Resource Networks of Pet Scam Websites

Benjamin Price
Department of Computer Science
University of Bristol
Bristol, UK
bp17492@bristol.ac.uk

Matthew Edwards
Bristol Cyber Security Group
University of Bristol
Bristol, UK
matthew.john.edwards@bristol.ac.uk

***Abstract*—The pet scam is a form of online fraud in which scammers leverage victims' emotional attachment to fictitious pets as a means for extorting money. Both fraudulent pet seller sites and fraudulent delivery sites are involved in the scam. When sites of either kind are taken down, scammers create new sites, often reusing effective content from previous scams.**

We explore connections within the largest current collection of pet scam websites, examining four distinct types of resource sharing that are indicative of shared authorship. We find that 90% of all accessible sites share at least one form of connection to another known site, including many identifiable links between seller and delivery sites, and that some scam authors could be behind hundreds of individual scam websites. We partially validate our linkage methods using domain registration data, and discuss the implications of using different connection types to analyse online fraud more generally.

***Index Terms*—online fraud, pet scam, clustering, link analysis**

I. Introduction

A pet scam website is a fraudulent website that claims to sell pets. Scammers will create a website that appears to be a legitimate seller of pets, and advertise through social media and traditional advertising platforms. Scammers will attract potential victims by advertising pets for far less than the market price. Their aim is to direct potential victims to their website and to get them emotionally invested in a fictitious pet. These fraudulent websites often appear to be legitimate at first glance. Many claim to be associated with real organisations such as the International Pet and Animal Transportation Association (IPATA), and some will have testimonials from what appear to be previous customers. The website will showcase the pets for sale and present contact information that allows victims to message the scammers so they can purchase the pet. If the victim chooses to purchase the fictitious pet, the scammer will only accept non-refundable payment methods such as Western Union and MoneyGram, which makes it difficult for victims to recover their lost money. However, once the victim has paid, the scam is not over. Pet scammers deploy a number of ploys to further extort money from their victims who are now emotionally and financially invested in a fictitious pet.

Alongside the pet advertisement website, scammers will create fraudulent pet delivery websites, and once a victim has paid for the pet, he or she will be given a fake tracking number and the URL for the fake delivery website. Here, the victim can track the status of the delivery of their pet. Shortly after the purchase, troubles with shipping will arise that can only be resolved by the victim paying the scammer more money. These include logistical and medical issues such as a pet being stuck in customs, or needing emergency veterinary care. The sunk-cost fallacy, along with an emotional message to the victims explaining how their poor pet is stuck somewhere or is ill, persuades the victim to pay. Additional fees can also be created after the initial purchase, such as fees for vaccinations or a ventilated cage. Much like in advance-fee fraud, the scammer will continue to invent new fees and hurdles as the transaction drags on. If the victim becomes apprehensive about paying more money, then the scammer can threaten to get law enforcement involved. This can frighten vulnerable people into cooperating. The scam ends when the victim either runs out of money, or realises that they have been scammed. At this point, the separation between the pet advertisement site and the delivery site means that pet scam websites attracting victims can sometimes pretend to not be responsible for the delivery website or involved in the shipping company's malfeasance, and then the delivery website can be reported or taken down, whilst the pet scam website continues operating.

While some pet delivery websites represent fictional companies, others take advantage of well-known brands and either pretend to be associated with real companies, or pretend to be the companies themselves. Since these websites are similar to, or exact copies of, legitimate transportation companies' websites, it can be difficult for victims to realise that they are fraudulent. In 2017, Delta Airlines filed a federal lawsuit in the USA against a number of fraudulent delivery websites associated with pet scams, including DeltaPetTransit.com and DeltaPetAirways.com, for breaching their trademark [1]. These sites were designed to look similar to the legitimate Delta Airlines website and even used their trademarked logo in order to trick victims into thinking that they were paying Delta instead of the scammers.

Pet scam websites are often targeted at particular breeds, to give the impression that they represent a legitimate breeder in a particular niche. At the same time, the websites are mass-produced in order to target as many types of pets as possible. Sites are taken offline by authorities once victims report them, only to be re-hosted under a new domain name. Online tools designed to make websites quickly and cheaply make this process even easier for the scammers. Many will also use

services such as WhoisGuard™ by Namecheap Inc. in order to protect the identities used in domain registration details, which makes it difficult for law enforcement to identify the perpetrator(s).

The number of complaints related to pet scams received by consumer protection organisations such as the Better Business Bureau (BBB) has been increasing every year. In the three year period from 2017 to 2019, pet fraud complaints to the BBB increased by 39% from 4,664 to 6,466 a year. Victims usually lost between $100 and $1,000, although some lost as much as $5,000. The majority of victims are from the USA and are in their 20s and 30s [2].

There have been efforts by organisations to keep track of the names of people, websites and emails involved in pet scams, so that potential victims can be warned. PetScams.com is a website run by volunteers that is dedicated to maintaining and hosting the largest public list of pet scam websites. Users are able to report websites via an online form. Volunteers working for PetScams.com will review submissions to decide whether or not the complaint is legitimate. If enough legitimate complaints are made, the domain is added to the appropriate list of scam websites. They maintain two lists of domains: one for those fraudulently advertising and claiming to sell pets, which we shall refer to as *pet scam websites*, and one for those fraudulently claiming to deliver pets which we shall refer to as *delivery scam websites*. This paper uses both lists from PetScams.com as a source of known pet scam and delivery scam websites.

Pet scam websites are usually constructed as cheaply as possible to minimise operating costs, so operators often duplicate resources from previous instances in a similar or related campaign. For example, some of the testimonials on different websites are almost exact copies of each other, with the only differences being the name of the pet and website. Many websites also reuse identical images of pets under different names. These similarities suggest that multiple websites are made by the same person or group of people. Scammers who find successful techniques and methods for scamming people will want to reuse them on their next website. This suggests that clustering pet scam sites into connected campaigns based on shared resources is viable. This would also serve to identify the most prolific scammers, prioritising them as targets for law enforcement action.

