The E-Commerce Market for "Lemons": Identification and Analysis of Websites Selling Counterfeit Goods

John Wadleigh
Computer Science & Engineering Department
Southern Methodist University
Dallas, TX, USA
jwadleigh@smu.edu

Jake Drew
Computer Science & Engineering Department
Southern Methodist University
Dallas, TX, USA
jdrew@smu.edu

Tyler Moore
Computer Science & Engineering Department
Southern Methodist University
Dallas, TX, USA
tylerm@smu.edu

ABSTRACT

We investigate the practice of websites selling counterfeit goods. We inspect web search results for 225 queries across 25 brands. We devise a binary classifier that predicts whether a given website is selling counterfeits by examining automatically extracted features such as WHOIS information, pricing and website content. We then apply the classifier to results collected between January and August 2014. We find that, overall, 32% of search results point to websites selling fakes. For 'complicit' search terms, such as "replica rolex", 39% of the search results point to fakes, compared to 20% for 'innocent' terms, such as "hermes buy online". Using a linear regression, we find that brands with a higher street price for fakes have higher incidence of counterfeits in search results, but that brands who take active countermeasures by filing DMCA requests experience lower incidence of counterfeits in search results. Finally, we study how the incidence of counterfeits evolves over time, finding that the fraction of search results pointing to fakes remains remarkably stable.

1. INTRODUCTION

The economist George Akerlof won a Nobel Prize for his work describing markets with asymmetric information [3]. In such a market for "lemons", consumers cannot reliably distinguish between high- and low-quality goods. This inability triggers a crucial market failure, whereby prices are driven down and low-quality goods dominate.

More recently, Anderson argued that the market for secure software is also a lemons market, as buyers cannot reliably observe whether or not the software they are buying is in fact secure [4]. Researchers working in security economics now recognize that information asymmetry is one of the fundamental barriers facing cybersecurity today [5].

In this paper, we study a related market for lemons: the online sale of clothing and luxury goods. Miscreants selling knockoff versions of popular goods have proliferated online in recent years. Frequently, counterfeit-goods shops are manipulating search engines to get high placement in search results for legitimate terms, thereby duping consumers into thinking they can get a good deal on the real thing. If search engines cannot distinguish legitimate sellers from those selling counterfeits, why would we expect people to be capable of doing so?

The online sale of knockoffs matters, and not just for those brands whose goods are impersonated. Bad experiences with e-commerce carry large indirect costs, in that they can turn people off from online participation and erode trust in the Internet [28]. Consequently, in this paper we set out to investigate the prevalence of counterfeit goods online. We make the following contributions.

- We present a methodology for collecting data on the prevalence of counterfeit goods in web search (Section 2).
- We build an accurate classifier using features automatically extracted from website content that distinguishes legitimate from fake sellers based upon data returned by search results (Section 3).
- We conduct extensive empirical analysis of data collected between January and August 2014 (Section 4). We apply the trained classifier and find that 5 407 websites are selling counterfeits, out of a total of 18 756 websites. Overall, 32% of search results point to fakes, and 79% of queries issued included at least one fake in the first page of results.
- We show that while search engines do refer customers to counterfeit sellers when the search terms ask for it, they also refer customers to counterfeit sellers in large numbers even when they give no indication that they want fakes.
- We present a linear regression that demonstrates that the higher the selling price is for fakes for a given brand, the more search results point to fakes. The regression also demonstrates that active enforcement via DMCA takedowns can reduce the prevalence of fakes in search results by 9 %-pts.
- We show that the prevalence of fakes among brands is relatively stable over time, and furthermore that some sellers respond to their website dropping out of search results by adding copies registered at different URLs.

2. DATA COLLECTION METHODOLOGY

Researchers have extensively studied the online sale of unlicensed pharmaceuticals [17, 19, 20, 23], as well as the unauthorized acquisition of digital goods such as music, movies and software [15, 30]. We decided to focus instead on the sale of counterfeit consumer physical goods, due to its prevalence and relative lack of attention from the research community.

2.1 Constructing Search Queries

We first had to decide on which brands to focus our investigation. We began by collecting data on five seed brands: Ugg, Coach, Rolex, Hermes, and Oakley. After gathering search results on associated queries, we scraped all product listing pages from 68 stores manually identified as selling counterfeits. We then decided to focus on the 25 brands observed most often in the inventories of the confirmed counterfeit stores.

We are very interested in measuring the extent to which consumers intending to buy from authorized retailers are instead presented with links to websites selling fakes. To that end, the 25 brands were then paired with search terms of varying levels of "innocence". We selected three search terms for each innocence level: innocent, grey, and complicit. We deem as innocent the keywords "fast delivery," "buy online" and the lack of keyword (meaning the query is only the product name). We deem the keywords "replica," "fake," and "knockoff" as complicit, as any shopper using those terms is clearly seeking out counterfeited goods. Between these extremes lie grey keywords "cheap," "discount," and "sale", since there is ambiguity as to the intention of the shopper. In total, we combine all 25 brands with each of the 9 keywords to yield 225 unique search queries.

2.2 Gathering Data on Websites in Search Results

We automatically issued queries to the Google Custom Search API to obtain the top 100 results for each of the 225 terms. We then visited the URLs using an automated browser, storing the HTML and a screenshot.

When the automated browser visited the search results, a script checked for the presence and properties of elements we believe to be indicative of malicious intent. The elements observed were iFrames, currencies, and pricing information.

We took steps to remove price outliers that were triggered by parsing errors or ambiguities in brand names (e.g., searches for Coach purses can also return advertisements for buses). Also, because pricing data was pulled for currencies besides US dollars (specifically Euro, Pound, and Yen), the prices extracted were converted to USD in order to make meaningful analysis.

Pricing data was found using regular expressions and a combination of currency symbols, and price associations were found (for example a product might contain both an "original price" and a "discount price") by locating individual prices and climbing the DOM structure. We also fetched the WHOIS data for each website, extracting the date and country of registration, as well as whether or not a privacy or proxy registration had been used [9].

We gathered data in two distinct periods. An initial study carried out in January 2014 identified 21 646 search results and 6 979 distinct websites. To inspect for changes in behavior over time, we reissued the same queries weekly between June and August 2014.

