Purchased Fame: Exploring the Ecosystem of Private Blog Networks

Tom Van Goethem
imec-DistriNet, KU Leuven
tom.vangoethem@cs.kuleuven.be

Najmeh Miramirkhani
Stony Brook University
nmiramirkhani@cs.stonybrook.edu

Wouter Joosen
imec-DistriNet, KU Leuven
wouter.joosen@cs.kuleuven.be

Nick Nikiforakis
Stony Brook University
nick@cs.stonybrook.edu

ABSTRACT

For many, a browsing session starts by entering relevant keywords in a popular search engine. The websites that users thereafter land on are often determined by their position in the search results. Although little is known about the proprietary ranking algorithms employed by popular search engines, it is strongly suspected that the incoming links have a significant influence on the outcome. This has lead to the inception of various black-hat SEO techniques that aim to deceive search engines to promote a specific website.

In this paper, we present the first extensive study on the ecosystem of a novel type of black-hat SEO, namely the trade of artificially created backlinks through private blog networks (PBNs). Our study is three-pronged: first, we perform an exploratory analysis, through which we capture intrinsic information of the ecosystem and measure the effectiveness of backlinks. Next, we develop and present an ML-driven methodology that detects PBN sites with an accuracy of 98.7% by leveraging various content-based and linking-based features intrinsic to the operation of the ecosystem. Finally, in a large-scale experiment involving more than 50,000 websites, we expose large networks of backlink operations, finding thousands of websites engaged in PBNs.

ACM Reference Format:
Tom Van Goethem, Najmeh Miramirkhani, Wouter Joosen, and Nick Nikiforakis. 2019. Purchased Fame: Exploring the Ecosystem of Private Blog Networks. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS '19), July 9–12, 2019, Auckland, New Zealand*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3321705.3329830

1 INTRODUCTION

Every second, users enter more than 60,000 search queries on Google [7]. In more than half of the cases, users click through to one of the first three results that are returned, according to a study by Advanced Web Ranking [20]. As search engines are one of the critical drivers of organic traffic, being ranked higher than similar businesses can provide significant competitive advantages. The prices for Search Ads, advertisements that are shown on top of the result pages and may cost more than $50 per click [32], are indicative of the monetary incentive to score high in search results.

There exist numerous black-hat SEO techniques that can be leveraged to improve the rank of a website. For instance, by overloading a page with related keywords, the relevance score of the search algorithm can be manipulated [19], or by ranking low-quality websites, often filled with malicious content or bloated with advertisements, for trending search-terms [16]. As a result of the continuously improving detection of SEO abuse [23, 26], many of the black-hat SEO techniques are rendered mostly ineffective.

As a consequence of this ongoing arms-race between search engines and miscreants who try to artificially boost their websites higher in the search results, the black-hat SEO techniques keep evolving. In this paper, we explore the underlying infrastructure of a relatively new technique, called private blog networks (PBNs), that shows a change in the paradigm of SEO abuse: instead of exploiting single aspects of the ranking algorithms, PBNs leverage the way these algorithms are intended to operate, i.e. promoting websites with links originating from legitimate, trustworthy sources. To this extent, PBN operators set up networks consisting of websites that are purposefully created to appear legitimate. To analyze the different actors in this ecosystem, we develop a multi-step machine learning classifier that leverages both content-based and linking-based features inherent to the operations of PBNs with an accuracy of 98.7%. As part of a large-scale experiment on 52,777 websites, our classifier manages to detect 3,552 PBN sites. Furthermore, we study the PBN customers and find that the decision to purchase backlinks is often financially motivated, either to promote a business or to drive more users to websites that generate money through advertisements. Finally, despite the efforts of PBN providers to hide their network, we manage to detect several clusters, ranging from a handful of websites to several hundred websites that are controlled by a single entity. We conclude that the backlink ecosystem is highly lucrative both for the providers, who can generate a turnover of more than $100,000 per month, as well as for their customers, who can attract many more visitors to their websites at the expense of their competitors.

In summary, we make the following contributions:

- We perform the first comprehensive study of the ecosystem of artificially created backlinks on PBNs, identifying the involved entities and analyzing their interactions.

AsiaCCS '19, July 9–12, 2019, Auckland, New Zealand

ACM ISBN 978-1-4503-6752-3/19/07…$15.00
https://doi.org/10.1145/3321705.3329830

- We develop a novel methodology driven by a multi-step machine learning algorithm that can be used to detect PBNs and associate the involved domains with a high accuracy.
- Leveraging this method, we perform a large-scale scan on more than 50,000 websites, and discover thousands of domains that aim to boost the reputation of their customers. We manage to cluster together several networks and find that providers may employ up to several hundred websites to promote their customers.
- As search engines are continuously improving their techniques to detect backlink abuse, PBN providers have to resort to more extensive measures to avoid detection.

2 EXPLORATORY ANALYSIS

In order to evaluate the current state-of-practice of SEO abuse, we perform an exploratory experiment and report on a novel black-hat SEO technique named private blog networks (PBNs).

2.1 Experimental setup

As the first step of our analysis, we explored which types of backlink services are currently provided. To this end, we searched for backlink-related phrases such as "buy backlinks" and "improve SEO ranking" and analyzed the most prominent results. Furthermore, we evaluated the backlink packages that were offered on marketplaces specialized in SEO techniques, such as SEOClercks [24] and KonKer [9]. Although most services include the number of links that would be created, only a few services reveal which technique they use to create backlinks. To obtain more information about the type of backlink services that are provided and gain in-depth insights into the workings of the ecosystem, we purchased several services that are representative of the market and followed the process presented in Figure 1. In total, we purchased 12 backlink services, covering three different price ranges: low-end backlinks costing $6, mid-end backlinks that we purchased for $24-30, and, lastly, high-end backlinks which cost $59-86.

After purchasing the backlink services, we were asked to provide information such as the target URLs and keywords that are related to the promoted website. To this end, we set up 12 websites, which we created by registering recently expired domain names, and serving a prior version of the website that was obtained through Internet Archive's Wayback Machine. The reason for reviving expired websites is twofold: first, this method yields legitimate-looking websites, preventing the providers from discovering they are being analyzed and second, as search engines may treat new websites differently than existing ones, this could increase bias in our results. To further reduce the consequences a change in domain ownership can have on a website's reputation, the websites remained in an idle state, i.e. no content was added or modified, for at least 4 months before any backlink service was purchased.

2.2 Backlink creation

As soon as the backlinks have been created, the backlink provider reports back the completion of the service and typically includes a list of pages that contain a link back to the customer's domain. This allows the customer to verify the fulfillment of the order. In Table 1, we show the number of links that were delivered by each

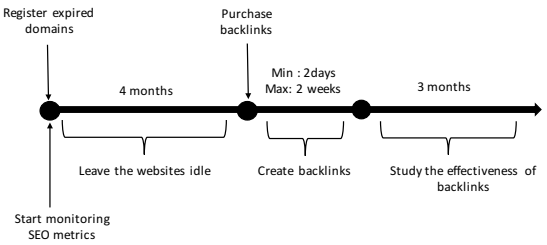


Table 1: Overview of purchased backlink services

Site #	Service/marketplace	Backlink type	Price	# Links
1	Fiverr	Profile abuse (.edu, .gov)	$6	26
2	SEOClercks	Profile abuse & social	$6	74
3	Fiverr	Profile abuse	$6	37
4	Fiverr	PBN	$6	27
5	KonKer	PBN	$24	5
6	SEOClercks	PBN	$30	5
7	KonKer	PBN	$30	5
8	KonKer	PBN	$30	5
9	Sape Links Network	PBN & links on homepage	$72	-
10	BackLinks.com	PBN	$86	16
11	Marketing1on1	Social, guest post, PBN directory listing, comments	$72	281
12	KonKer	PBN	$59	7

backlink provider. By manually inspecting these links, we classified the types of backlinks created. As can be seen in Table 1, the cheaper backlink services mainly create backlinks on existing websites by creating fake profiles and adding a link in the profile section of each user to their client's URL. For all other backlink services, the providers created links on their private blog network (PBN). These are networks of websites that are specifically used for creating links to their customers, and thus boost their ranking. Typically, the websites are structured as blogs, where each newly created link is associated with a blog post entry containing several hundred words of content related to the linked website.