In this paper, we explore the links suggested by the different resources reused between pet and delivery scam websites, drawing on the largest and most up-to-date collection of known sites to identify connected campaigns and investigate which shared resources are most suggestive of common authorship. In particular, we investigate:

1) *How is agglomerative clustering of pet scam websites into connected campaigns affected by the type of resource used to 'link' sites?*
2) *Do different shared resources confirm or complement each other in establishing links between sites?*
3) *Which shared resource links are most likely to be validated when referring to domain registration details?*

We begin with a brief survey of related work. In Section III we describe our data collection and some features of the resulting pet fraud website corpus. Following that, Section IV discusses four means of identifying shared resources between scam websites, and outlines our validation strategy. Section V presents results for the main aims of our investigation. We conclude with a discussion of our findings, their limitations, and implications for future work both on pet scams and online fraud more generally.

II. Related Work

Pet scams specifically have not been extensively covered in previous work. The only prior art we are aware of is work by Norazman & Zamin [3], who report on their efforts refining email filters specifically for pet scams, and present some details about the operation of pet scams, drawn from support forums and victim interviews. Although this form of fraud in particular has not been well covered, the pet scam stands as an example of a common pattern of internet-enabled fraud wherein the victim is attracted via an online advertisement, and then further groomed into payment in private correspondence. Other examples include rental scams [4], dating fraud [5], cryptocurrency trust trading scams [6], high-yield investment programmes [7] and technical support scams [8].

Connections between online malicious actors have been used in a variety of contexts as an aid to a study of those actors, and particularly how they make use of the web. For example, work targeting extremist organisations in the US [9] and internationally [10] has made use of link analysis to identify unknown groups and forums and situate them within networks of interest. These earlier works focused on direct links to different websites, and particularly violent extremist sites. However, connections in the form of re-used images [5], common text [5], [7], analytics identifiers [6], [11], domain registration details [6], [8] and even replicated webpage structure [7] have been observed between instances in a variety of crimes, including many related to online fraud.

Drew & Moore [7] identified replicated criminal websites related to particular frauds using text and webpage structure features. Their approach exploits criminals' need to re-use material in order to keep the setup costs for their fraud low. They found that different fraud types exhibited different replication behaviours, with escrow-fraud websites producing two large clusters, while high-yield investment programs were more diverse and disconnected. Our analysis explores similar questions for pet scam websites, along with a broader discussion of the relative usefulness of different materials for identifying connected scam sites.

Clustering of scam instances by their shared resources can have several applications. Edwards et al. [5] made use of connections in resources shared between fraudulent dating profiles, such as images and text, to identify the geographic origins of scam profiles which used proxies to disguise their connection. Leontiadis et al. [12] clustered unlicensed pharmacies using their inventories, identifying that a large number of such online pharmacies relied on a small number of suppliers

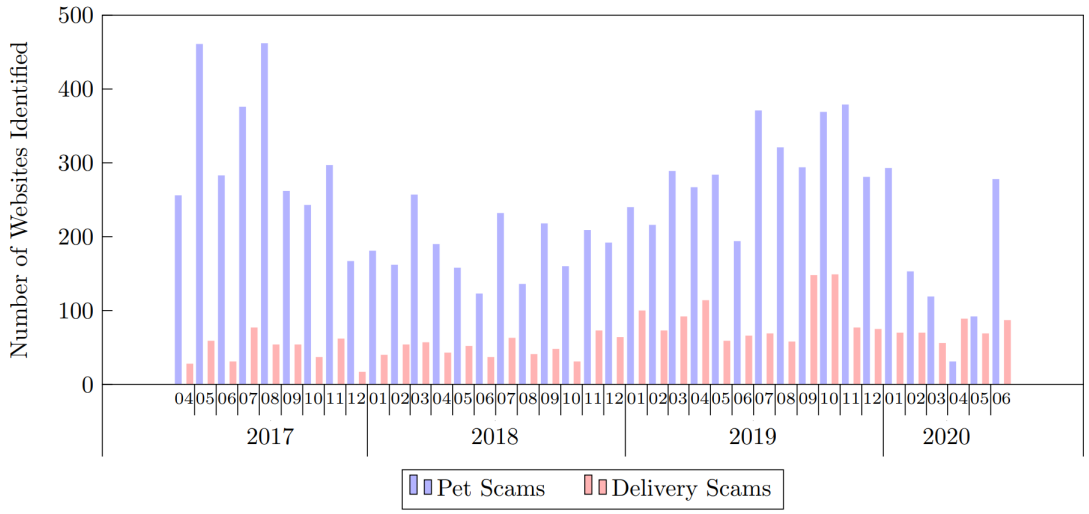


Fig. 1: Number of pet scam and delivery scam websites identified each month from April 2017 to June 2020.

– pointing to an area of effective action for law enforcement intervention. Phillips & Wilder [6] cluster advance-fee fraud sites and connected Bitcoin addresses to understand the typology of scams and the degree to which different entities are operating connected campaigns. In general, identifying connected clusters of fraud has pointed to opportunities for more effective interventions, and enabled the targeting of limited resources for enforcement.

III. Data Description

Our dataset of pet scam and delivery scam websites was obtained from the complete listing hosted by PetScams.com as of the end of June 2020. Since 2017, 12,050 scam websites have been identified, at an average rate of 309 new domains identified per month. Figure 1 shows how the number of pet scam websites identified varies over time. The period between February and May 2020 shows a notable decrease in the number of pet scam sites identified, while figures for June 2020 seem to demonstrate a return to pre-pandemic levels. While the period strongly suggests a relationship to the COVID-19 pandemic, the causality is unclear – we do not know if volunteers verifying reports were distracted by pandemic-related issues, if a drop in pet purchasing behaviour led to a decrease in user reports, if the criminals themselves were less active as a consequence of global disruption, or if some combination of these or other explanations is to blame. We note, however, that it appears to be pet scam websites specifically that are affected during the pandemic period, and the rate of identification for delivery scam websites does not appear to have been affected. This, together with other organisations reporting a *rise* in pet scamming activity during the pandemic period, connected to increased pet-buying behaviour [13], [14], suggests that the effect may be through impact on volunteer activity in verifying site reports.