3. CLASSIFYING WEBSITES SELLING COUNTERFEITS

We first describe the features used in building the classifier in Section 3.1 and then outline the methods used in Section 3.2.

3.1 Feature Selection and Extraction

We constructed several features after inspecting many websites selling counterfeit goods. We focus on three classes of features: URL-level, page-level, and website-level features.

URL-level features The least resource-intensive approach is to select features based upon only the URL. This approach has been used to identify phishing websites [7] and malware [21].

1. **Replica In FQDN**
   This Boolean feature identifies when the term "replica" is contained as part of the web page's fully qualified domain name (FQDN). We also considered the term "knockoff", but that was often associated with articles and blogs decrying counterfeits.
2. **Length of FQDN**
   This numeric feature denotes the number of characters present in the URL's fully qualified domain name.

Page-level features We also inspected the scraped HTML content to identify additional features indicating that counterfeits may be sold.

1. **Number of Currencies Seen**
   Unlike most big box retailers which offer custom websites for each country serviced, counterfeit goods websites typically offer a single unified site which is designed to service any number of countries. Furthermore, it is uncommon for these sites to implement their own custom payment vehicles. Most of the third party payment solutions deployed offer checkout alternatives which include providing payment in a large variety of currencies.
2. **Large iFrames**
   Unusually positioned and sized iFrames are used to obfuscate malicious scripts and redirections common in criminal websites [8, 22]. We define large iFrames as having unusually large height and width in addition to containing a different top level domain which is also not part of the Alexa top 1000 domains [1].

3. **Percent Savings Average**
   This numeric feature indicates the average percentage of savings on a given webpage. This is relevant on online stores whose pricing data was able to be scraped automatically, in the case where two prices were listed in association with at least one item (an "original" price and their price, to demonstrate the savings). The average of the savings percentages on a given page is stored in the hopes of finding counterfeit stores offering ludicrously high savings.
4. **Number of Times Duplicate Price Seen**
   This numeric feature specifies the number of times a duplicate price was seen on a given page. For example, a page with various products listed at prices $40, $45, $40, $50, $40, and $45 has 3 duplicate prices present ($40 is repeated twice and $45 is repeated once). This feature was included to catch lazy counterfeit store owners who copy and paste products, changing titles but not prices.
5. **Page Contains Webmail Address**
   This Boolean feature identifies when a webpage contains an email address which includes the text "@yahoo.com", "@gmail.com", or "@hotmail.com". It would be highly unusual for a legitimate brand reseller to utilize a free webmail account.
6. **Unique Brand Term Count**
   The unique search term count represents the number of unique search terms identified within the webpage's HTML content.
7. **Top-Level Page Mentions Brand**
   We also visited the top-level web page to look for the mention of any brand (a string search for any of the brands in the root page's text). It indicates that a website may have been hacked if the top-level page makes no mention of any brands while the page listed in the search results does. A website hacked to host a store is almost certainly selling counterfeits.
8. **Content Consistent with Takedown Page**
   This Boolean feature identifies when a webpage has been taken down and replaced with content from brand-enforcement companies.

Website-level features The final category of features were those describing characteristics of the website itself, as opposed to the displayed content.

1. **Private or China-registered WHOIS**
   While legitimate companies use private and proxy WHOIS registrations [10], it is frequently employed by those conducting dubious operations such as selling counterfeit goods. Furthermore, we observed that many websites selling replicas have operations based in China.
2. **WHOIS Registration Age Under 1 Year**
   This Boolean feature identifies when a webpage is less than one year old.
3. **Website In Alexa Top 100K**
   This Boolean feature identifies when a webpage is ranked in the top one hundred thousand websites by Alexa based on the webpage's web traffic.

	GLM		SVM		ADA	
	#	%	#	%	#	%
TP	175	29.1%	180	29.9%	125	20.8%
TN	337	66.0%	340	56.5%	318	52.8%
FP	31	5.1%	28	4.7%	50	8.3%
FN	59	9.8%	54	9.0 %	109	18.1%
Accuracy	85.0%		86.4%		73.6%	
Precision	85.0%		86.5%		71.4%	
Recall	74.8%		76.9%		54.4%	

Table 1: Truth tables and accuracy measures for each classifier using 10-fold cross-validation.

3.2 Building and evaluating the classifiers

Counterfeit websites were classified using Logistic Regression, Adaptive Boosting, and Support Vector Machine models. Each of the machine learning models were trained against our specialized features developed to identify counterfeit goods websites. Finally, the models were validated against a manually constructed ground truth dataset to assess each model's detailed classification accuracy characteristics.

We implemented the three models using the R programming language. Three of these packages include Support Vector Machines (SVM) [24], Generalized Linear Models (GLM) [29], and Adaptive Boosting (AdaBoost) [11]. While all three packages can be highly accurate for various types of classification problems, each package performs very differently when modeling (i.e. learning) different volumes of input data [13].

Ground Truth Identification One of the fundamental challenges to classifying logical copies of counterfeit goods websites is the lack of ground-truth data available for evaluating the accuracy of automated feature selection and classification methods. Some researchers have relied on expert judgment to assess similarity, (e.g., [18]) but most forego any systematic evaluation due to a lack of ground truth. We now describe a method for constructing ground truth datasets for samples of counterfeit goods websites.

In order to classify stores, 602 unique screenshot/HTML pairs were pulled at random from the data collected for manual inspection. The screenshots were then examined to determine whether the store appeared to be selling fakes – a task which is easier, and much slower, to do by hand. Of the 602 stores sampled, 234 were determined to be counterfeit and 368 were determined to not be counterfeit.

