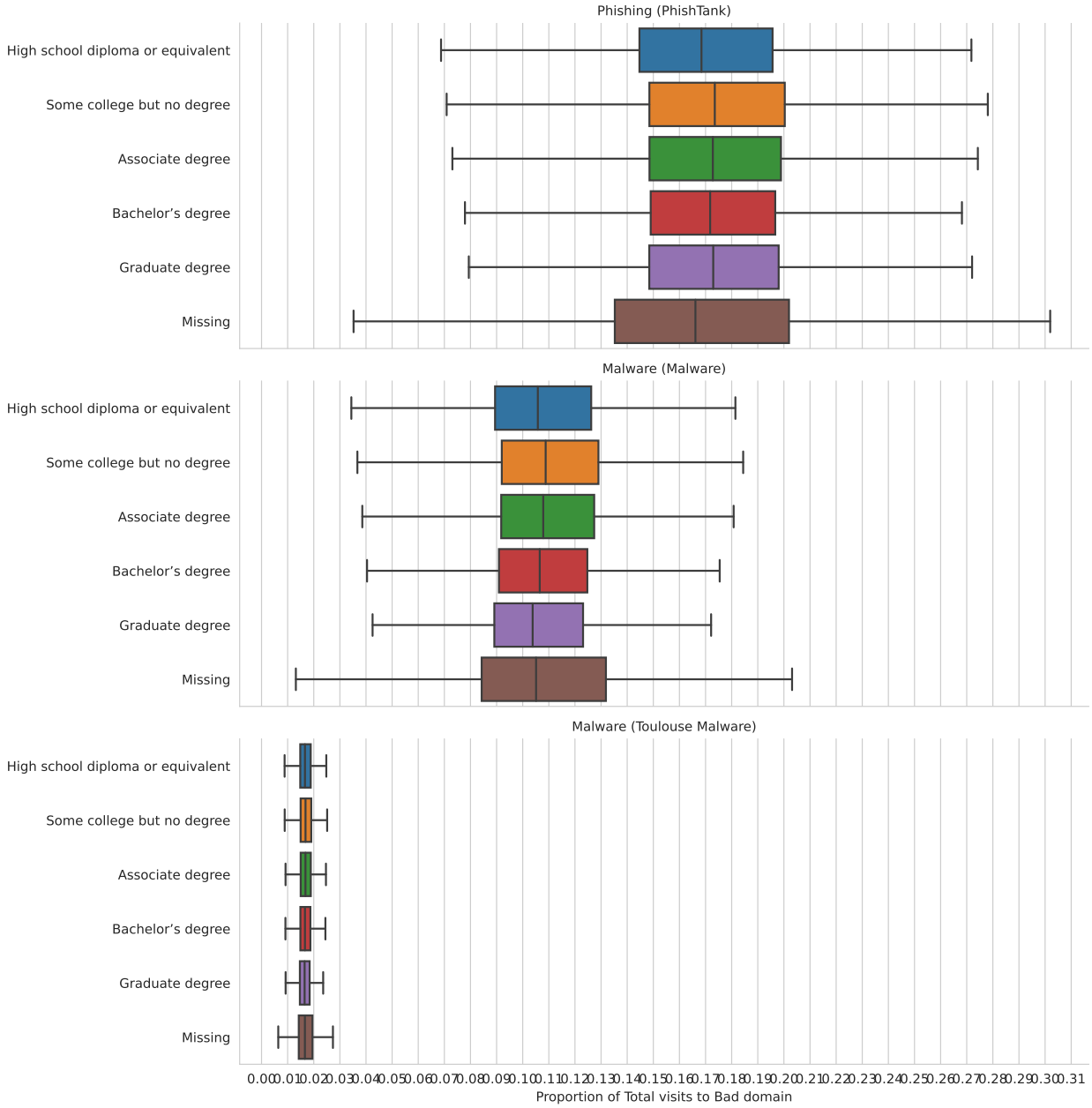


Figure SI 1.8: Time Spent on Pornographic Domains by Education

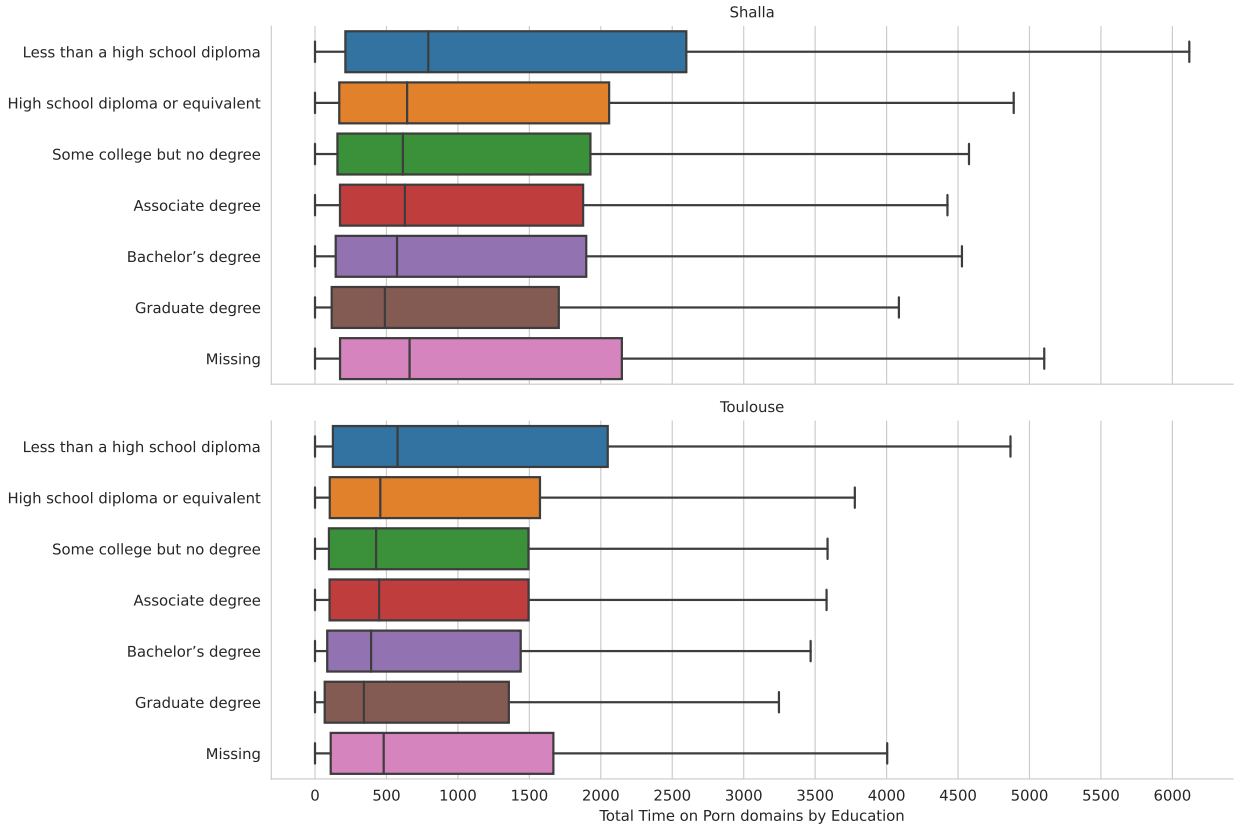