2.3 Backlink effectiveness

To develop an intuition of how effective backlinks are in promoting their customers, we obtained various heuristics that reflect a domain's reputation on a daily basis throughout the duration of our experiment. More precisely, from the Moz service [17], we obtained our test websites' backlinks and their Domain Authority, which is a proprietary search engine ranking score that predicts how well a website is ranked within search results [18]. From the Majestic service [15], we obtained the Citation Flow and Trust Flow metrics [15]. Citation Flow ranges from 0 to 100, and reflects the "power" of a link from any given website. In conjunction, Trust Flow demonstrates the quality of a particular website and is increased when trusted websites link to this domain.

While these metrics aim to represent the ranking algorithm employed by search engines, it is possible that significant discrepancies occur, especially under the conditions of attempted manipulation. As popular search engines no longer provide any ranking score or metric [28], we evaluated the direct difference in ranking for

relevant keywords. More concretely, for each of our test websites, we composed three relevant queries and queried Google through their website and Bing through their API for each query. To avoid bias, we cleared all browser cookies before making a new query and ensured the service was contacted from the same geographical location over the duration of the experiment. We performed all queries on a daily basis and recorded the top 50 results.

In Figure 2, we show four of the aforementioned metrics of websites' reputation. When considering Citation Flow, we can see that for the low-end and mid-end backlink services, there is no apparent change in ranking after the links were created. For specific services, e.g., the one purchased for Site #2, the backlink service had an adverse effect, thus *decreasing* the ranking. For all of the high-end backlink services, we observed an increase of the Citation Flow metric around the time the backlinks were created. However, for all except one, this effect was temporary and after two months, the ranking dropped back to a value lower than the original.

The Domain Authority metric, reported by Moz, shows a similar trend as Citation Flow. For the low-end and mid-end backlink services, the decrease in ranking over a longer time was more pronounced. Interestingly, for the high-end backlinks, the two services that improved Citation Flow the most (for Site #10 and Site #11), did not have the same positive effect on the Domain Authority ranking. This result highlights that search engines use different metrics to compute ranking, so techniques that have little to no impact on one search engine may prove successful with another.

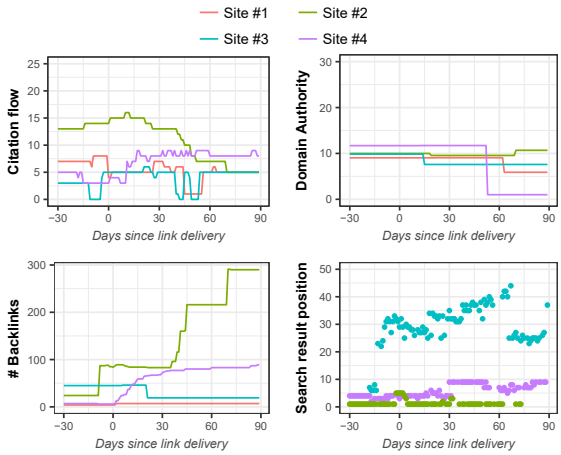
An interesting outlier with regard to the number of backlinks is Site #12: the backlink provider only reported 7 web pages on which backlinks were created. However, the Majestic service found that a month after the transaction was completed, 281 backlinks had been created, which is most likely due to a mistake by the backlink provider. As content is often automatically generated and placed on the backlinking websites, we suspect that the backlink provider mistakenly placed backlinks on all, or at least on a larger-than-intended fraction, websites in their network.

Last, we evaluate the rank of a website in the search engine results pages. In Figure 2, we show the evolution of the first occurrence of each website in the results of a single query. For brevity, we only show the results of a single search query, as returned by Google. This query was selected based on the number of correctly obtained results for which the target website could be found within the top 50 results. It should be noted that on multiple occasions, our IP address was blocked from the Google Search Engine because of the automated nature of our experiment. As a result, these metrics could not be obtained daily.

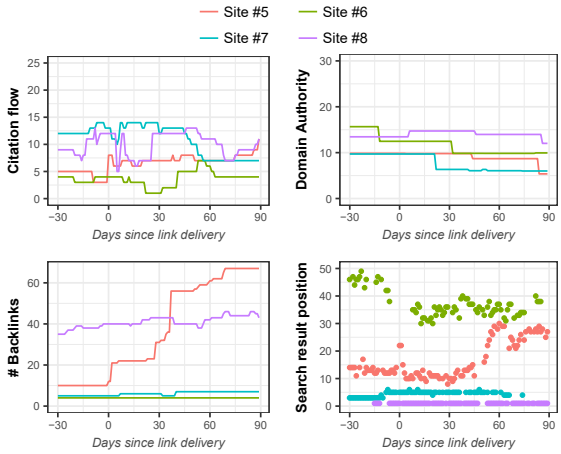
In contrast to the Citation Flow and Domain Authority metrics, there is no apparent improvement in the ranking of search results for high-end backlink services. Moreover, Site #9 and Site #12 experienced a *decrease* in search ranking, i.e. the position at which they are displayed increases after the backlinks had been created. Possibly, this is because Google's ranking algorithm is able to detect that the back-references originate from websites involved in SEO-abuse, thus leading to a penalty for the targeted websites. Note that even though this behavior is the opposite of what was intended, it could be abused to negatively affect the ranking of competitors.

For most websites, the Bing Search Results showed a similar pattern as Google's. However, for 4 websites, namely Site #2, #7,

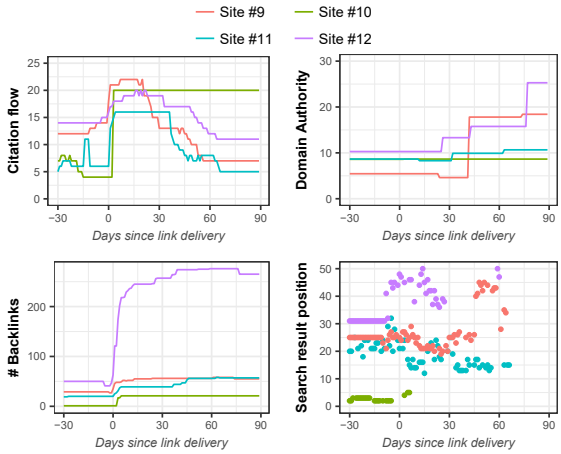
Figure 2: Evolution of Citation Flow, Domain Authority, number of backlinks discovered by Majestic and position in Google search results (lower is better), ranging from 30 days before until 90 days after link delivery.



(a) Heuristics for low-end backlink services

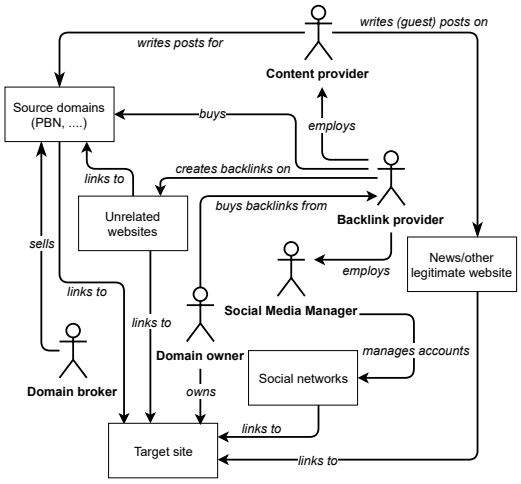


(b) Heuristics for mid-end backlink services



(c) Heuristics for high-end backlink services

Figure 4: Overview of the backlink ecosystem.



#8, and #12, we found that the position of the website in the search results improved from being outside of the top 50 to be the first result after the backlinks had been created. Interestingly, backlink services in all price ranges exhibited this behavior which indicates that the mechanisms that detect artificial backlinks and penalize websites purchasing them, are specific to each search engine.

Although our set of evaluated backlink services is not sufficiently large to make conclusive claims on the effectiveness of purchased backlinks in general, our results indicate that in several cases, backlinks are able to affect the reputation of a website. Furthermore, there is an apparent difference in the ability of search engines and companies that provide SEO-related reputation metrics to detect new types of abuse. This shows that there is a need to improve the detection of websites arranged in these types of private blog networks. In Section 4, we propose such a detection mechanism based on a multi-step machine learning classifier.