Results We independently trained and evaluated logistic regression (GLM), Support Vector Machine (SVM) and, adaptive boosting (ADA) models using 10-fold cross-validation. Table 1 shows the detailed truth tables for each model, along

Feature	Coef.	Odds ratio	p-value
Page Contains Webmail Address	0.697	2.007	0.1722
Unique Brand Term Count	0.167	1.182	< 0.0001
# Currencies Seen	0.240	1.272	0.0017
Large iFrames	5.320	204.3	< 0.0001
Private or China WHOIS	0.285	1.330	0.384021
Replica in FQDN	1.442	4.227	0.0002
WHOIS Registration $<$ 1 Year	1.505	4.504	0.0001
Percent Savings Average	0.044	1.045	< 0.0001
# Times Duplicate Price Seen	0.005	1.005	0.4471
Top-Level Page Mentions Brand	-0.701	0.496	0.0097
Website on Takedown Page	2.892	18.05	0.0005
Length of FQDN	0.044	1.045	0.0782
Website in Alexa Top 100K	-2.626	0.072	< 0.0001

Table 2: Coefficients and odds ratios for the logistic regression classifier (terms in bold are statistically significant).

with figures for accuracy, precision and recall. Logistic regression and SVM produced more accurate results than adaptive boosting.

To get a sense of the relative importance of different features in the classifier, we can examine the coefficients and odds ratios from the best-fit logistic regression trained on the ground-truth data. Table 2 presents the results, with terms that are statistically significant in the model highlighted in bold. As expected, the presence of a large iFrame loading an external website is highly associated with the website selling counterfeits. In fact, websites exhibiting this behavior face 204-times greater odds that they are selling counterfeits! Newly registered domains, using the term 'replica' in the FQDN, and appearing on a takedown page are all associated with selling fakes. The more currencies available on a website, the greater the advertised savings and the longer the FQDN, the more likely the website is to be fake. Surprisingly, however, a private or Chinese WHOIS registration address was not found to be statistically significant. Finally, two features are negatively associated with selling fakes – websites with a top 100,000 Alexa ranking and those whose top-level index page also mention the brand are less likely to sell fakes. The latter reflects the fact that the website is more likely to be an actual merchant and not a compromised host.

4. EMPIRICAL ANALYSIS

We now apply the best-performing SVM classifier to thousands of search results gathered between January and August 2014. In Section 4.1 we examine how the prevalence of stores selling fakes varies by brand and type of search query, while in Section 4.2 we describe a regression to help explain why the search results for some brands include more fakes than for others. In Sections 4.3 and 4.4 we study how knockoff search results vary over time. Finally, in Section 4.5 we examine characteristics of the fake websites themselves.

4.1 How prevalent are counterfeits in search?

Recall from Section 2 that in January 2014 we gathered up

	% Fake Search Results	# Fake Websites	% queries page 1 fake	% queries result 1 fake
Innocent	20%	631	64%	6%
Grey	35%	875	86%	28%
Complicit	39%	780	86%	49%
Overall	32%	1 587	79%	28%

Pairwise χ^2 comparison	% results fake		% queries page 1 fake		% queries result 1 fake	
	adj. p	Sig.?	adj. p	Sig.?	adj. p	Sig.?
Innocent vs. Grey	0.0000	✓	0.0067	✓	0.0004	✓
Innocent vs. Complicit	0.0000	✓	0.0067	✓	0.0000	✓
Grey vs. Complicit	0.0000	✓	1.0000		0.0150	✓

Table 3: Comparing the prevalence of counterfeits by search query intent. The top table reports the results, while the bottom establishes whether or not the differences are statistically significant according to a pairwise χ^2 test with FDR-adjusted p-values.

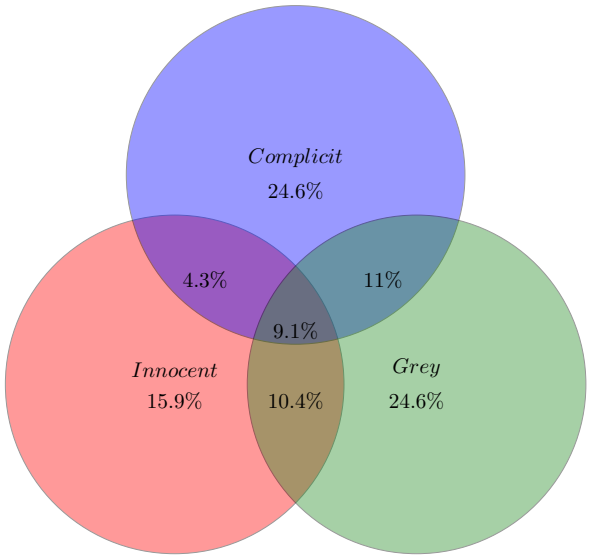

Figure 1: % unique counterfeit stores found through innocence level of query

to the first 100 results to 225 queries spanning 25 brands targeted by counterfeiters. Overall, 32% of these search results linked to websites selling knockoffs, spanning 1 587 distinct websites. In 79% of the search queries, at least one link in the first page of results (top 10) pointed to fakes. Nearly one third of the time, the very first hit was to a website selling fakes!

Table 3 reports these overall measures, but it also breaks down the results by the intent of the search query: innocent (brand only, brand+"buy online", brand+"fast delivery) grey (brand+"cheap", "discount", or "sale"), or complicit (brand+"replica", "fake", or "knockoff"). Unsurprisingly, fewer innocent search queries linked to knockoffs than grey or complicit ones (20% vs. 35–39%). Nonetheless, it is striking that the fraction of fake results remains so high for obviously benign search terms, as well as how similar the proportion is for grey and complicit terms (e.g., "discount"

vs. "replica").

The bottom rows of Table 3 report a statistical test for significance of the difference between these proportions using pairwise χ^2 tests. From these tests, we can see that the differences in proportion of fakes in search results across each query type are in fact statistically significant.

But what about the differences for high-ranking results? 64% of innocent queries have at least one fake result in the first page, compared to 86% for grey and complicit queries. The difference in page one fake prevalence between innocent and grey queries is statistically significant, but the difference between grey and complicit is not. Finally, in 6% of the innocent search queries the first result is fake, compared to 28% for grey and 49% for complicit queries (all differences statistically significant). Hence, we can conclude that regardless of the intent of the person issuing searches, there remains a good chance she will inadvertantly be exposed to a website selling knockoffs.

Another way to compare the results based on query intent is to check whether stores appearing in one type of query also appear in other query types. Figure 1 illustrates this overlap. While roughly 9% of the counterfeit sites appeared in the results of all three search categories, 65.1% were found under only one query type. In other words, many stores selling fakes may target different types of users: those targeting customers who know they are buying fakes may present a different shopping experience than to those selling to people who think what they are buying is legitimate.