Figure SI 1.9: Time Spent on Pornographic Domains by Age

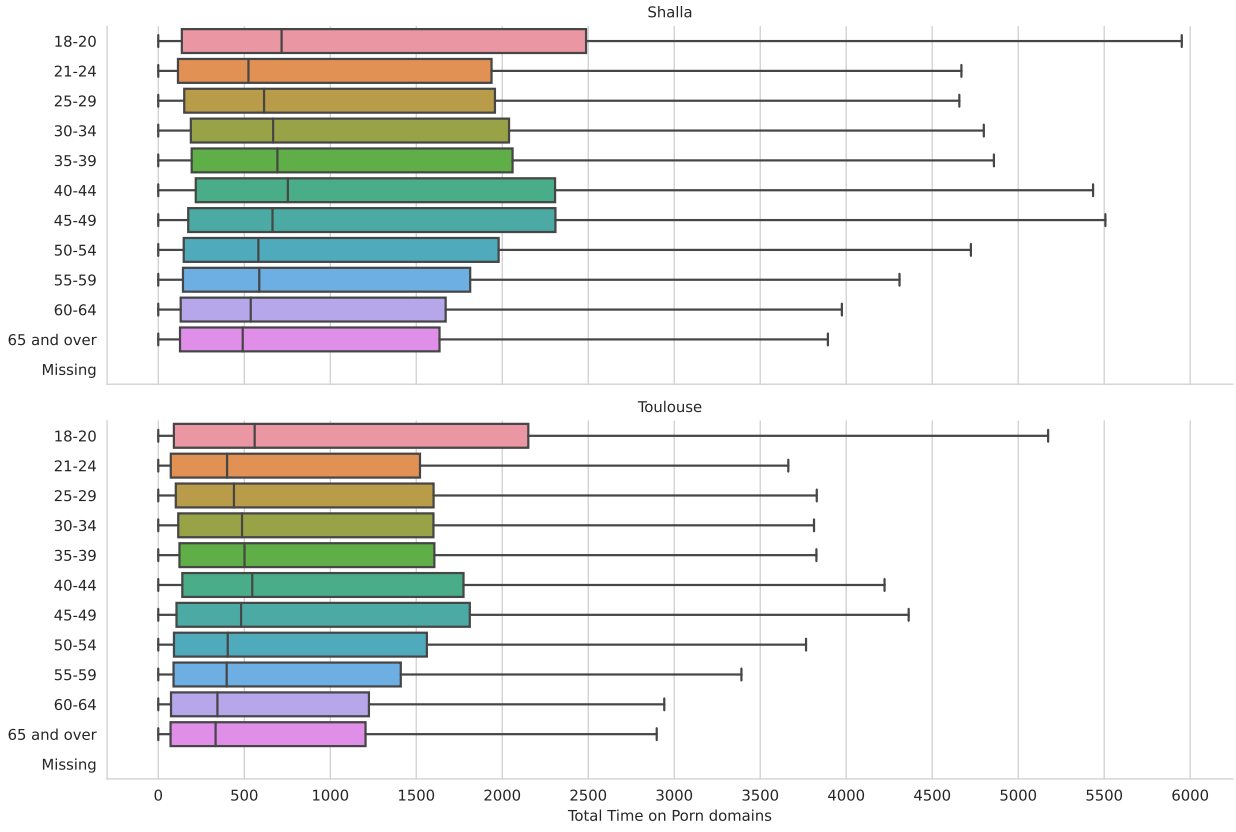