3 ECOSYSTEM OVERVIEW

As a result of our exploratory analysis, we identified multiple parties who each offer specialized services within the backlink ecosystem. Furthermore, we identified that there are two main classes of abusive backlinks provided: one class aims to leverage the reputation of existing websites by insidiously creating links to customers, e.g. by leaving comments or creating fake user profiles. The other class leverages the reputation of domains that are owned by the backlink providers themselves. In this section, we provide an overview of the backlink ecosystem and focus on private blog networks.

3.1 Roles in the backlink ecosystem

Figure 4 shows an overview of the backlink ecosystem. To arrive at this ecosystem, we read a large number of posts on SEO forums and blogs as well as the documentation available on the websites offering backlink services. Finally, all of our findings were interpreted in light of our backlink-purchasing experiments described in Section 2. It is important to note that not all roles described in the following paragraphs need to be simultaneously involved in every operation, and that a single party may assume multiple roles.

Domain owners are the entities who want to promote their website (*Target site*) by purchasing backlink services.

Backlink providers offer a service where, in exchange for money, they create artificial links pointing to the website of their customers. These links can originate from various sources: backlink providers may try to leverage authoritative domains that are not under their control, by creating fake profiles or inserting comments. One of the goals of backlink providers is to build backlinks that resemble organic linking behavior, in order to avoid detection. As such, backlink providers may employ the specialized services offered by "Social Media Managers" and "Content providers", who respectively generate links on social media sites and create content in the form of blog posts related to the targeted website. Finally, backlink providers may offer links originating from their own domains (referred to as "Source domains"). These are websites under the control of the backlink provider.

Content providers write original high-quality text that accompanies the blog post linking to the PBN customer. This is done because search engines may take into account several features about the page that an outgoing link is placed on, such as quality of the content and relevance to the topic of the targeted website. Alternatively, to save on costs, backlink providers may use content generation algorithms, such as text spinners [25].

3.2 Private Blog Networks

To create links to their customers, backlink providers own a number of domains that host websites containing a blog or another type of content-management system (CMS). This allows the backlink provider to automate the process of placing articles that contain links to their customers. In most cases, backlink providers attempt to grow their network with domains that already have a number of incoming links from unrelated websites, and thus are reputable from the viewpoint of search engines.

When a new website is added to the network, a backlink provider may try to improve its ranking by having the other sites link to it. While this practice may have advantages for the whole network of websites — especially when it is large enough such that not every website links to all other websites — it makes detection easier for search engines. In 2013, Google uncovered a large network named Anglo Rank [22]. Consequently, the websites involved in the network were penalized, rendering all domains useless for the purpose of SEO boosting. Additionally, the domains of the customers who purchased this service were penalized as well which again demonstrates that backlinks can be purchased to either boost one's own domain or damage the reputation of a competitor website.

4 DETECTION OF PBN SITES

In this section, we present a technique to detect websites belonging to PBNs. A naive approach to this would be to randomly select websites and classify these. However, because the vast majority of websites on the web are unrelated to the PBN ecosystem, this approach would result in a highly unfavorable base rate, causing even a very good classifier to produce a high number of false positives. Therefore, in this work, we take advantage of an efficient guided search approach which starts from an initial seed of known PBN websites and leverages the linking behavior intrinsic to the

backlink ecosystem. Next, using a two-step classifier, we identify PBNs with an accuracy of 98.7%. By applying the classifier on a set of 5.8M pages on 52.7K websites, we detect thousands of PBN sites.

4.1 Guided search to find PBNs

Our data collection methodology consists of multiple forward and backward searches to incrementally build a *backlinks graph* where nodes are websites, and edges represent links between them. In a forward search step, we visit cross-domain links extracted from a website while in a backward search step, we visit the backlinks of a given node. The details of the algorithm are as follows:

(1) Backward search to build the initial seed of PBNs
As the first step, we start to compose a set of websites that have been determined to be PBN sites: websites that created backlinks to our domains as part of our exploratory analysis.

(2) Forward search to find potential PBN customers
For every PBN site in the initial set, we crawl up to 200 web pages and record all cross-domain links. These contain links to other PBN customers as well as to legitimate websites that are not involved in the backlink scheme. Consequently, the linked-to websites are labeled as *potential* PBN customers.

(3) Backward search to find potential PBN sites
Next, we obtain a list of backlinks pointing to any of the potential PBN customer domains. This list includes links from PBN sites (in case the customer did in fact purchase backlinks), and links from other websites that link to the alleged PBN customer for legitimate reasons. These websites can then be classified into PBN and non-PBN sites. Note that these have a higher probability of being PBN compared to a random sample of the web, as these are linking to potential PBN customers.

The above interleaved forward and backward searches can be repeated, where after the completion of the three steps, the detected PBNs can serve as a new seed for the first step. In a first iteration, we started with a seed of 50 PBN sites, and manually labeled the output of step (3), i.e. 1,027 sites linking to potential customers. We found 252 of these to be PBN sites, and use these as the seed for a second iteration. By applying the three phases of the aforementioned method, we find 52,777 new websites that link to potential customers. We use a multi-step classifier to label these, as described in the following sections.

4.2 PBN classifier

We develop a two-class (PBN and non-PBN) multi-step classifier built upon two sets of features, namely *content-based* and *linking-based*, which are intrinsic to the two key characteristics of PBNs operation. The insight behind *content-based* features is to capture common structural similarities of PBN websites while the second set of features represent linking behavior of the backlinks ecosystem.

Content-based features rely on the fact that PBN operators need to rapidly and automatically create content and include links to their customers while evading search engines. These types of development practices result in having similar content structures. The key insight used to design *linking-based* features is that PBN operators, by definition, have to link to their customers. We take advantage of such linking behavior of the network and formulate them as classifier features. For example, the feature that captures the number of links to PBN customers, should have a completely different distribution in PBNs and non-PBNs.

4.3 Content-based classifier (step 1)

PBN operators develop their websites rapidly, mostly using popular CMSs such as WordPress, use a template to automatically insert the links of their customers, and try to balance between the content and links to stay under the radar of search engines. From our analysis in Section 2, we observed that because of the way they deploy their websites, PBN websites share common structures. As such, *content-based* features are designed to capture this common structure.

4.3.1 Content-based features. There exists an abundance of features that can be collected from websites. In the iterative process of feature engineering, we opted for features that are an intrinsic part of the backlink ecosystem. Furthermore, we selected features that either would be hard to circumvent, e.g. the customers that are linked to, or come at a certain cost. For instance, while it is possible for PBN providers to create many distinct-looking web pages, this would require a significant cost and potentially make the backlink service unprofitable.

c_1: Number of words in anchor text for cross-domain links
When purchasing a PBN service, the customer is asked to enter a number of keywords of their promoted website. Typically, customers enter only one or two words, which are then used as the anchor text. As a result, the median number of anchor words is oftentimes lower than with regular sites.

c_2: Alexa ranking of cross-domain links
The cross-domain links of PBN sites are mainly to their customers, which are typically not widely popular. As such, the average Alexa rank of cross-domain is relatively low. Moreover, we found that several PBN sites also link to high-profile sources, i.e. for every customer link, a link is created to a high-profile site. Most likely, this is done to appear more authoritative. By incorporating the first and third quantiles of the Alexa ranking, we also capture this information. As PBN sites are required to link to their customers, this feature is difficult to evade without incurring a significant cost, either in the number of PBN sites or the effectiveness of the PBN boosting.

c_3: Number of cross-domain links
The main goal of PBN sites is to link to customers, so it is not uncommon for them to have many outgoing links. Furthermore, because a new page is created for every new customer that is linked to, most pages have only a single cross-domain link. This makes the distribution of cross-domain links per page significantly different from most non-PBN sites. Although this feature can be evaded by randomizing the number of links on every page, this does require more effort by the PBN operator.