We next examine the results broken down by brand. The first numerical column in Table 4 reports the proportion of search results pointing to fakes by brand in descending order. Bvlgari (or "Bulgari"), an Italian luxury goods company known for high-end watches, topped the list with 49.39% of its search results pointing to fake stores. Following close behind is Hublot, a Swiss luxury watchmaker. In fact, the nine brands with the most fakes are watch companies. By contrast, just 7.39% of results for Coach (maker of luxury handbags) point to fakes.

The brands with proportions listed in green bold-face experienced disproportionately low rates of counterfeit search results, according to a χ^2 test, while those at the top in red bold-face face disproportionately high rates of counterfeits in their search results. For the brands in the middle, the deviations from the 42% overall average are not statistically signficant.

In the next section, we present a linear regression that helps explain why we see such variation in search results across brands. From the right-hand columns in Table 4 we can already see some potential explanations. First, those brands at the top tend to not be as aggressive in pursuing infringers via DMCA notice-and-takedown requests. By contrast, many brands near the bottom have filed many such takedown requests for brand infringement. Furthermore, the right-most column reports the median sales price for goods sold on counterfeit websites for each brand. For the luxury watches at the top, even the fakes can sell for thousands of dollars. By contrast, for many of the less-targeted brands, the fakes sell for a few hundred dollars at most.

4.2 What explains the variation in counterfeit prevalence?

We have established that there is substantial variation in how prevalent stores selling fakes are in web search results (as low as 7.39% for Coach and as much as 49.39% for Bvlgari). We now describe a linear regression to help identify what characteristics may explain these differences.

Because much of the variation can be attributed to differences in brands, we group results together by brands for this regression. The response variable is the percentage of search results pointing to fakes (left-most column in Table 4). We have constructed a number of explanatory variables that we hypothesize influence the prevalence of fakes:

1. **Churn**: When search results exhibit lots of turnover, we expect that it is easier for websites selling unauthorized goods to penetrate the results using search-engine optimization. We measure the average weekly churn for the June–August data sample by computing each week's churn as the sum of the FQDNs added and dropped in the period divided by the total number of observed FQDNs in adjacent periods.
2. **Popularity**: We hypothesize that the popularity of a brand is corrleated with how many people will try to sell fakes. To estimate popularity, we used the Google Trends to compute the relative search popularity of every brand compared to Nike, which is the most popular brand we studied according to Google.
3. **Active DMCA Enforcement**: Some companies vigorously defend their brand online, while others do not. One countermeasure available to brandholders is to issue DMCA takedown requests on websites selling counterfeits because they display copyrighted images owned by the brandholder. We searched the Chilling Effects Database [1] of DMCA requests for each brand. If we observed more than 25 DMCA takedown notices associated with the brand, then we deemed the brand to be actively employing the DMCA in protecting its brand.
4. **Average Counterfeit Sales Price**: The merchants peddling counterfeit goods have a strong financial incentive to sell fakes that command a higher street price. Hence, we computed for each brand the average asking price for all stores selling that brand's fakes.

Table 5 presents the results of the regression. Two explanatory variables are statistically significant. First, the presence of active DMCA enforcement is negatively correlated with the percentage of search results pointing to fakes. This is an encouraging result for any brand that has taken enforcement action but has wondered if it was a worthwhile endeavor or a futile game of whack-a-mole. These results

[1] http://www.chillingeffects.org

Brand	% fake search results	# fake websites	% queries page 1 fake	% queries result 1 fake	# DMCA reports	Avg. fake site churn %	Median fake price
bvlgari	49.39	193	88.89	55.56	0	18.49	$588.3
hublot	47.67	201	100	77.78	8	19.98	$2060.42
panerai	45.86	188	100	55.56	18	18.61	$1381.99
patek philippe	44.75	181	100	37.5	0	18.92	$3117.87
tag heuer	42.88	171	100	25	5	19.41	$991.75
breitling	42.56	188	100	11.11	12	18.49	$1928.46
cartier	39.78	173	66.67	33.33	1	19.65	$1066.68
iwc	39.5	179	100	50	3	21.06	$1339.76
fendi	33.9	141	71.43	14.29	1	19.45	$279.47
hermes	33	147	88.89	55.56	5	20.47	$261.33
dior	32.4	183	66.67	33.33	39	22.72	$221.98
gucci	29.67	178	77.78	22.22	16	23.85	$227.62
rolex	28.88	120	100	62.5	33	21.69	$4316.39
oakley	28.62	122	77.78	22.22	16	31.78	$112.85
prada	28.56	150	88.89	22.22	23	24.75	$297.38
versace	28.26	138	66.67	0	1	23.15	$182.95
air jordan	28.15	131	100	28.57	1439	29.36	$91.77
armani	27.44	153	66.67	11.11	1	20.89	$166.23
burberry	27.11	156	66.67	11.11	336	24.97	$210.45
louis vuitton	26.67	147	77.78	33.33	93	29.64	$284.41
ugg	25.14	95	77.78	11.11	10	28.68	$160.94
nike	20.67	125	55.56	0	1439	30.16	$88.99
adidas	17.46	99	50	0	9	23.32	$88.09
chanel	14.89	90	55.56	11.11	202	23.77	$630.27
coach	7.39	50	33.33	11.11	186	24.74	$223.07
Average	31.62	147.96	79.08	27.83	155.84	23.12	$812.78

Table 4: Counterfeit stores found in search results broken down by brand (left columns); additional per-brand characteristics such as DMCA enforcement activity anand the median advertised price among stores selling fakes (right columns). The entries in bold in the first column indicate a statistically significant difference in the brand's proportion of fakes in search results compared to the 42% average (using a χ^2 test with 95% confidence).

indicate it's the former; more precisely, brands actively enforcing copyrights using the DMCA see an 8.59 percentage point reduction in the fraction of results that link to fakes.

The second significant explanatory variable is the counterfeit price. The greater the asking price for fakes, the more prevalent counterfeit stores are in the search results for a given brand. In particular, every doubling of the average asking price corresponds to a 2.88 percentage point increase in the proportion of fakes in search results. This points to a second potential intervention brands might consider: identifying ways to undermine the street price for fakes. While somewhat counterintuitive, allowing some fakes to be sold at very low prices might deter others from entering the market.