Overall, there are fewer delivery scam websites (2,551) than pet scam websites (9,499). This suggests that not all pet scam websites have a unique corresponding delivery scam website. It is not always possible to determine which delivery scam websites are associated with which pet scam websites until a victim has paid money and is told who will be shipping the pet. In our later analyses, we discuss means by which delivery scam sites can be directly connected to pet scam sites as a result of shared resources.

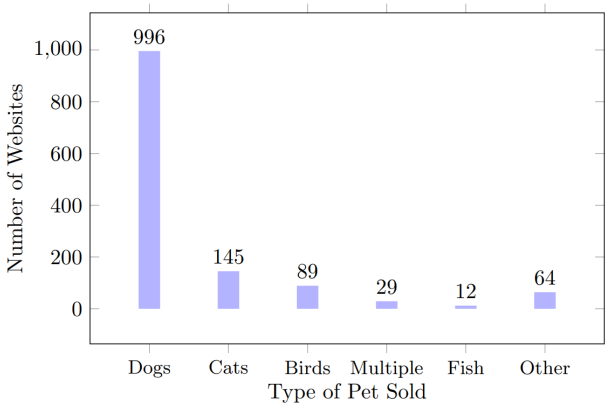


Fig. 2: Distribution of pets sold on 1,335 websites.

We crawled all 12,050 known domains from both of the PetScams.com lists and downloaded the 1,780 websites that were still online, including over 70,000 images. Of these, 1,335 were pet scam websites and the remaining 445 were delivery scam websites. Figure 2 shows that out of the 1,335 pet scam websites that were downloaded, the vast majority exclusively sold dogs. Cats and birds were the next most popular pets to be sold, with only a small number of websites selling multiple types of pets. Pet scam sites tend to target specific breeds. Previous analyses carried out by PetScams.com, and confirmed through our observation, suggest that the most targeted dog breeds are French Bulldogs and Yorkshire Terriers.

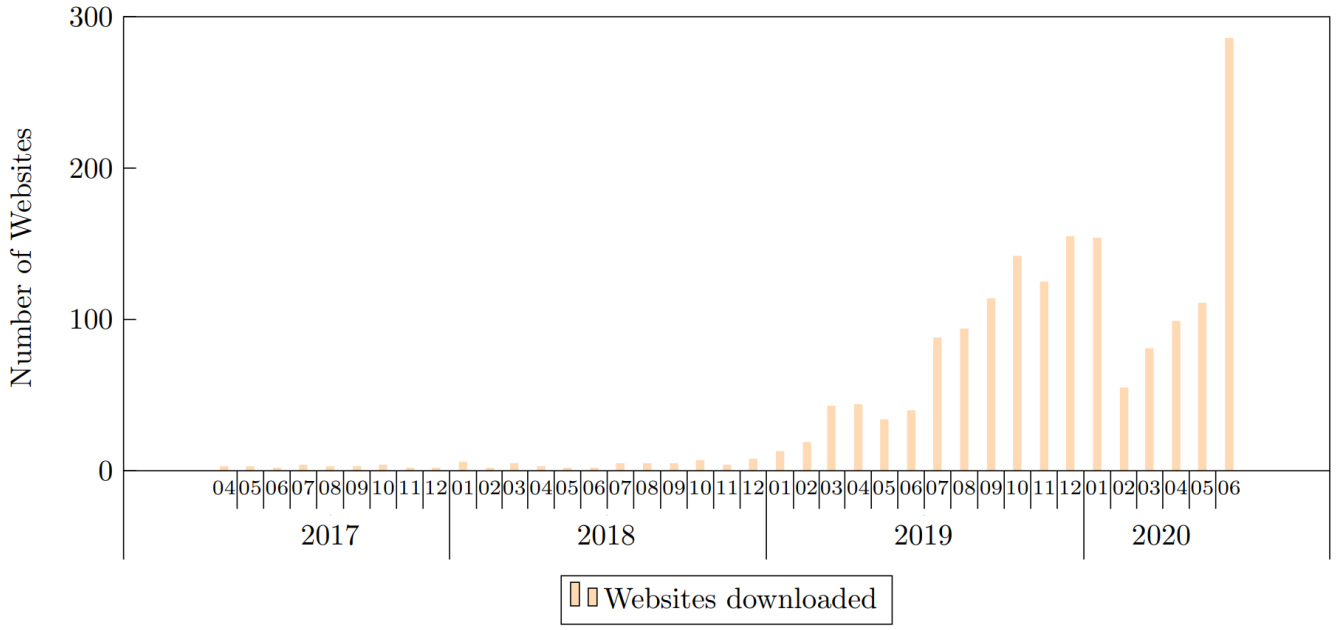


Fig. 3: Sites captured in our sample by the date they were first identified.

Figure 3 shows the distribution of our sample over time. While a recency bias is to be expected given the nature of the PetScams.com reporting efforts, significant numbers of scam sites first identified more than a year ago were still online at the time of our collection, a result which appears to show surprisingly slow responses from hosting and domain providers. Even more interesting are the small number of highly persistent scam sites first identified as far back as 2017, and still (or once again) online. For example, shihtzupuppysforsale.com, first identified in October 2017, is still online at the time of writing. This site, along with some 89 other sites in our sample, is hosted on a Google Cloud IP address associated with a range of domain registrars. Manual investigation turned up no clear traits in common between the scam domains still online as of our collection date–no registrars seemed particularly more or less likely to still be hosting scam sites three years after reporting began.

Following crawling, we collected public domain registration data for each site scraped. While 66 different registrars were observed, there was significant clustering around a few dominant services, with 64% of sites hosted by the top 5 most popular registrars. Figure 4 shows the breakdown of sites per registrar. Namecheap Inc. was the most popular registrar, accounting for 652 different scam sites (37%), and NameSilo LLC. was the second most popular, with 250 domains (14%). Both of these companies offer services which hide the address and contact details given by the person who registered the website. Many details within the WHOIS response data are therefore hidden, which frustrates the use of domain registration details themselves as a means of clustering fraud campaigns. In our results, we discuss which content-based features correlate best with links verifiable in domain registration data, suggesting potential workarounds for this issue.