c_4: Visually similar pages
As we showed in Section 5.1.5, many PBNs make use of a CMS (70.48% of the PBN sites are powered by WordPress), allowing them to easily create new pages. As a side-effect, many of the pages on the website look visually similar. We make use of a perceptual hash to determine the visual similarity of web pages. Although this feature can also be circumvented by creating unique-looking pages, this does require a significant effort by the PBN operator,

on the one hand on the design of unique-looking web pages and on the other hand to orchestrate publishing new content.

c_5: Visually similar pages (after removing images)
This feature is similar to c_4, but accounts for PBNs that include a unique image for every new "blog entry".

c_6: Similar DOM structure
This feature also leverages the fact that many PBNs make use of CMSes. In contrast to c_4 and c_5, which exploit the perceptual information of different web pages, this feature looks at the similarity of how the different HTML elements are structured. In addition to the visual similarity, it also makes it more difficult for PBN operators to evade, since they can not simply alter CSS properties but really have to alter the way they compose web pages.

c_7: Cross-domain link URL length
Upon purchasing a PBN package, customers are asked to provide a link to the domain that needs to be promoted. Customers often provide a link to the home page. As such, PBN sites are more likely to have cross-domain links to URLs with an empty path.

c_8: Cross-domain links in DOM siblings
This feature was included to improve distinguishing PBN sites with sites that suffer from comment abuse. Sites whose comment mechanism is abused to create many new links may have certain features that resemble PBN sites. Something that distinguishes them is that most of the links are in DOM siblings, as these are all comments at the end of the document.

c_9: Links to domains outside of Alexa 1M
This feature is similar to c_2, but only takes into consideration the number of domains that are not present in the list of 1 million most popular sites according to Alexa, i.e. presumably PBN customers.

c_{10}: Unique domains linked to
PBN sites have to link to their customers to boost their SEO score. Consequently, the unique number of domains that a PBN site links to is significantly higher compared to non-PBN sites.

c_{11}: URL length, sorted by number of cross-domain links
Many CMSes provide a functionality where they group together posts within a certain category, or posts that were made in a certain month. For the displayed post entries, a brief summary of the beginning of the post is shown. These aggregation pages have typically a short endpoint, combined with a large number of cross-domain links. Evasion of this feature is feasible, and require PBN operators to disable this aggregation function. On the other hand, this would make all pages even more similar.

c_{12}: Links with `text-decoration: none`
This feature was introduced to distinguish PBNs from other types of SEO abuse, namely when adversaries are creating pages with many links that do not appear as such, i.e. the default underlining of links has been disabled by setting the CSS property `text-decoration` to the value `none`.

c_{13}: Links with `rel=nofollow`
Overall, PBNs have relatively little links with the `rel=nofollow` attribute, as they intend to promote their linked customers. Additionally, other types of SEO abuse, such as comment abuse have a very high number of links with this attribute.

c_{14}: Unique words in the document
As we showed in Section 5.1, PBN sites have a different distribution of the number of words used per web page. This feature captures the median, average, minimum and maximum number of words found on web pages of a single website.

c_{15}: Number of web pages
As PBN operators typically create a new blog post entry for each customer they link to, the number of unique web pages found on a single PBN site can be considerably high. Circumventing this feature would require PBN sites to either create many more sites, or remove old blog entries (which would stop the SEO boosting); both options are unfavorable.

c_{16}: Number of words near links
For each link that is added, PBN operators add some text that is related to keywords provided by the customer. As such, links are typically found in the middle of a paragraph with PBNs.

c_{17}: Links that are close to each other
Similar to c_8, this feature was added to distinguish between different types of SEO abuse, in this case, link stuffing and comment abuse, where links are oftentimes placed very close to each other.

c_{18}: Alexa rank of the domain
This feature captures the popularity of PBN domain names.

For each website, we extract features from up to 200 pages. For the set of 52,777 sites from the guided search, we visit at total of 5,845,048 pages with a headless browser (Chromium) over the course of one month (September 2017) by leveraging a distributed setup of 15 VMs, each provided with 4 vCPUs and 4GB RAM.

4.3.2 Classifier Implementation. As a ground truth, we leverage the websites found after the first iteration of our three-phased guided search. By manually labeling these, we found 252 PBN sites and 775 non-PBN sites. We combine this list with 50 PBN sites that linked to one of our sites, as part of the exploratory analysis (the PBN sites are sampled to prevent an over-representation of a single provider), and a sample of 200 randomly selected non-PBN sites to ensure a representable ratio. As our dataset is sparse and contains outliers, we opt for Random Forest, which benefits from the strength of ensemble learning and is robust against overfitting. To evaluate our model, we use 10-fold cross-validation on the labeled dataset which reports an accuracy of 91% and an area under the ROC curve of 93.8%. However, as the dataset is imbalanced and we aim to label PBN websites (the minority class) precision and recall are more important evaluation metrics; these are 87.7% and 73.8% respectively. In the next section, we leverage the output of the content-based classifier to construct linking-based features, which are then used to construct another classifier, which improves precision and recall to more than 98%.

In table 2, we ranked the importance of the top 5 features according to their average impurity decrease. This metric calculates how much each feature decreases the weighted impurity of the trees. In general, cross-domain links features impacted the accuracy of our classifier more than the features that are based on the content of a website such as c_{14}, c_{15} and c_{18}.

4.4 Linking-based classifier (step 2)

In the second phase of our multi-step classifier, we construct a graph where the nodes are the sites that need to be classified (in this case 52,777 sites obtained from the second iteration of our three-phase guided search), complemented with a latent class *potential customer*. Instances of this latent class are websites that are being linked to by

Table 2: Importance of content-based features ranked by their average impurity decrease (AID)

Feature	AID
# words in anchor text for cross-domain links (c_1)	0.53
Alexa ranking of cross domain links (c_2)	0.44
# cross domain links in DOM siblings (c_8)	0.42
# Links with rel=nofollow (c_{13})	0.42
largest group of visually similar pages (c_4)	0.40

likely PBN sites (as determined by our content-based graph). Note that because our content-based classifier is not perfect and because PBN sites also link to unrelated sites, instances of this latent class are websites that have a larger likelihood of being a PBN customer.

4.4.1 Linking-based features. We extract five features from this graph and leverage these to build a second classifier. The features are mainly related to the linking behavior of PBN sites to their customers, i.e. the ratio of links pointing to PBN customers compared to non-customers links is significantly higher for PBN sites.

l_1: non-customers that are linked to
In our algorithm, websites are marked as a customer, i.e. as soon as a single site that is more likely to be a PBN site (according to our content-based classifier) links to it, we consider the site a potential PBN customer. As such, PBN sites will have relatively few non-customers they link to. To circumvent this, PBN sites would have to include many links to unrelated websites, a tactic that we observed on a few PBN sites during our manual analysis.

l_2: Average number of non-PBN links to linked domains
This feature captures the average number of non-PBN sites that link to the same sites as the site for which the feature is being computed. As PBN sites mainly link to their customers, which in turn are less likely to receive links from sites unrelated to the PBN ecosystem, the value for this feature will be significantly lower for PBN sites than for non-PBN sites.

l_3: Linked customers out of Alexa top 1M
As PBN operators can not choose their customers, and customers tend to be not-widely-known sites, which sometimes fall outside the Alexa top 1M, PBN sites will typically link to more unpopular, potential PBN customers than non-PBN sites. As these customer sites fall outside of the Alexa top 1M, they are less likely to receive links from sites outside of the backlink ecosystem. Although it has been shown that the Alexa list can be manipulated [11, 21], we consider it unlikely that PBN operators were aware of this at the time of our data collection. Alternatively, the Tranco list [11], which is more resilient to manipulation, could be used instead.

l_4: Number of customers linked to
Obviously, PBN sites will link to significantly more potential PBN customers. Similar to l_3, this feature is highly impractical, as the links to the customers need to remain in order to keep their SEO score boosted. The only alternative would be that PBN operators create significantly more PBN sites, which comes at a high cost.

l_5: Average number of PBN links to linked domains
This feature is similar to l_2, but instead focuses on other PBN sites that link to the same websites. The rationale of this feature is that a PBN site links to PBN customers, which in turn receive a lot of incoming links from other PBN sites. This feature is hard to evade because PBN customers require multiple incoming links in order to sufficiently boost their SEO score.