Finally, we note that despite the simplicity of the regression and the relatively small number of brands involved, a large amount of variation can be explained using this model. With an R^2 value of 0.5855, this indicates that 65% of variation in the percentage of results leading to fakes can be explained by this simple linear model.

	Coefficient	Std. Error	p-value
Intercept	6.81	21.6	0.756
Churn	0.286	0.896	0.753
Popularity	-0.218	0.164	0.1982
Active DMCA Enforcement	-8.59	4.02	0.002245
$\log_2$**(Counterfeit Price)**	2.88	1.11	0.00087
	$R^2 = 0.6546$ (adj. $R^2 = 0.5855$)		

Table 5: Linear regression on counterfeit prevalence by brand. Significant variables are shown in bold.

4.3 How does the prevalence of counterfeits vary over time?

We now investigate how counterfeit peddling evolves over time by examining weekly collections of search results from June–August 2014. We first examine how the prevalence of fakes in search changes for brands over time. Figure 2 shows the relative position of each brand over the 12 week period, specifically charting their percent of counterfeit search results. The most striking observation from the bump chart is how consistent the positions were, despite moderate shifting. For example, Ugg remained at the bottom throughout the period (though we note its position improved considerably since the January data collection described earlier). Meanwhile, Bvlgari stayed near the top position throughout the period. We do observe a few substantial movements, however. For example, Louis Vuitton fell four positions, from 16th to 20th most counterfeited, while Air Jordan rose from 20th to 15th most counterfeited brand.

Figure 6 presents another way to look at the data, which may help reveal what is driving these fluctuations. It shows the average number of counterfeit websites added and re-

Figure 2: Bump chart tracking the prevalence of counterfeit search results in brands over time.

	Avg # counterfeit stores per week				Avg # counterfeit stores per week (page 1)	
	added	removed	Δ stores	Δ rank	added	removed
nike	32.4	27.8	4.6	1	1.6	1.2
air jordan	36.9	33.9	3	5	4.1	4.1
hermes	28.1	25.6	2.5	1	4.1	3.8
panerai	31.4	30.7	0.7	-2	4.6	4.3
armani	39.1	38.5	0.6	1	4.4	3.6
gucci	33.3	32.7	0.6	1	2.4	3.1
louis vuitton	35.8	35.3	0.5	-4	4.9	4.4
fendi	32.7	32.3	0.4	1	2.2	2.2
breitling	35.2	34.9	0.3	5	3.6	3.7
hublot	34.1	33.8	0.3	1	3.8	4.4
chanel	23.6	23.3	0.3	0	3.1	2.9
versace	31.3	31	0.3	2	3.8	3.6
bvlgari	36.2	36	0.2	0	4.9	4.6
adidas	22.3	22.2	0.1	-1	2.1	2.4
rolex	27.2	27.1	0.1	2	2.8	2.8
coach	14.8	14.9	-0.1	0	2.2	1.8
ugg	12.9	13	-0.1	0	0.5	0.6
tag heuer	29.2	29.4	-0.2	1	1.9	1.9
cartier	34.4	34.9	-0.5	-1	3	2.9
burberry	34	34.6	-0.6	0	4	4.2
oakley	45.5	46.6	-1.1	-2	4.3	4.6
patek philippe	30.3	31.5	-1.2	-3	2.7	3
iwc	41.9	43.2	-1.3	0	4.1	4.2
dior	38.6	40.3	-1.7	-4	3.8	4.6
prada	31.2	35.1	-3.9	-4	1.6	2.2
average	31.7	31.54	0.15	0	3.22	3.24

Table 6: Turnover of bad stores by brand over 12 week collection

moved, per brand, during the 12-week period. Brands that remove more fake websites than are added are likely to move down the rankings, while those who add more fake websites than are removed tend to rise in the rankings. The table shows this by plotting this delta, as well as the net change in rank from the first and last observation.

Overall, across brands the replacement-rate of counterfeit websites appears to be consistent over time, even within the top page of results. This suggests that whenever bad websites are taken down, they are consistently replaced by other bad websites rather than by legitimate ones. Likewise, as legitimate stores fall out of the search results they are generally replaced with other legitimate stores. Overall, the brands do appear to be engaged in an endless game of whack-a-mole with no end in sight.

4.4 Does replicated content replace removed websites?

We have just established that there is substantial turnover in websites selling fakes. We now study the relationship between the old websites when they fall out of the search results and the new websites that replace them. To do that, we clustered the 3 622 websites in the June–August dataset that appeared in the results 4 weeks or less.

We clustered the websites based on image similarity. We compared vertical and horizontal similarity using luminosity histograms implemented using the Eye.Open image library for C#[2]. We then conservatively clustered websites together using hierarchical agglomerative clustering with cut-height of 25%. In total, 1 389 websites were placed into clusters of at least size 2.

Figure 3 plots the timing fake websites first appear and disappear in the search results, grouped by cluster, for all clusters containing at least 10 websites. Each group is assigned a different color and line style for a given cluster size. The purpose of this graph is to visualize the extent to which websites selling fakes are published in parallel or serialized. We can see, for example, that the cluster with 22 websites is highly parallel, with 14 near copies of the same page appearing in the search results at once. One of the clusters of size 16 (the blue dashed grouping), by contrast, has no more than four website copies in operation at once, suggesting a more serial operation of replacing websites that are removed.