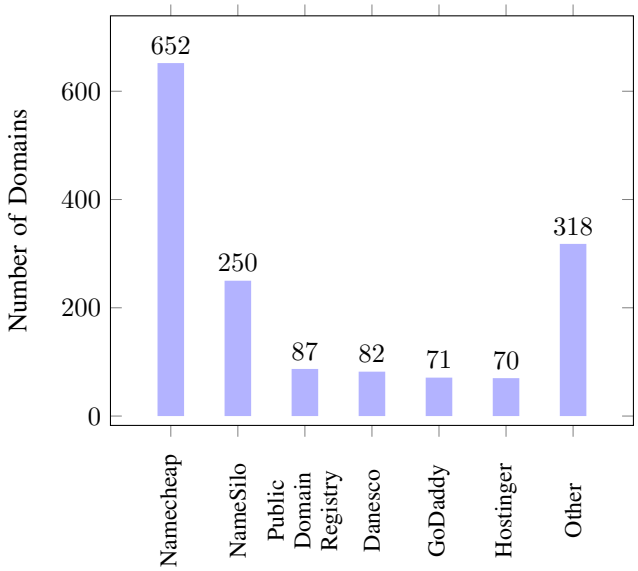


Fig. 4: Number of websites registered with the most popular registrars.

The location of pet scam sites is an important element of the fraud script used. A pet does not need to be delivered if a customer can come and collect the pet themselves, so scammers first enquire about the victim's location before revealing themselves to be a prohibitive distance away, justifying the use of the delivery scam website. We used a geolocation service[1]

[1]http://freegeoip.app

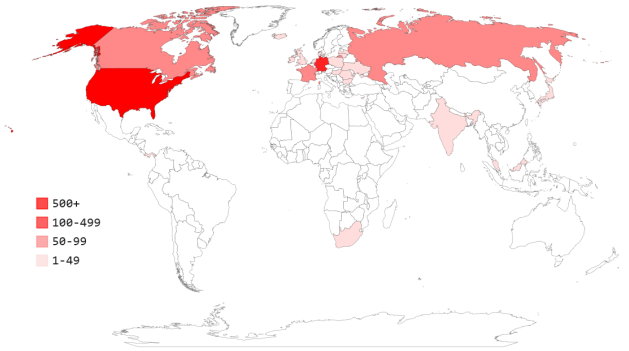


Fig. 5: Number of pet scam IP addresses in each country.

to identify the national origins of IP addresses belonging to pet scam websites. Figure 5 shows them plotted on a world map. The vast majority of domains were hosted in the USA, with other common origins including Germany, Russia, France and Canada. The evidence this carries about the true national origins of scammers is quite limited – many websites are cloud-hosted, so the location of the server may have little bearing on the origin of the scam. Nonetheless, the choice of many servers in Germany and Russia for English-language websites is interesting, and may reflect a search for cheaper or less-well-regulated hosting providers.

IV. Method

We examine four different methods for identifying connections between pet scam sites based on their content, detailed below. We then partially validate these connections using the subset of sites for which domain registration details are accessible.

A. Direct Links

The first type of connection we looked for was direct links – URLs pointing from one domain to another anywhere in the source of a site. If one scam website has a direct reference to another scam website's domain, this is a strong indicator of collaboration, and both websites may even have been created by the same person(s). While such links are directional, from a source to a target, the other forms of connection we examine are not, so we treat all connections as undirected.

B. Shared Images

The second connection type we examined was images held in common between sites. These are most typically images of pets, often collected by scammers from social media or other public sources, and reused for their appeal value. As our dataset included some 72,288 images, we required a fast and robust method of comparison. We also wished to compare the perceived visual match between images rather than use an exact cryptographic hash such as SHA or MD5, to avoid being misled by minor compression artefacts or small alterations introduced through editing. To meet these requirements, we used a perceptual hashing algorithm.

Perceptual hash functions differ from cryptographic hash functions in that they are designed such that two similar bit strings will result in a similar hash value. The greater the difference in bit strings, the greater the difference in the hash values. In particular, we chose a perceptual hashing algorithm using the discrete cosine transform (DCT) method. There is a tradeoff in perceptual hashing centred on the length of the hash chosen. Smaller hashes are quicker to compute, but less accurate, as each bit of the hash reflects an average across a broader section of the image. Following experimentation with different hash-lengths, we found 64 bits of hash to be optimal.

Each image had its perceptual hash computed, and images with the same perceptual hash were considered to be matched, indicative of a connection between sites. When generating the resource network for shared images, we formed an edge between two nodes if they shared at least two images. Increasing the threshold to a higher number of shared images provides more evidence that the websites were created by the same person(s), but results in fewer connections.

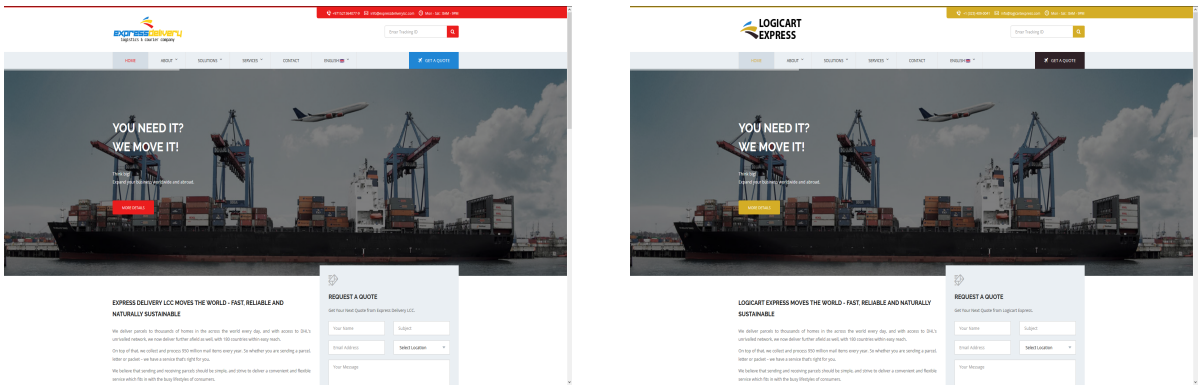
Some of the most common hashes were from images that would not be useful, such as completely black images, or images of logos of popular companies that deal with transactions such as PayPal and Western Union. We wanted to exclude these types of images from consideration since payment processor logos, similar icons and background colour blocks are too common to provide a good indication of affiliation or shared authorship. In order to exclude these unwanted images, we created a script that iterated over the most common image hashes and displayed several images with each hash. We then manually checked if the images were valid or invalid. The validity checker script also marked all images with a height or width less than 64 pixels as invalid automatically. In this manner, we were able to determine the validity of the images with the most popular image hashes, and excluded over 3,000 invalid images from further consideration.