Figure SI 1.10: Proportion of Visits to Phishing/Malware Domains by Race

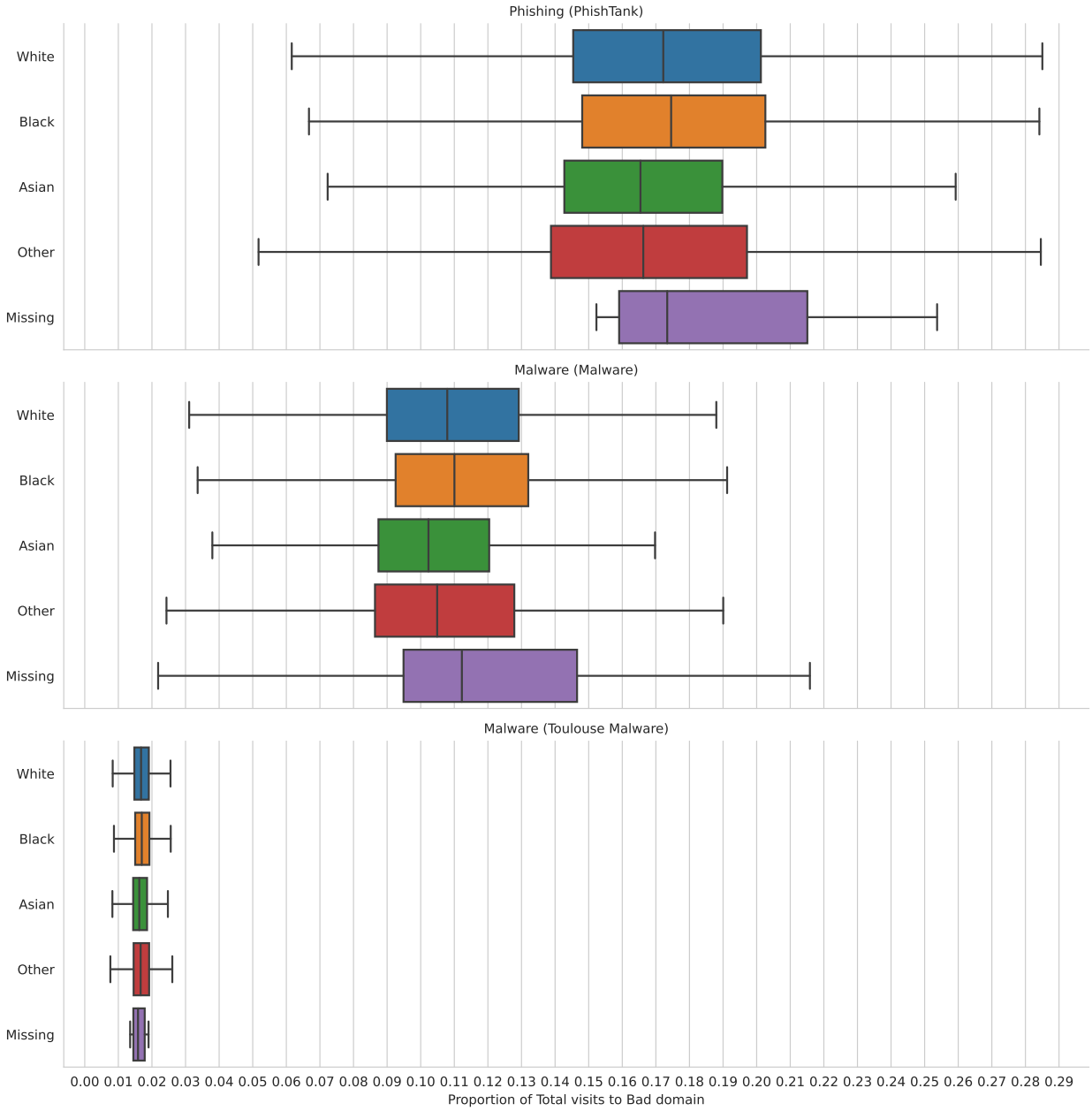


Figure SI 1.11: Proportion of Visits