We can quantify the tendency towards serial or parallel operations by calculating the maximum *depth* clusters have. Here depth is defined as the largest number of websites with overlapping time intervals appearing in search results. Clusters with larger depths are more parallelized, while those with smaller depth values are more serialized. Table 7 reports the average depths for different cluster sizes. We can see that, in general, the average depth for clusters is much smaller than the cluster size (e.g., depth 2 for the 3 clusters of size 13). This suggests that a modestly parallel, but primarily serial, approach is being used by websites selling fakes. Once their websites drop out of the search results, they quickly replace them with new ones with content copied

[2] `https://similarimagesfinder.codeplex.com/`

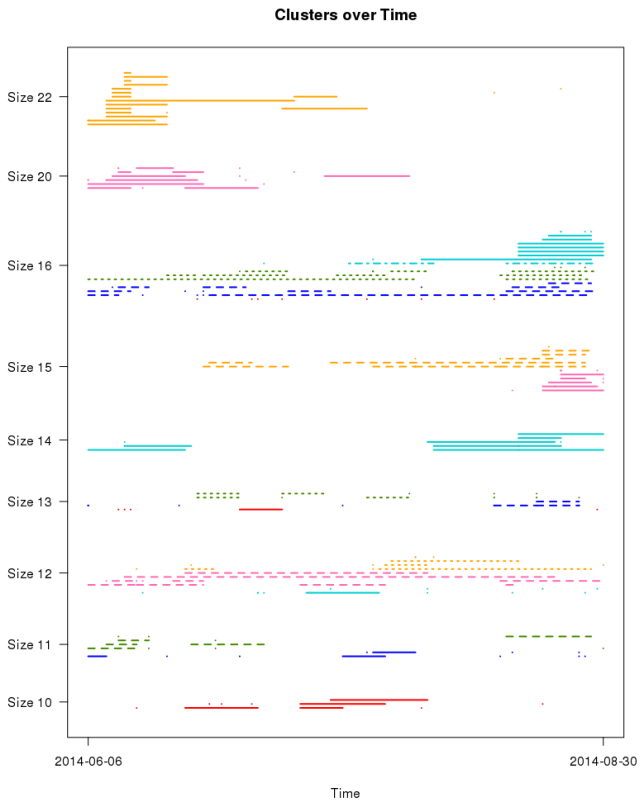

Figure 3: Clustered websites selling counterfeits grouped by order of appearance in the search results.

over from the dropped ones.

4.5 What are the characteristics of websites selling fakes?

Up to now, the analysis has focused on how the prevalence of counterfeits in search results varies. We now examine the counterfeit-hawking websites themselves to explore how they differ from websites that do not sell counterfeits.

First, Figure 4 plots the number of brands a given store appears in the search results for. Most stores were found by searching for only only one brand (969 stores), suggesting that most websites specialize in selling a single type of counterfeit good. This may reflect a search-engine optimization strategy moreso than a supply chain issue, as it is not expensive to create distinct storefronts. Some stores do sell multiple brands, however, which is to be expected when there are multiple brands in a given category (e.g., watches). Only a very small number of stores are like bazaars, however, selling more than 10 of the 25 brands we monitored.

We next study how the time since a website has been registered affects the likelihood that it will be selling counterfeits. Out of the 6 979 unique FQDNs encountered during data collection, we could extract the website's creation date from the WHOIS in 3 933 cases (56.35%). Figure 5 illustrates the subset of data for which this data was collected. Websites registered for less than a year were much more

Cluster size	# in cluster	Avg. depth
1	2092	1.0
2	250	1.1
3	64	1.3
4	33	1.7
5	15	1.9
6	10	2.2
7	8	1.9
8	7	2.4
9	5	2.2
10	1	3.0
11	2	3.0
12	3	3.3
13	3	2.0
14	1	5.0
15	2	6.0
16	4	4.8
20	1	6.0
22	1	14.0

Table 7: Cluster sizes and average depths for similar fake websites appearing in search results between June–August.

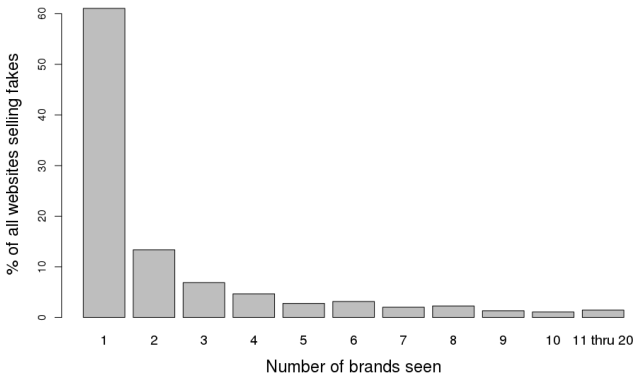

Figure 4: Brands per store (as measured by search queries).

likely to be identified as selling counterfeits, whereas the websites identified as not selling counterfeits tended to be older. Note that despite the trend, there are a large number of exceptions to this rule.

The last website-specific characteristic we study is its associated country. There are two ways to identify a website's geographic location. One is to use the IP address to find out where the website itself is hosted. Another approach is to identify the registrant's country from the WHOIS details. Of course, these countries need not be the same. It is quite common to register a web address in the owner's home country, but host the content elsewhere, particularly in countries with widespread web hosting infrastructure (e.g., the US, Netherlands and Germany).

Thus, we use both methods when examining all January search results for websites selling knockoffs and others. Ta-

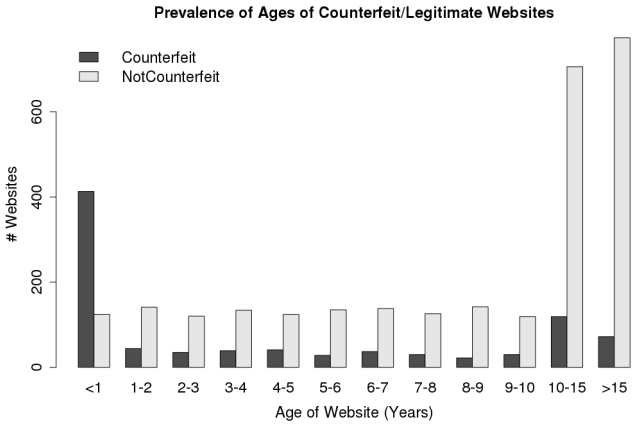

Figure 5: Websites' ages and counterfeit status.