C. Textual Similarity

Our third connection type aimed to find significant blocks of shared text between pairs of websites. Scammers are known to reuse text between scams to save on the labour involved in website authorship, often with only minor adjustment e.g., in the form of altered site or pet names in fake testimonials.

As we focused on visible text body elements, we extracted all text content from the ⟨p⟩ tags on each site, forming natural blocks of textual content. Guided by previous work [7], our first approach was to then use a tokeniser to split the text into either a list of words or a list of sentences. The Jaccard index between the lists from two websites could then be computed to quantify similarity.

However, we found poor performance for this token-based method. The similarity measures for words were all extremely low, and did not particularly reflect our observations of duplicated content. Websites selling the same breed of animal naturally had a higher Jaccard index, but even large passages of duplicated text struggled to outweigh the dissimilar portions. Computing the index using sentences, while matching larger



(a) expressdeliveryllc.com (b) logicartexpress.com

Fig. 6: An example of two websites with a HTML tag similarity of 1.0.

blocks of text, provided even less favourable results overall. The largest Jaccard index value found after a trial analysing 1,000 paired sites was 0.38 for words and 0.15 for sentences.

The second approach we took to finding shared text between pairs of websites was to compute the longest common substring (LCS) between sites. This was implemented through the `find_longest_match` function from the `difflib` library[2], and was the most computationally expensive connection type to extract of the four used in this paper. To form a link in the resource network, we set a threshold of 400 characters for the LCS, which would typically equate to the majority of a paragraph being exactly duplicated.

The most common LCS was a paragraph of 586 characters (110 words) that was found in 67 websites and discusses the (extremely suspect) logistics of shipping a puppy by air.

> "Shipping a puppy by itself to a new location always sounds cruel and embarrassing, but actually I think it is harder for us than the puppy(s). With my many years of shipping experience, I know for a fact that all of the pups are well taken care of. So if you stop and think about it, the airlines are not going to mistreat the puppy(s) for fear of lawsuit and customer dissatisfaction. I tape puppy(s) food and feeding instructions to the top of the crate and put frozen water in the crate, so it will gradually thaw out for the puppy(s) and the puppies are offered food along the ride."

D. HTML Structural Similarity

The fourth and final connection type we examined relied upon similarities in the HTML structure of scam websites. This often occurs due to scammers' use of site templates and online tools that enable the quick setup of new scam sites, and has been identified as a useful feature for clustering other scam websites [7].

HTML tag frequencies were extracted from each website. We then computed the Jaccard index of the dictionaries between all pairs of websites. Figure 6 shows an example of two websites with a Jaccard index of 1.0 – an exact match for tag frequencies. The visual similarity of the webpages is correspondingly acute. There were 78 such pairs of websites, most of which were delivery scam websites. Scammers may be more inclined to clone delivery scam sites than pet scam sites because they are not a primary landing-site for the fraud, and do not need to be customised with images or customer testimonials. For the purposes of forming a connection in the resource network for HTML similarity, we set a Jaccard similarity threshold of 0.9, permitting for minor differences in tag frequency.

E. WHOIS Validation

All four methods used to find shared resource connections between scam websites in this paper use on-site data observable in webpage content. In order to validate these connections, we used off-site data. This data came from domain registration records obtained through WHOIS queries, and contained information including the IP address, registrar, date of registration and address. If a connection was found between a pair of websites, then in order to validate that connection, we compared the names of the registrars, the addresses and the IP addresses and checked for exact matches. If both the name of the registrar and address matched (i.e., the same registration details were used), or if the IP addresses matched (i.e., the same server is being used for both of the sites), then the pair of websites were marked as validated, since there is external evidence of these websites being created by the same person(s).

A number of the sites in our collection suffered from incomplete domain registration data. Some were missing values in fields such as the address, or the website had been taken down so the WHOIS data was no longer available. Many domains made use of identity protection services. This meant that it was not possible to validate every link found in a resource network. In total, 1,361 of the 1,780 domains revealed registration details viable for validation of on-site connections.

[2]https://docs.python.org/3/library/difflib.html

V. Results

A. Direct Links

There were 172 direct links from one website to another domain in our collection of 12,050 websites[3]. Some of these were linking to images that were hosted on another pet scam website's domain. 17% (296/1,780) of all scraped scam websites appear in this resource network. The vast majority of edges are only between websites of the same type. There is only one edge between a pet scam website and a delivery scam website. This link is because an image of a dog on a pet scam website is hosted on a delivery scam website.

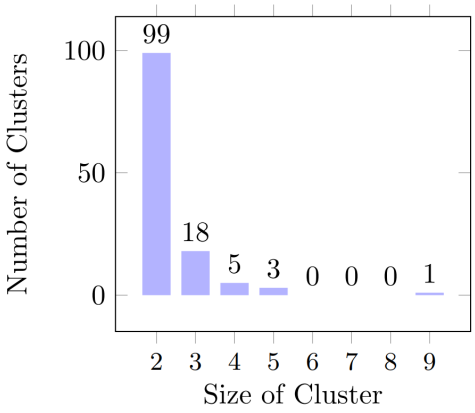


Fig. 7: Size of clusters in the resource network generated by finding direct links in HTML.

Figure 7 shows that the majority of clusters have two websites in them. The average degree (average number of edges each node has) is 1.16, also indicating that few websites had more than one other pet scam URL in them. The largest cluster identified contains nine websites, which constitutes around 3% of all the websites in the resource graph. In validation, we find that 47% (21/45) of the links with WHOIS data available are validated.