We use the manually labeled set of 1,277 sites (302 PBN, 975 non-PBN) to train a Random Forest classifier. Using 10-fold cross-validation for evaluation, we find precision and recall as 98.5% and 98.4% respectively, showing a significant improvement over just using the content-based classifier. By applying our model to the 52,777 sites collected from the guided search, our classifier marked 3,552 (6.73%) websites as a PBN site. As a sanity check, we manually inspected 10% of classified websites and found that all non-PBN sites were classified correctly. For the PBN sites, we found 6 (1.71%) misclassified instances. These were either legitimate websites that were themselves abused (e.g. through the commenting system), or strongly resemble the characteristics of a PBN site.

4.5 Limitations

Although our approach allows us to accurately discover new PBN sites without many false positive results, there are a few limitations that are inherent to our approach. First, our guided search requires an initial set of PBN sites. In our analysis we obtained these by purchasing backlinks from a variety of services for a total of $427 USD. Second, as a direct result of our guided search, new PBN services can only be detected when they provided backlinks for customers that also purchased the services of PBN providers that are being considered in the current iteration. As such, under the assumption that a customer of a high-end backlink provider is unlikely to have purchased services of a low-end provider, it is only possible to discover PBN services that are in a similar price range as the ones from the initial set. Finally, as a more general limitation of the ecosystem, it is prohibitively difficult to train a classifier that solely based on isolated features of a website is able to determine whether a site is associated with a private blog network. Consequently, we required a second step of our machine learning method that considered the features related to the linking behavior of a large number of sites. As such, our approach only allows to discover new PBN sites with only limited control on which ones will be detected. Nevertheless, as our approach can be applied iteratively, each iteration will yield a new set of PBN sites (which can then serve as the seed for the next iteration).

5 CHARACTERISTICS OF PBN ECOSYSTEM

Leveraging the 3,552 PBN sites we detected in the large-scale experiment by using our classifier, we present an in-depth analysis of the characteristics of these sites as well as information on the customers they serve.

5.1 PBN sites

5.1.1 Outgoing links. To analyze how links from PBN sites to their customers are created, we compute the total number of unique domains that are linked to by PBN and non-PBN sites. The cumulative distribution of this information, which is shown in Figure 5, clearly shows that PBN sites link to significantly more unique domains (median: 112) compared to non-PBN sites (median: 21).

To further evaluate how links are created, we performed an additional experiment: for a period of 43 days, we visited a subset of

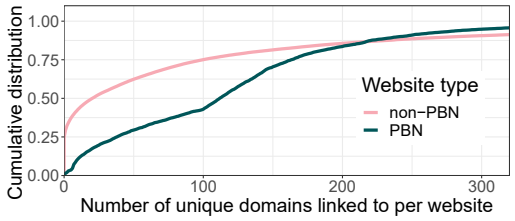


Figure 5: Cumulative distribution of unique number of domains linked to by PBN and non-PBN sites.

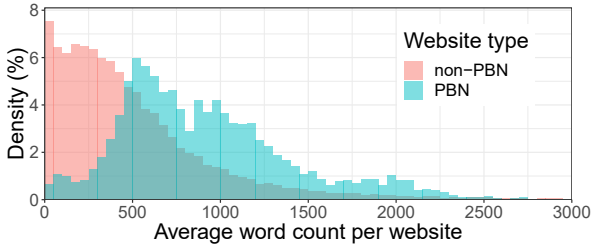


Figure 6: Average word count for PBN sites and websites unrelated to the ecosystem.

PBN sites on a daily basis and collected all outgoing links. We consider the domains that were encountered during the first five days as a baseline, and after this phase, every previously unseen domain is considered a new link. We found that the median number of links created per day over a 38 day period was 0.49, i.e. approximately one link every two days.

5.1.2 PBN site content. If links to PBN customers would be posted without an accompanying text or description, the PBN sites could be easily detected as such. Consequently, PBN sites typically create a blog post for each link that is created. Depending on the quality of the PBN, the content of the blog posts are either generated automatically by so-called text spinners, or are written manually by a content provider. An example of a PBN site with blog posts targeting health-related topics is `healthy-ch.org`, a screenshot of this website is shown in Figure 9 in the Appendix.

According to many online guides and tutorials on how to improve the SEO score of a website, the length of the posted article has a significant impact on the ranking. While there is no consensus on what is the ideal length of a page, most guides advise a length ranging from 500 to 1,000 or even 2,000 words. Interestingly, when analyzing the average word count across websites, as shown in Figure 6, PBN sites contain significantly more text (median: 833) than websites that we considered not to be PBN sites (median: 395).

5.1.3 Domain name registration. In order to analyze how backlink providers manage their domain names, we obtained historical WHOIS information on each of them through the SecurityTrails API[1]. As a comparable baseline, we also obtained historical WHOIS data from an equally sized set of websites that were classified as non-PBN. For each domain, we calculated the age based on the creation date of the domain since its last owner. Figure 7 shows the cumulative distribution for both PBN and non-PBN sites. From this

[1]https://api.securitytrails.com/

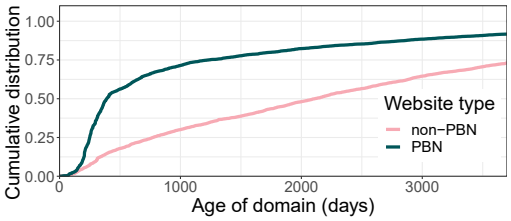


Figure 7: Cumulative distribution of the age of PBN sites and non-PBN sites expressed in days.

graph, it is clear that the distribution of non-PBN sites is roughly distributed evenly over time, whereas PBN sites have a much higher concentration of short-lived domains. There can be several possible explanations for this. First, the phenomenon of private blog networks is relatively new, as they are a response to the recently improved detection rates of other backlink abuse by popular search engines. Alternatively, the high number of new domains could be due to PBN sites being detected as backlink abuse by search engines, requiring them to cycle through new domains.

5.1.4 Leveraging residual trust. When a new domain is purchased, it generally takes some time before it is considered authoritative by search engines, a prerequisite to boost the reputation of other sites. However, to fast-track this, PBN providers may leverage residual trust from existing domains that have expired [10, 13]. More specifically, PBN providers can buy expired domains that had already gained the trust and reputation with search engines. Furthermore, it is likely that websites that previously linked to it do not remove the links, and thus keep boosting the domain's search engine ranking.

From the historical WHOIS data, we determined whether a domain was held by a different entity before it was registered to the PBN provider. We found that of all PBN sites, 78.31% were owned by a different party, whereas this is considerably less for non-PBN sites: 48.26%. It should be noted that these values are upper bound estimates, as a change to the registrant information in the WHOIS would also be considered a change of ownership. Nevertheless, these numbers indicate that PBN providers are taking advantage of the residual trust of expired domains to improve their blog network.

5.1.5 Website infrastructure. As backlink providers manage tens or hundreds of websites on which they regularly post new content, they are likely to automate or simplify this process. One of the ways they do this, is by leveraging content management systems: we found that 70.48% of the PBN sites are powered by WordPress, in contrast to 28.52% of the non-PBN websites.

5.2 PBN customers

In contrast to PBN sites, customers of backlink services do not exhibit any site-specific characteristics that can be used to detect them, as any type of website can purchase backlink services. However, it is still possible to leverage linking-based features that are inherent to the backlink ecosystem: when buying backlink services, the customer's domain will receive multiple links from various PBN sites. In our exploratory analysis, we found that the number of links on PBN sites that were provided ranged from 5 to 27. Following this information, we determine a website to be a customer of a PBN service if there are at least 5 incoming links from sites that our

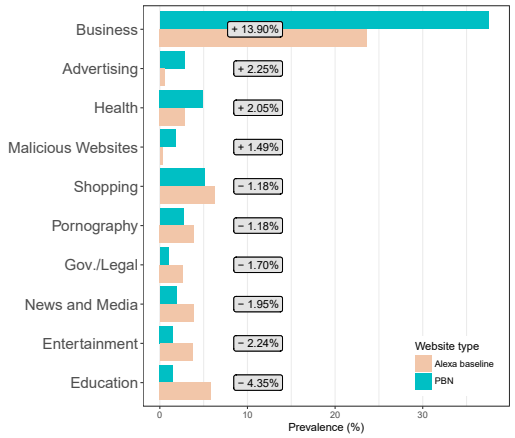


Figure 8: Categories of PBN customer sites which prevalence differs the most with the baseline.

classifier determined to be a PBN site. Out of the approximately 3 million unique domains that were linked to by the 52,777 analyzed websites, we found 12,848 PBN customers. It should be noted that this is a lower bound as our dataset only captures a subset of the entire backlink ecosystem, and therefore not all PBN sites that link to a single customer are present in the dataset.