Country	Hosting Country Odds ratio	95% C.I.	WHOIS Country Odds ratio	95% C.I.
United States	1.00	–	1.00	–
Australia	1.12	(0.666,1.81)	**3.89**	(2.420,6.20)
Belgium	–	–	3.58	(0.874,12.99)
Canada	1.17	(0.713,1.86)	1.05	(0.496,2.02)
China	**2.41**	(1.296, 4.38)	**17.88**	(13.492,23.96)
Czech Republic	**2.20**	(1.170, 4.01)	**2.67**	(1.307, 5.18)
Denmark	1.73	(0.772, 3.62)	2.72	(0.539, 10.71)
Estonia	**14.57**	(7.967,28.93)	–	–
France	**0.37**	(0.172, 0.70)	0.95	(0.468,1.76)
Germany	0.72	(0.492,1.04)	0.72	(0.343,1.34)
Hong Kong	0.43	(0.063, 1.50)	2.48	(0.851, 6.41)
India	1.23	(0.334, 3.60)	2.10	(0.903,4.46)
Ireland	1.62	(0.886, 2.85)	–	–
Italy	**0.48**	(0.196,0.99)	0.88	(0.400,1.72)
Japan	–	–	**8.65**	(2.027,45.11)
Malaysia	**9.42**	(4.804,19.89)	–	–
Netherlands	**4.57**	(3.293,6.37)	0.47	(0.070,1.61)
Panama	**6.21**	(1.824,24.64)	**6.55**	(4.281,10.09)
Private/Proxy	–	–	1.25	(0.967, 1.60)
Russia	**8.94**	(4.659,18.32)	–	–
Singapore	1.31	(0.573, 2.74)	–	–
Spain	0.85	(0.311, 1.95)	1.79	(0.693, 4.09)
Sweden	**7.52**	(5.243,10.96)	0.38	(0.016, 1.85)
Switzerland	**0.25**	(0.058,0.68)	**0.19**	(0.045, 0.51)
Thailand	–	–	**4.27**	(1.007, 16.84)
Turkey	**3.61**	(1.302,9.98)	–	–
United Kingdom	1.20	(0.936, 1.53)	**1.96**	(1.496, 2.56)

Table 8: Odds ratios indicating the relative prevalence of websites selling fakes compared to the US (for both hosting and WHOIS registration). Bold figures indicate statistically significant risk factors.

ble 8 presents odds ratios for the most popular countries from the results for both hosting and WHOIS countries associated with websites. The odds ratios are calculated relative to incidence of fake and legitimate websites hosted and registered in the US. Hence, any odds ratio greater than one (and highlighted in red bold font to indicate statistical significance at 95% confidence) indicate that the country is a positive risk factor for having counterfeit websites. Likewise, any odds ratio less than one (and highlighted in green bold font) is a negative risk factor for having counterfeit websites.

A number of trends are apparent from inspecting the odds ratios in Table 8. Websites selling fakes have 17-times greater odds of being registered in China than the US. However, while the fake website may be registered to a Chinese person or business, the website itself is only twice as likely to be hosted in China than in the US. Japan is also at much greater odds (8 times), but there are fewer Japanese websites in the results. Private and proxy WHOIS registrations are more likely to be associated with fakes, but at only 25% greater odds (and just outside statistical significance.

Countries with large legitimate luxury goods industries (France, Italy, Switzerland) are much less likely to either host websites selling fakes or have them registered by entities in those countries. In contrast, countries with large web hosting industries (Netherlands, Estonia, Sweden) are more likely to host websites selling fakes, even though the websites are not likely to be registered there.

In sum, we can see that the infrastructure for supporting websites selling counterfeit goods is global, and that counterfeit producing countries (e.g., China) are more likely to register the websites and countries with strong IT infrastructures are more likely to host the websites.

5. RELATED WORK

In very recent work carried out concurrently to our own efforts, Wang et al. use clustering techniques to identify "campaigns" of similar websites advertising counterfeit goods [32], using methods described in [12]. They also issue search queries for brands, but they identify websites selling knockoffs by looking for signs that the hosting website has been hacked and is demonstrating cloaking behavior. Their analysis in turn focuses on linking together disparate websites into groupings. Our work complements theirs in that we focus on the more general problem of classifying all search results as selling knockoffs or not. Cloaking behavior is indicative of many, but certainly not all, of today's websites selling counterfeits. One way we can see this is to note that 45% of the websites we identified as selling counterfeits also mentioned the brand on their homepage. This suggests that many of these websites are not hacked, but instead are brazenly selling fakes. Furthermore, we focus our analysis on examining differences in the prevalence of counterfeits in web search by user intent and brand characteristics.

More broadly, a number of papers have investigated abuse in search-engine results. Provos et al. presented a mechanism for identifying drive-by-downloads in web search results [27]. Moore et al. [25] and John et al. [14] report on the poisoning of trending search terms to distribute malware and host ad-laden, auto-generated content. Leontiadis et al. document search-poisoning by those peddling counterfeit pharmaceuticals [17]. The same authors recently reported on a longitudinal study of such search-engine poisoning promoting unlicensed pharmacies [16]. Notably, they compared the prevalence of search poisoning based upon the intent

of the search queries, finding that both innocent and complicit queries turn up unlicensed pharmacies. This complements our own findings regarding the presence of knockoffs in search results, regardless of query intent.

A number of papers have proposed classifiers to identify malicious web content. Abu Nimeh et al. compare several methods for classifying phishing websites [2]. Many others have constructed features for classifying malicious web pages based upon website content or behavior [6, 8, 26, 31, 33]. Our paper continues in this tradition, but builds a classifier based upon features specific to websites selling knockoffs (e.g., selling in multiple currencies, pricing information).

6. CONCLUSION

The web has revolutionized commerce, giving consumers access to more choice at lower prices. Unfortunately, it can be hard to determine whether the great deal found online is truly a bargain or is actually cheap because the merchant is selling knockoffs. In this paper, we have conducted a large-scale empirical analysis of 25 counterfeit goods found through web search. We designed a purpose-built classifier to predict whether a given website found through search likely sells genuine merchandise or counterfeit goods.

We have found that 32% of inspected search results point to fakes overall, but we have also observed wide variation. Innocent queries such as "hublot buy online" are less likely to lead to fakes, but introducing the word "cheap" can lead to nearly 40% of the results pointing to stores selling counterfeits. Furthermore, some brands are targeted more often than others. Brands who sell high-end goods such as luxury watches tend to have their search results polluted with more knockoffs. Not all the news is bad for brands, however, as we have presented a linear regression that indicates those who actively protect their brand via DMCA enforcement experience much lower rates of fakes in search.