B. Shared Images

With a threshold of two shared images as a condition for forming a link, 2,417 connections can be found between 949 websites. The smaller clusters in this graph are more likely to be strongly connected. These complete sub-graphs were a result of groups of websites that all shared the same two images. Review of the domains involved revealed that the majority of these websites had similar names and were selling the same breed of animal. The largest fully-connected sub-graph of this kind has 8 nodes. One image appeared on 29 different domains. There is one pair of websites with more than 900 images in common. Both websites sell parrots and equipment needed to keep birds.

Figure 8 shows the largest cluster of domains when drawing connections through two shared images. The cluster contains 199 websites. These are for the most part separable into two

[3]123 sites linked to a domain no longer online.

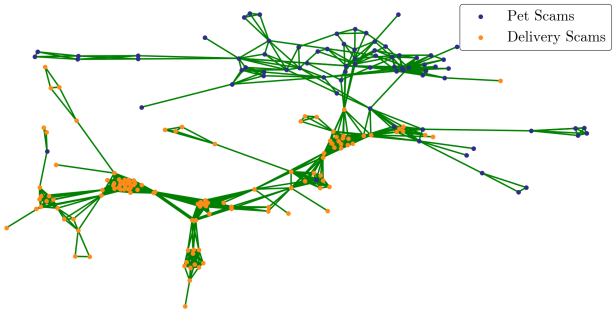


Fig. 8: The largest cluster in the resource network generated by shared images, containing 199 websites. Of these, 77 are pet scam websites and 122 are delivery scam websites.

distinct clusters, one half containing pet scam websites, and the other half containing mostly delivery scam websites. We can see that there are two websites preciousairsales.com and parrotsabode.com that join the two halves, sharing images with both pet scam and delivery scam websites. In validation, 50% (997/1,994) of connections were validated with common IP addresses or shared registration data.

C. Textual Similarity

There are 14,378 pairs of websites with a longest common substring (LCS) of at least 400 characters. This generated the largest resource network of all connection types, containing 1,325 domains, covering 74% (1,325/1,780) of all scraped websites. There are two main clusters in the resulting network. The largest cluster covers 70% (923/1,325) of all websites in the resource network, and mostly consists of pet scam websites. The next largest cluster contains 203 websites and mostly consists of delivery scam websites. The validation rate for LCS connections was the lowest of the four connection types, with only 34% (4,089/12,144) of the connections being validated with off-site data.

D. HTML Structural Similarity

There are 604 pairs of websites with a high HTML structural similarity. The vast majority of clusters were strongly connected and were of the same website type, with the largest cluster containing 15 delivery scam websites. Clusters only containing delivery scam websites were more likely to be strongly connected than clusters only containing pet scam websites. In validation, 77% (391/506) of these connections were validated using WHOIS data. This was the highest validation rate of a single connection type.

E. Combining Connection Types

To understand the degree to which connection types complement each other, we considered hybrid graphs combining evidence from different types of connection. We can set a threshold for the minimum number of distinct connection types required to be in agreement in order to form an edge in a hybrid resource network. Table I shows that setting the threshold to one—using all possible connections—results in

the largest graph, covering 90% of all sites in our collection, but with the lowest validation rate, with just 34% of those connections that can be checked in off-site data being validated by registration details or shared IP addresses. If we increase our threshold, we can increase the validation rate, but reduce the number of nodes captured in the resource network.

Connections used in resource network	Percentage of scraped websites included in resource network	Percentage of pairs of websites validated by WHOIS data
Direct links only	17%	47%
Shared images only	53%	50%
Shared text only	74%	34%
Similar HTML only	26%	77%
At least one connection	90%	34%
At least two connections	45%	62%
At least three connections	20%	74%

TABLE I: The percentage of pairs of websites with matching WHOIS data, given the type of connection that was used to generate the resource network.

	% validated by connection			
Connection	**Link**	**Images**	**Text**	**HTML**
Direct Link	-	20.9%	19.8%	8.1%
Shared Images	1.5%	-	55.3%	20.2%
Shared Text	0.2%	9.3%	-	3.7%
Similar HTML	3.2%	81.0%	89.2%	-

TABLE II: Percentage of edges with one type of connection that have an additional type of connection.

Table II shows the percentage of edges with one type of connection that also have an additional type of connection. It shows that websites with a high HTML similarity have a 89.2% chance of also having a LCS of least 400 characters, and a 81.0% chance of having at least two shared images. Since pairs of websites with a high number of shared resources are more likely to have been created by the same person(s), we can see that HTML similarity is the most reliable guide of the individual connection types, its connections being the most likely to be confirmed by any other connection type. If choosing a single connection to build a network, HTML similarity would provide the most accurate resource network.

Conversely, shared text edges are the least likely to be supported by any other types of connections. The most likely connection is shared images with a 9.3% chance. This, in combination with the low validation rate of text connections (34%) suggests that some text connections may be spurious, a possibility we discuss further below.

There is a positive correlation between connection types that share edges with other connection types, and the validation rate of connection types. Connection types that are more likely to share a second type of connection were more likely to have a higher WHOIS validation rate. The WHOIS validation rates for the shared text, direct links, shared images, and similar HTML were 34%, 47%, 50% and 77% respectively. The maximum percentage of edges with a second connection type were 9%, 21%, 55% and 89% respectively. This suggests a rank ordering of connection types by the likelihood they will be verified, providing stronger evidence of shared authorship.

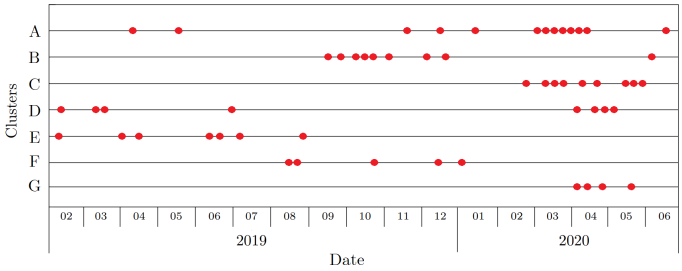


Fig. 9: The date when domains in the largest 7 clusters were registered according to WHOIS data. Clusters are from the resource network generated from websites with connections verified through domain registration data and at least three of the following: A direct link in HTML, at least two shared images, a LCS of at least 400 characters, or a HTML tag similarity of at least 0.9.