We find that PBN customers receive on average 14.27 incoming links from actual PBN sites. Considering that we only analyzed a sample of all potential PBN domains, the average number of PBN domains per customer is higher than what we experienced in our exploratory experiment. A possible explanation for this is that customers place multiple orders to counter the seemingly temporary effect of purchased backlinks that we encountered for the high-end backlink services as part of our effectiveness analysis. Alternatively, website owners who purchase backlinks may try their chances with different providers.

Although most PBN customers have a number of incoming links from PBN sites that do not significantly deviate from the average, we find that there are a few outliers. For instance, one website is being linked to by 390 unique PBN domains. This website is `marketing1on1.com`, the primary site of the backlink provider we used to purchase a backlink package from (for Site #11, ref. Table 1). Most likely, this provider is using their own PBN to boost the visibility of their website. However, this also makes it possible to easily trace all the PBN domains back to this single provider.

Finally, to gather more insight in which website owners are mainly interested in purchasing backlinks for their domains and what their motivation is, we look at the categories of PBN customers. We use FortiGuard Labs' mechanism[2] to obtain the category for each PBN customer domain. As a baseline of websites that are not promoted by a PBN, we collect category information of a random sample of 50,000 websites from the Alexa top 1M list. Figure 8 shows the 10 categories that differ the most from the baseline. Interestingly, the *Business* category is the most popular, with a prevalence of 37.54% among PBN customers, and also differs the most from the baseline. Along with the *Advertising* category that showed to second most significant increase in prevalence with regard to the baseline, this confirms our intuition that buying backlinks is often financially motivated. Counterintuitively, the *Health* category is more prevalent with PBN customer sites. However, a manual analysis of a sample of these websites showed that the majority were either health-related businesses, such as a plastic surgery clinic, or were related to questionable practices such as online pharmacies.

Malicious Customers To get a better understanding of the malicious activities of PBN customers, we make use of Virus Total which aggregates many antivirus products and online scan engines. After querying all the 12,848 PBN customers and searching for malicious activities performed after the date we detected the customers, we found that 717 (5.58%) of the customer websites have been used for various malicious activities including phishing, distributing malware or serving other malicious contents such as scams, adult contents or offering illegal services. The majority of malicious customers abuse PBN SEO techniques to promote malware distribution websites (44.4%) and phishing website (19.6%).

We take a closer look at the date a customer purchases PBN backlinks to understand the dynamics of abusing this black-hat SEO technique. For this purpose, we run a separate experiment in which we monitor the detected PBNs for 35 days and record the backlinks created by these providers on a daily basis. In total, we find that 25,271 domains purchased backlinks in this period of which 1,081 (4.3%) customers started to abuse the purchased links in order to promote a variety of malicious contents. We observed that these websites wait for an average of 138 days after the creation of backlinks before they start serving malicious content.

6 UNCOVERING PBN NETWORKS

In the previous sections, we primarily focused on analyzing PBN sites and their customers in isolation. As part of our exploratory experiment, we found that PBN providers may have a substantial number of websites, up to several hundreds, allowing them to provide links originating from various websites. However, as we found in our large-scale analysis of PBN sites, administrators of private blog networks employ various tactics to evade detection, as this could render their entire network obsolete at once. Nevertheless, as the backlink providers may forget about certain aspects, or out of convenience or necessity create direct connections or similarities between their PBN sites, it is still possible to associate them to a single network. In this section, we discuss several techniques that can be used for this purpose.

6.1 Webserver location

One of the most straightforward ways to link together a network of PBN sites is through shared hosting. Of the 3,552 domains we classified as PBNs, we found a total of 11 different groups where at least ten PBN sites were hosted on a single IP address that was not considered to be associated with shared hosting, following the methodology proposed by Tajalizadehkhoob et al. [29, 30]. However, when analyzing these groups in more detail, we find that several originate from the same subnet, and the domain names share a common structure, e.g. for one group all domain names are in the `.biz` TLD. When we consider the `/24` subnet of the 11 different groups, we can join 3 groups, totaling to 8 distinct groups

[2] https://fortiguard.com/

of providers consisting of 87, 63, 41, 37, 24, 16, 15 and 10 PBN sites. It is important to note that the sizes of these providers is a lower bound, as a direct result of the sampling step of our guided search.

As backlink providers may be aware that hosting all their domains on a single IP address makes it straightforward for search engines to blacklist them, they may opt to leverage dedicated hosting. To evaluate to what extent this tactic is used and whether that indeed makes them more resilient to detection, we map each IP to its autonomous system and considered the ASes hosting the most PBNs. Through this, we find a single organization, SEO Ways, that hosts 103 PBN sites from our sample on 50 unique IP addresses covering several subnets.

6.2 WHOIS information

Another technique that can be used to link together domains belonging to the same provider, is by analyzing WHOIS information of each domain. In the previous section, we found that more than half of the PBN sites make use of WHOIS privacy protection, showing that many are well aware of this technique. Nevertheless, we found 15 private contact email addresses that were used for the registration of at least ten PBN sites. The groups of these networks show interesting information about how different the operations of PBN providers are. Whereas some providers try to diversify the domain names in their network e.g. by using several TLDs, other providers focus on a specific campaign. For instance, the largest group of PBN sites targets a specific niche, namely law practices. The domain names of all 23 PBN sites owned by this provider (the largest group we found with this technique) were composed of a specific law term such as "DUI", "child custody" or "misdemeanor" and the term `lawyergo`.

In order to get a more accurate estimation of the size of PBN operations, we performed a reverse lookup to obtain all domains registered by all the email addresses that were used to register PBN sites. Through this technique, we find 2,629 new domains that are highly likely to be involved in the PBN ecosystem. Furthermore, we find that ten email addresses registered more than 100 domain names; the top five registered 475, 466, 212, 179 and 160 domains. This gives a clear indication that a single private blog network may consist of several hundred websites.

6.3 Shared customers

Whereas the previous techniques leverage information from the web infrastructure of PBNs, in this section we introduce a general technique that exploits the intrinsic linking behavior of PBN sites within their network. More concretely, when a customer purchases a backlink package, the provider will create multiple links from several websites out of her network. Consequently, PBN sites from the same network are more likely to link to the same customers. To leverage this behavior, we first create a graph where the nodes represent PBN sites, and edges are created when two PBN sites link to the same target website. We excluded links to websites that are in the list of 1 million most popular sites according to Alexa, as these are generally more likely to be linked to.

The weight of each edge is set to the total number of customers the two PBN sites have in common. Next, to create clusters of PBN sites, we remove edges that have a weight lower than a certain threshold and determine the connected components of the graph. Of course, the composition of these clusters is directly related to the imposed threshold. To define this value, we leverage the PBN networks we uncovered in the previous sections. More concretely, we define a cost function that reflects how well the previously discovered networks are represented by the connected components for a specific threshold value. By iterating over all possible thresholds in the range of 1 to 50, we find that the optimal value is 12, i.e. two PBN sites are connected if they share at least 12 customers.

In total, we find 108 connected components representing 63.71% of all detected PBNs. The remaining PBN sites did not share any customers with other PBN sites, which could be either due to the sampling process, or the limited number of web pages we visited on each site. Of the 108 connected components, most are relatively small; only 20 consist of more than ten PBN sites. Nevertheless, this evaluation does reflect the magnitude of certain private blog networks; we find that the five largest connected components consist of 540, 460, 330, 152 and 80 websites. These numbers should be considered rough estimations: on the one hand, we may have missed sites from a network because of the sampling process, on the other hand, if there are a sufficient number of customers who purchased backlinks from multiple providers, the networks of the providers would be joined. We find that the cluster of 330 PBN sites is related to Marketing1on1, as all domains in this cluster link to it, supposedly to promote the website of this PBN provider. In total, we found 390 PBN sites that link to Marketing1on1, indicating that the reported network sizes are likely to be an underestimation.