By and large, merchants selling fakes take advantage of reliable web hosting by operating in countries with strong infrastructure. They also tend to replace removed websites with copied content on new URLs.

In future work, we hope to continue measuring progress in combating the sale of counterfeit goods by carrying out longitudinal studies. More work can be done to improve the classifier's accuracy so that it can be used in an ongoing basis by operators in the field. We also hope to investigate similarities between websites selling fakes in greater depth.

Acknowledgments

This work was partially funded by the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD) Broad Agency Announcement 11.02, the Government of Australia and SPAWAR Systems Center Pacific via contract number N66001-13-C-0131. This paper represents the position of the authors and not that of the aforementioned agencies.

7. REFERENCES

[1] Alexa top 1 million websites. `http://s3.amazonaws.com/alexa-static/top-1m.csv.zip`.

[2] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. A comparison of machine learning techniques for phishing detection. In *Proceedings of the 2nd APWG eCrime Researchers Summit*, pages 60–69. ACM, 2007.

[3] George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):pp. 488–500, 1970.

[4] R. Anderson. Why information security is hard - an economic perspective. In *Proceedings of the 17th Annual Computer Security Applications Conference (ACSAC'01)*, New Orleans, LA, December 2001.

[5] R. Anderson and T. Moore. The economics of information security. *Science*, 314(5799):610–613, October 2006.

[6] Sushma Nagesh Bannur, Lawrence K Saul, and Stefan Savage. Judging a site by its content: learning the textual, structural, and visual features of malicious web pages. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, pages 1–10. ACM, 2011.

[7] Aaron Blum, Brad Wardman, Thamar Solorio, and Gary Warner. Lexical feature based phishing url detection using online learning. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security*, AISec '10, pages 54–60, New York, NY, USA, 2010. ACM.

[8] Davide Canali, Marco Cova, Giovanni Vigna, and Christopher Kruegel. Prophiler: a fast filter for the large-scale detection of malicious web pages. In *Proceedings of the 20th international conference on World wide web*, pages 197–206. ACM, 2011.

[9] Richard Clayton. WHOIS data extracted from templates (deft-whois), 2014. `http://www.deft-whois.com`.

[10] Richard Clayton and Tony Mansfield. A study of whois privacy and proxy service abuse. In *13th Workshop on the Economics of Information Security*, 2014.

[11] Mark Culp, Kjell Johnson, and George Michailidis. ada: An r package for stochastic boosting. *Journal of Statistical Software*, 17(2):9, 2006.

[12] Matthew F. Der, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Knock it off: Profiling the online storefronts of counterfeit merchandise. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1759–1768, New York, NY, USA, 2014. ACM.

[13] Jake Drew. Machine learning in parallel with support vector machines, generalized linear models, and adaptive boosting, 2014.

[14] John P. John, Fang Yu, Yinglian Xie, Arvind Krishnamurthy, and Martin Abadi. deseo: Combating search-result poisoning. In *USENIX Security Symposium*. USENIX Association, 2011.

[15] Markus Kammerstetter, Christian Platzer, and Gilbert Wondracek. Vanity, cracks and malware: Insights into the anti-copy protection ecosystem. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 809–820, New York, NY, USA, 2012. ACM.

[16] N. Leontiadis, T. Moore, and N. Christin. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of ACM CCS 2014*, Scottsdale, AZ, November 2014.

[17] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In *USENIX Security Symposium*, 2011.

[18] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. Pick your poison: pricing and inventories at unlicensed online pharmacies. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 621–638. ACM, 2013.

[19] K. Levchenko, N. Chachra, B. Enright, M. Felegyhazi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, A. Pitsillidis, N. Weaver, V. Paxson, G. Voelker, and S. Savage. Click trajectories: End-to-end analysis of the spam value chain. In *Proceedings of IEEE Security and Privacy*, Oakland, CA, May 2011.

[20] C. Littlejohn, A. Baldacchino, F. Schifano, and P. Deluca. Internet pharmacies and online prescription drug sales: a cross-sectional study. *Drugs: Education, Prevention, and Policy*, 12(1):75–80, 2005.

[21] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. Beyond blacklists: learning to detect malicious web sites from suspicious urls. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1245–1254. ACM, 2009.

[22] Niels Provos Panayiotis Mavrommatis and Moheeb Abu Rajab Fabian Monrose. All your iframes point to us. 2008.

[23] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. Voelker, S. Savage, and K. Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proceedings of USENIX Security 2012*, Bellevue, WA, August 2012.

[24] David Meyer. Support vector machines. *The Interface to libsvm in package e1071. e1071 Vignette*, 2012.

[25] Tyler Moore, Nektarios Leontiadis, and Nicolas Christin. Fashion crimes: trending-term exploitation on the web. In Yan Chen, George Danezis, and Vitaly Shmatikov, editors, *ACM Conference on Computer and Communications Security*, pages 455–466. ACM, 2011.

[26] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, pages 83–92. ACM, 2006.

[27] N. Provos, P. Mavrommatis, M. Rajab, and F. Monrose. All your iFrames point to us. In *Proceedings of the 17th USENIX Security Symposium*, August 2008.

[28] Markus Riek, Rainer Boehme, and Tyler Moore. Understanding the influence of cybercrime risk on the e-service adoption of European Internet users. In *13th Workshop on the Economics of Information Security*, 2014.

[29] German Rodriguez. Generalized linear models, 2014.

[30] Michael D. Smith and Rahul Telang. Competing with free: The impact of movie broadcasts on dvd sales and internet piracy1. *MIS Q.*, 33(2):321–338, June 2009.

[31] D. Wang, S. Savage, and G. Voelker. Cloak and dagger: Dynamics of web search cloaking. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, pages 477–490. ACM, 2011.

[32] David Y. Wang, Matthew Der, Mohammad Karami, Lawrence Saul, Damon McCoy, Stefan Savage, and Geoffrey M. Voelker. Search + seizure: The effectiveness of interventions on SEO campaigns. In *ACM Internet Measurement Conference (IMC)*. ACM, 2014.

[33] Steve Webb, James Caverlee, and Calton Pu. Predicting web spam with http session information. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 339–348. ACM, 2008.