to Phishing/Malware Domains by Age

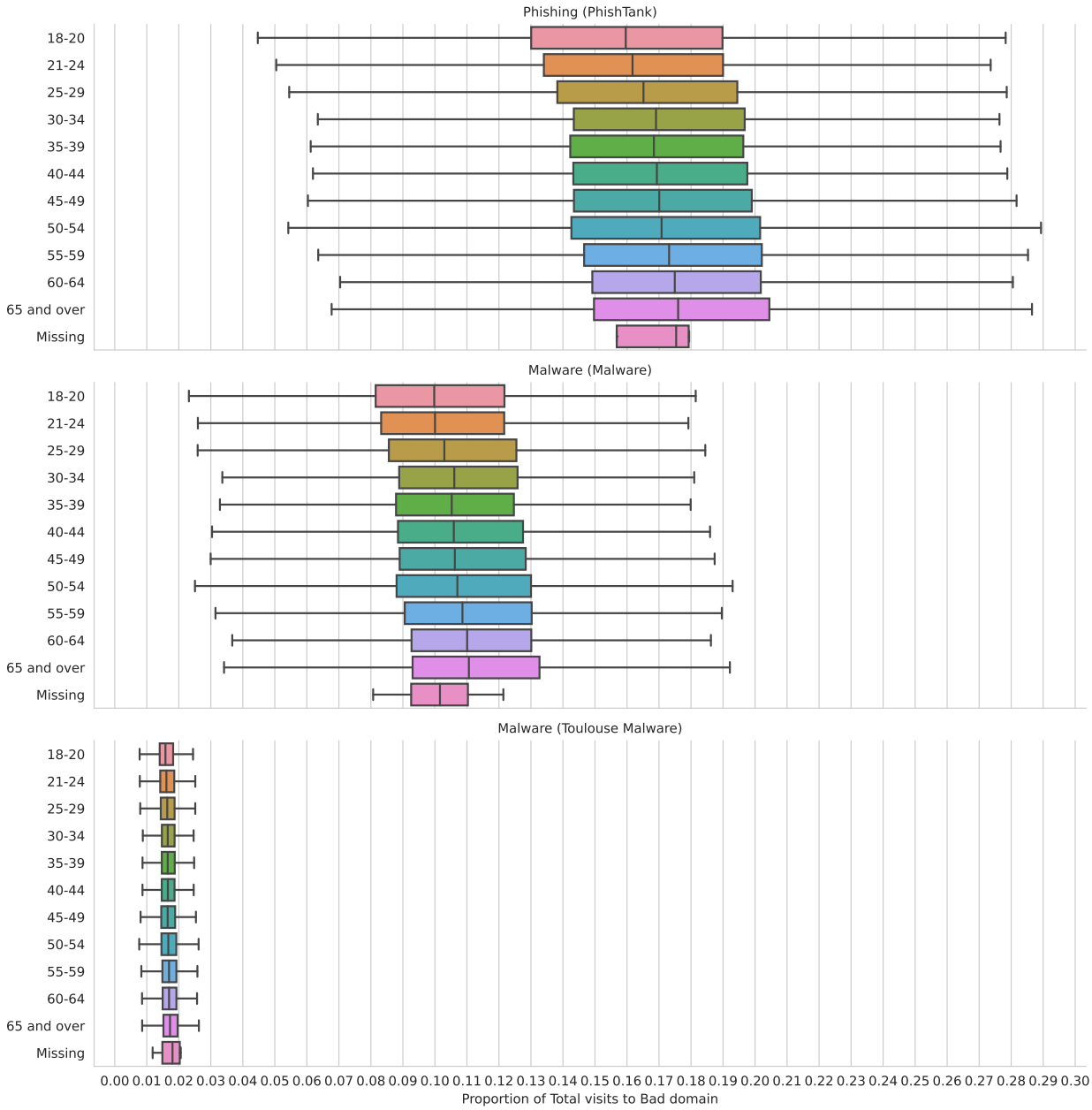


Figure SI 1.12: Proportion of Time Spent on Pornographic