6.4 PBN Revenue Estimation

As a result of our analysis, we find two PBN networks that can be linked back to a specific provider: i) our sample of analyzed websites contains 103 PBN sites hosted on an autonomous system owned by SEO Ways, ii) we find that 390 PBN sites link to a single provider (Marketing1on1). For the other detected networks, we were unable to link these back to a specific operation, and therefore could not include these in our revenue estimation. In this section, we leverage this information to create an estimation of the revenue made by the PBN providers.

To compute the monthly revenue of the PBN providers, we crawl their PBN sites (up to 200 visits per website) on a daily basis during 35 days and record all outgoing links that have been observed. Next, we consider the first 5 days as the baseline period, and only take the new links discovered during the following 30 days into account. For SEO Ways, we find that during one month, there were 796 new, unique customers being linked to by a total of 50 websites. Following the prices reported on their website[3], they offer packages of $150, $289 and $559, depending on the number of posts that are created. As a lower-bound estimation, we consider that customers only purchase the cheapest package and find that this PBN provider grosses roughly $110,000 per month. The operating costs of SEO Ways are mainly determined by the cost of their server infrastructure: we found 50 IP addresses under their control that hosted a website (using a reverse DNS database, we found 198 domain names pointing to one of their IP addresses). When we consider the price of a dedicated server to be $100/server/month,

[3] http://www.seoways.com/pbn.html

and a domain name and IP address $5/month, we estimate their operating costs to be roughly $6,000 per month.

Using the same method, we observed that in one month 3,520 links to new customers of Marketing1on1 were created, originating from 266 PBN sites. As listed on their website, Marketing1on1 provides four backlink packages ranging from $49 to $269. Again considering only the cheapest package, we find a lower-bound revenue estimate of $172,480 per month. In contrast to SEO Ways, Marketing1on1 mainly leverages shared hosting providers for its hosting needs, which are considerably cheaper. To compute the operating costs, we first need to determine the total number of PBN sites hosted by Marketing1on1 (because of the sampling step in our large-scale analysis, not all domains are included in our dataset). For this, we first find all backlinks to `marketing1on1.com`, using the service provided by Majestic. Next, we visit all pages that link to the PBN provider's website and determine whether the link is still present, and count the number of other domains that are linked to. We only consider a domain to be part of the Marketing1on1 network when there are less than 50 unique domains linked on that site. The latter step is important because a manual inspection of sample of the backlink pages showed several instances where the `marketing1on1.com` domain was used in comment abuse. In total, we find 1,544 domains that likely belong to the Marketing1on1 network. Note that because for every package at least 10 links created throughout the network, it is unlikely that this would significantly affect the revenue estimation. Considering an operating cost of $20/month/domain (shared hosting options can be found for $5/month, but we find that the PBN provider tries to diversify its hosting, so not only the cheaper options are available), we find an operating cost of $30,880, and thus a monthly profit of approximately $140,000.

Interestingly, the profits generated by a single PBN are in the same range as that of trending keywords abuse, which boosts the ranking of Made-for-Adsense (MFA) websites and sites promoting fake anti-virus software [16]. However, in contrast to most other black-hat SEO techniques, the revenue made by PBN operators largely originates from their paying customers. Consequently, the findings of Wang et al., namely that undermining the malicious monetization techniques are a potent response against black-hat SEO campaigns [31], are no longer applicable to PBNs.

7 RELATED WORK

7.1 Black-hat SEO techniques

There exists a wide variety of black-hat SEO techniques that aim to deceive search engines such that the promoted website is ranked higher in the search results. The vast majority of the black-hat SEO techniques that have been studied to date try to manipulate a specific aspect of the ranking algorithm. For example, the contents of web pages may be specifically crafted such that it includes many specific keywords for popular searches (a technique known as *keyword-stuffing*) [19]. Because these techniques exhibit behavior different from typical web pages, they are relatively quickly detected by search engines and miscreants have to resort to new techniques. For instance, in 2011, John et al. report on a technique where trending keywords were abused to lure users to malicious web page [8]. As a part of a longitudinal study by Moore et al., the researchers found that an update to Google's ranking algorithm rendered this SEO technique largely ineffective [16]. Later, in 2016, Liao et al. report on a new keyword-based black-hat SEO technique, namely including long-tail keywords on pages hosted by cloud web hosting services such as Amazon S3 or Google Drive [14]. This technique abuses the intuition that a search engine will give preference to results that contain a more specific search keyword. Du et al. explore another novel SEO technique, where wildcard DNS entries are abused to create virtually infinite sites in order to deceive search engines [3]. The novel black-hat SEO technique we report on in this paper, private blog networks, differs from the prior findings in the sense that it does not try to manipulate specific intricacies of the ranking algorithm. Instead, it tries to manipulate the search results by leveraging the way the ranking algorithm is intended to function, i.e. favoring websites that have incoming links from authoritative sources (these are the websites which are artificially created and maintained by the PBN operator).

Another class of nefarious SEO techniques try to exploit existing websites, which are already considered reputable by search engines. Examples of these types of abuse include comment spam, where miscreants post comments to blog posts with a link to their own website [34], profile abuse, where spun content is added to the biography section of an automatically registered user account [35]. In some instances, adversaries go as far as to compromise existing websites and place SEO-boosting on them [12, 31]. In contrast to these techniques, we find that PBN operators orchestrate their own websites, and do not directly exploit or interfere with other sites.

7.2 Detection of SEO abuse

In 2000, Davison was the first to introduce a method, based on the C4.5 decision tree algorithm, to detect *nepotistic links*, i.e., links purposefully created to mislead search engines [2]. In the following years, many researchers have reported on "spam links," and techniques to detect them. For instance, in 2004, Gyöngyi et al. report on *web spamming*, a technique in which malefactors create thousands of links containing a large number of related keywords in order to rank higher in search engines [6]. In a follow-up work, they explore the effect of various types of web spamming on the Page Rank algorithm, which was at that time used by Google [5]. Similarly, in 2005, Wu and Davison introduce a technique to detect *link farms*, a network of densely connected websites that aim to promote a target website [33]. Among other things, they find that 9% of search queries have at least one spam page in the top 10 results, showing that at that time link farms were still highly effective.

Over the following three years, various new detection techniques have been proposed [1, 4, 36]. These methods leverage topological relationships, aim to detect suspicious nodes and use that to grow their seed dataset, use machine-learning, or are based on graph regularization techniques. For an overview, we refer the reader to the survey by Spirin and Han [27]. It is important to note that prior work has mainly focused on detecting link spam, which typically creates an abundance of interconnected links between websites. Contrastingly, PBNs try to resemble regular websites and are careful to only create links that appear legitimate. As such, web spam and PBNs exhibit very different properties and thus the detection methods are not directly interchangeable.

8 CONCLUSION

By ranking higher in search results, website owners can gain a significant advantage over their competitors. Next to following the standard guidelines to optimize a website for search engines, certain website administrators may resort to deceptive SEO methods, for instance by purchasing artificially created backlinks. Through an exploratory experiment where we ordered several of these backlink packages, and find that many of these backlinks are created through so-called private blog networks. In this paper, we present a comprehensive and in-depth analysis of the different entities involved in the PBN ecosystem.

We propose a novel methodology to detect PBN sites based on a guided search to identify websites that are more likely to be involved in a PBN, and a multi-step classifier that leverages both content-based and linking-based features inherent to the PBN ecosystem. Starting from a limited seed set of PBN sites, we perform a large-scale experiment on 52,777 websites and manage to detect 3,552 PBN sites. Our analysis of these PBN sites shows that backlink providers employ a wide variety of measures to avoid detection, which is highly indicative of the ongoing arms race between search engines and PBN owners. We conjecture that improving the mechanisms used to detect PBNs, e.g. through our proposed methodology, can drive up the operational costs of these services and thus reduce the financial motives to operate PBNs. Finally, despite the various measures taken by backlink providers to prevent all websites from their network to be associated, we find that several techniques are still effective at detecting clusters. Leveraging these methods, we find that a single PBN network may consist of several hundred websites that link to thousands of customers, allowing the operators to generate a revenue of more than $100,000 per month.