F. Targets of Interest

In what follows, we take the most conservative approach to building a resource network, requiring at least three connection types, and permitting only connections that are validated by domain registration details. This is severely restrictive, excluding many potential connections, but allows us high confidence that sites share an author. Seven clusters of four or more domains are recovered, the largest containing 13 domains.

Figure 9 shows the dates when websites in these seven largest clusters were registered. Most of the domains in a cluster were registered over a long time frame; only a few clusters had domains registered at a similar time. Cluster G had all four of its domains registered in a seven-week period in early 2020, and Cluster A had seven of its 13 domains registered within a seven-week period. Interestingly, this aligns with the COVID-19 pandemic period, as do half of the registrations for Cluster D and all of those for Cluster C. All of these clusters are homogeneous in composition, separating into pet and delivery scam sites. Delivery scam sites dominate, at five of seven clusters. Of the others, Cluster F reflects five sites that all claimed to sell bulldogs—suggesting a focus on working a particular market—while Cluster B connects nine sites advertising a number of different dog breeds.

Since these clusters share at least three types of resources and connections can be validated in domain registration data, we expect them to have been created by the same person(s). Domains being registered around a similar time supports this, suggesting an author preparing a number of alternatives at the same time, perhaps intentionally creating redundancy in case of takedowns. However, the spread of sites being registered, sometimes over more than a year, is also worth remarking upon, as it suggests an ongoing involvement from persistent criminals, reusing materials sometimes more than a year old.

To place these high-confidence clusters in the wider context of the resource networks that can be drawn, we note that links confirmed by only two connection types would join Cluster C and Cluster G into one 22-domain cluster of delivery scam sites, while growing the number of sites attached to the other clusters. Relaxing our criteria further, to allow for any single connection type, clusters A and E remain distinct groups of 15 and 8 domains, while the others appear to join as part of a large, 1,265-domain cluster of pet scam sites which may be authored by the same individual or group. This cluster connects pet scam and delivery scam sites and, though with weak evidence, suggests a common origin for 71% of the pet scam sites we have observed.

VI. Discussion

Our results show that the choice of connection used in building a resource network has important implications for the resulting clusters. Of the four connection types we examined, highly similar HTML tag frequency was the most often validated by domain registration details, and was also the most likely to be confirmed by a second type of connection. Given this, HTML similarity could be considered the most reliable single indicator of shared authorship, a finding that, if replicated, could have implications for the analysis of networks of online fraud other than pet scams. This is also, happily, a very lightweight comparison method, taking little computation time or storage space.

However, a high likelihood of being verified in other data is not necessarily the best measure for understanding the usefulness of connections. Significantly, only a quarter of domains appear in the resource network created through HTML similarity links. Moreover, our registration data for pet scam domains is limited, and these sites may not be fully representative of the wider population. Similarly, having multiple connection types confirm a link between domains is most valuable when these connection types are more usually uncorrelated, as they then provide independent indicators of common authorship when they do align.

Shared text is the most common type of connection, with over 14,000 edges found in our dataset of 1,780 nodes, and appears to link nearly three quarters of scam sites into clusters of varied sizes. It is possible that this reflects miscalibration —perhaps the LCS threshold is too low— but we should stress again that the threshold is 400 continuous characters of text, exactly matched between sites. This is unlikely to appear by chance in original writing, or to reflect snippets of website boilerplate text. It is therefore difficult to know how to interpret the low validation rate of shared text. Is shared text a misfiring connection type, showing correctly that large chunks of text are common between sites, but with this being primarily due to resources common to the industry, and not common authorship in a meaningful sense? Or is it the case, as shared text suggests, that perhaps as many as 923 pet scam domains are the work of one prolific scammer or group of collaborating scammers? Until conclusive evidence can be obtained regarding the individuals behind these online fraud sites, both interpretations may be considered plausible.

Our use of varied connection types has demonstrated that while pet scam sites and delivery scam sites are often mostly found in separate clusters, occasional mistakes cause certain domains to bridge these networks, highlighting reasons to suspect common authorship of both pet and delivery scam websites. These connections, if readily identified, could prove valuable evidence for prosecution purposes (or more pragmatically, administrative action like domain takedowns), contributing to the case that separate parts of the fraudulent activity were carried out by the same individual(s).

The anti-detection methods leveraged in pet scamming appear to be less mature than those deployed in phishing, suggesting an opportunity for techniques developed in anti-phishing research to be transferred to this domain. Such transfer should certainly be explored. However, pet scam websites differ from phishing sites in several important ways. In pet scamming, it is rarely the case that the site is impersonating a specific legitimate business – instead, scammers present themselves as being a legitimate small-scale vendor in a market containing many such vendors. Anti-phishing techniques that focus on identifying impersonation are therefore unlikely to be straightforwardly effective. Pet scam sites, while exhibiting signals of being fraudulent, are also not themselves the direct means by which the victim is exploited, with transactions between scammer and victim typically arranged over private communication channels. This disconnect poses additional—though by no means insurmountable—challenges for automated identification of pet scams, as classifiers must assess the likelihood of fraud based on soft signals in the site content, without mislabelling legitimate breeders' businesses.

Our finding was that 90% of sites could be connected to another pet scam site using one of the connection types we explore. This result seems to indicate that automatic classification of sites as pet scam sites may be achievable in a manner that does not also imperil legitimate pet breeding businesses — as new breeders should have no reason to use resources from known pet fraud sites, but new pet fraud sites almost always seem to. Such an automated classifier might prove more scalable than the current practices of manual investigation [2]. However, such an approach may have to be evaluated carefully against a registry of pre-existing legitimate breeders, to avoid them being mislabelled due to scammers copying their content.