Domains by Education

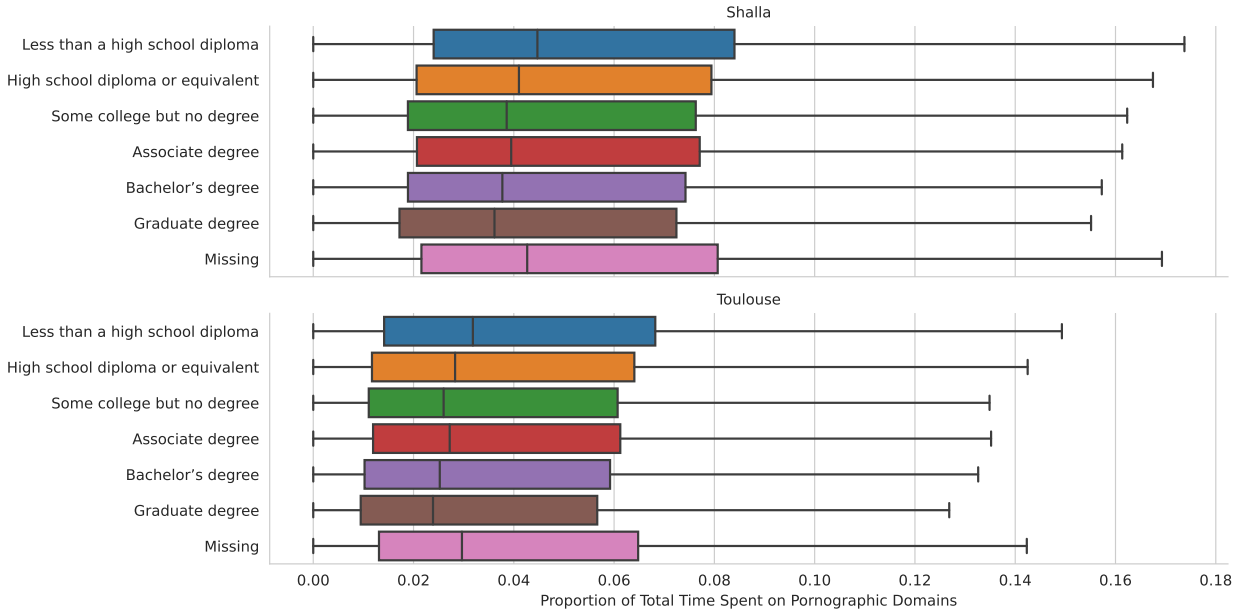


Figure SI 1.13: Proportion of Time Spent on Pornographic Domains by Age