Acknowledgments: We thank the anonymous reviewers for their helpful feedback. For Stony Brook, this work was supported by the Office of Naval Research (ONR) under grant N00014-16-1-2264 and by the National Science Foundation (NSF) under grants CNS-1813974, CMMI-1842020, CNS-1617593, and CNS-1617902.

REFERENCES

[1] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. 2007. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 423–430.

[2] Brian D Davison. 2000. Recognizing nepotistic links on the web. *Artificial Intelligence for Web Search* (2000), 23–28.

[3] Kun Du, Hao Yang, Zhou Li, Hai-Xin Duan, and Kehuan Zhang. 2016. The Ever-Changing Labyrinth: A Large-Scale Analysis of Wildcard DNS Powered Blackhat SEO.. In *USENIX Security Symposium*. 245–262.

[4] Qingqing Gan and Torsten Suel. 2007. Improving web spam classifiers using link structure. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. ACM, 17–20.

[5] Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Link spam alliances. In *Proceedings of the 31st international conference on Very large data bases*. 517–528.

[6] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 576–587.

[7] internet live stats. 2017. 60,180 Google searches in 1 second. http://www.internetlivestats.com/one-second/.

[8] John P John, Fang Yu, Yinglian Xie, Arvind Krishnamurthy, and Martín Abadi. 2011. deSEO: Combating Search-Result Poisoning.. In *USENIX Security*.

[9] Konker 2018. Konker - A Freelance Market For Those Who Bring Home The Bacon. http://www.konker.io/.

[10] Tobias Lauinger, Abdelberi Chaabane, Ahmet Salih Buyukkayhan, Kaan Onarlioglu, and William Robertson. 2017. Game of Registrars: An Empirical Analysis of Post-Expiration Domain Name Takeovers. In *26th USENIX Security Symposium (USENIX Security 17)*. USENIX, 865–880.

[11] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*.

[12] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. 2014. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 930–941.

[13] Chaz Lever, Robert Walls, Yacin Nadji, David Dagon, Patrick McDaniel, and Manos Antonakakis. 2016. Domain-Z: 28 registrations later measuring the exploitation of residual trust in domains. In *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 691–706.

[14] Xiaojing Liao, Chang Liu, Damon McCoy, Elaine Shi, Shuang Hao, and Raheem Beyah. 2016. Characterizing long-tail SEO spam on cloud web hosting services. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 321–332.

[15] Majestic. 2017. Glossary. https://majestic.com/support/glossary.

[16] Tyler Moore, Nektarios Leontiadis, and Nicolas Christin. 2011. Fashion crimes: trending-term exploitation on the web. In *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 455–466.

[17] Moz. 2015. Search Engine Ranking Factors. https://moz.com/search-ranking-factors.

[18] Moz. 2017. Domain Authority. https://moz.com/learn/seo/domain-authority.

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

[20] Advanced Web Ranking. 2017. CTR study. https://www.advancedwebranking.com/cloud/ctrstudy/.

[21] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D Strowes, and Narseo Vallina-Rodriguez. 2018. A long way to the top: significance, structure, and stability of internet top lists. In *Proceedings of the Internet Measurement Conference 2018*. ACM, 478–493.

[22] Barry Schwartz. 2013. Google Busts Yet Another Link Network: Anglo Rank. https://searchengineland.com/google-busts-yet-another-link-network-anglo-rank-179296.

[23] Barry Schwartz. 2015. Google Panda 4.2 Is Here; Slowly Rolling Out After Waiting Almost 10 Months. https://searchengineland.com/google-panda-4-2-is-here-slowly-rolling-out-after-waiting-almost-10-months-225850.

[24] SEOClerks 2018. SEO Marketplace - SEOClerks. https://www.seoclerk.com/.

[25] Usman Shahid, Zubair Shafiq, Shehroze Farooqi, Padmini Srinivasan, Raza Ahmad, and Fareed Zaffar. 2017. Accurate Detection of Automatically Spun Content via Stylometric Analysis. (2017).

[26] Amit Singhal. 2011. Finding more high-quality sites in search. https://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html.

[27] Nikita Spirin and Jiawei Han. 2012. Survey on web spam detection: principles and algorithms. *ACM SIGKDD Explorations Newsletter* 13, 2 (2012), 50–64.

[28] Danny Sullivan. 2016. RIP Google PageRank score: A retrospective on how it ruined the web. https://searchengineland.com/rip-google-pagerank-retrospective-244286.

[29] Samaneh Tajalizadehkhoob, Maciej Korczyński, Arman Noroozian, Carlos Gañán, and Michel van Eeten. 2016. Apples, oranges and hosting providers: Heterogeneity and security in the hosting market. In *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*. IEEE, 289–297.

[30] Samaneh Tajalizadehkhoob, Tom van Goethem, Maciej Korczyński, Arman Noroozian, Rainer Böhme, Tyler Moore, Wouter Joosen, and Michel van Eeten. 2017. Herding vulnerable cats: a statistical approach to disentangle joint responsibility for web security in shared hosting. *arXiv preprint arXiv:1708.06693* (2017).

[31] David Y Wang, Stefan Savage, and Geoffrey M Voelker. 2013. Juice: A Longitudinal Study of an SEO Botnet.. In *NDSS*.

[32] WordStream. 2017. How Does Google Make Its Money: The 20 Most Expensive Keywords in Google AdWords. http://www.wordstream.com/articles/most-expensive-keywords.

[33] Baoning Wu and Brian D Davison. 2005. Identifying link farm spam pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web*. ACM, 820–829.

[34] Jialong Zhang and Guofei Gu. 2013. NEIGHBORWATCHER: A Content-Agnostic Comment Spam Inference System.. In *NDSS*. Citeseer.

[35] Qing Zhang, David Y Wang, and Geoffrey M Voelker. 2014. DSpin: Detecting Automatically Spun Content on the Web.. In *NDSS*.

[36] Dengyong Zhou, Christopher JC Burges, and Tao Tao. 2007. Transductive link spam detection. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. ACM, 21–28.

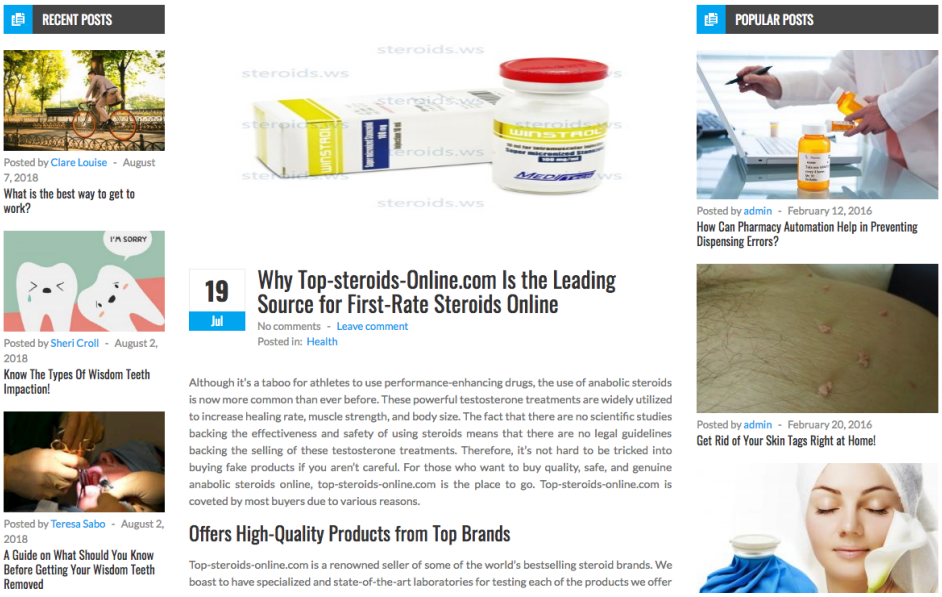


Figure 9: Screenshot of a PBN site targeting health-related topics.