Interest in pet scams appears to be rising, with some recent perpetrator arrests [15], [16]. Police attention may be drawn by rising levels of complaints during the pandemic [13], [14], with more customers searching for pets online and being more vulnerable to scammers' insistence on shipping pets. It may also be the case that scammers are more active during the pandemic, as has been observed more generally in cybercrime during national lockdowns [17]. Prosecution and traditional investigative methods have a valuable role in combatting pet scams, and may turn up further insight into the services and operational practices of fraudsters. To best serve the public

interest, law enforcement efforts should be directed towards high-volume repeat offenders. However, there is also a clear role for anti-fraud researchers, who should focus both on prioritising targets for law enforcement, assisting existing reporting and awareness approaches, and identifying other opportunities for disruption.

Future work could benefit from extending our resource-network analysis to different comparison features. Since the longest-common-substring feature was the only feature that looked at textual similarity and had the lowest validation rate, a new feature comparing the textual similarity of sites using a different method may be worthwhile. A promising candidate here might be stylometric comparison of the writing style between sites. This method would allow us to find connections between websites lacking exactly-matched substrings, but which do contain a similar writing style. Other connections that might be investigated include web analytics IDs [11], visual comparison of webpages [18], or properties of the contact details presented to victims to initiate the fraudulent transactions. More adventurously, research could benefit from attempts to contact pet scammers in the guise of victims, exposing reliable and current information about their contact details, email scripts, writing style and preferred payment methods.

VII. Conclusion

In this paper, we have examined over 1,780 domains involved with pet fraud, and then linked them to each other in networks of suspected common authorship through four different types of content-based connections. We found direct links from one site to another, images that are present on multiple websites, paragraphs of text that appear on multiple websites, and high structural similarity in the HTML tag frequencies of sites. We generated resource networks by setting a threshold on these connections. We then used off-site data from domain name registries to validate a subset of these connections.

Out of the four connection types we looked at, we discovered that comparing the similarity of HTML structure by using HTML tag frequency analysis had the highest rate of validation. At 77%, this was slightly higher than the 74% rate achieved by requiring at least three connection types. The connection type with the lowest validation rate was the longest common substring, with only 34% of these edges being validated. At our most permissive configuration, 90% of the scam websites in our collection could be linked to another by a single connection of any kind. When requiring three connection types to confirm a link, 20% of sites could still be linked. On balance, we find compelling reason to believe that a few individuals may be responsible for a large number of pet scam sites, suggesting some high-value targets for intervention.

For the interested reader, an interactive presentation of the resource networks we have built is available at https://petscams.herokuapp.com/. The data underlying this research, including the full crawled content of the observed pet scam sites, is available on request.

VIII. Acknowledgments

Our research would not be possible without the work of the volunteers who maintain the public lists of pet and delivery scam sites at http://petscams.com, and who work tirelessly to take down these sites and spread awareness of pet fraud.

References

[1] CBS News. (2017) Delta says bogus website tricks people who put pets on jets. Accessed: 10-05-2020. [Online]. Available: http://cbsnews.com/news/delta-says-mysterious-bogus-website-tricks-people-who-put-pets-on-jets/

[2] Better Business Bureau, "Puppy scams: How fake online pet sellers steal from unsuspecting pet buyers: A BBB study," 2018.

[3] N. S. A. Norazman and N. Zamin, "Development of scammed posts detector: A case study of pet scammed posting," in *Proceedings of the 18th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI)*. IIIS, 2014.

[4] M. Bidgoli and J. Grossklags, ""Hello. This is the IRS calling.": A case study on scams, extortion, impersonation, and phone spoofing," in *2017 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 2017, pp. 57–69.

[5] M. Edwards, G. Suarez-Tangil, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty, "The geography of online dating fraud," in *Workshop on Technology and Consumer Protection*. IEEE, 2018.

[6] R. Phillips and H. Wilder, "Tracing cryptocurrency scams: Clustering replicated advance-fee and phishing websites," *arXiv preprint arXiv:2005.14440*, 2020.

[7] J. Drew and T. Moore, "Automatic identification of replicated criminal websites using combined clustering," in *2014 IEEE Security and Privacy Workshops*. IEEE, 2014, pp. 116–123.

[8] N. Miramirkhani, O. Starov, and N. Nikiforakis, "Dial one for scam: A large-scale analysis of technical support scams," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2017.

[9] Y. Zhou, E. Reid, J. Qin, H. Chen, and G. Lai, "US domestic extremist groups on the web: Link and content analysis," *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 44–51, 2005.

[10] J. Qin, Y. Zhou, and H. Chen, "A multi-region empirical study on the internet presence of global extremist organizations," *Information Systems Frontiers*, vol. 13, no. 1, pp. 75–88, 2011.

[11] O. Starov, Y. Zhou, X. Zhang, N. Miramirkhani, and N. Nikiforakis, "Betrayed by your dashboard: Discovering malicious campaigns via web analytics," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 227–236.

[12] N. Leontiadis, T. Moore, and N. Christin, "Pick your poison: pricing and inventories at unlicensed online pharmacies," in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, 2013, pp. 621–638.

[13] Better Business Bureau. (2020) BBB warning: Puppy scam reports skyrocket during COVID-19 pandemic. Accessed: 29-08-2020. [Online]. Available: https://www.bbb.org/article/news-releases/22363-is-that-quarantine-puppy-real-puppy-scam-reports-skyrocket-during-covid-19-pandemic-bbb-warns

[14] ActionFraud. (2020) Animal lovers looking for pets in lockdown defrauded of nearly £300,000 in two months. Accessed: 29-08-2020. [Online]. Available: https://www.actionfraud.police.uk/news/animal-lovers-looking-for-pets-in-lockdown-defrauded-of-nearly-300000-in-two-months

[15] L. Parsons. (2020) Moment 'puppy scammer' is arrested after allegedly fleecing families out of thousands of dollars for adorable blue Staffordshire bull terriers that never arrived. Accessed: 29-08-2020. [Online]. Available: https://www.dailymail.co.uk/news/article-8642073/Puppy-scammer-arrested-allegedly-fleecing-families-THOUSANDS.html

[16] E. Ranley. (2020) Woman arrested, charged over string of online puppy scams. Accessed: 29-08-2020. [Online]. Available: https://www.news.com.au/national/victoria/crime/woman-arrested-charged-over-string-of-online-puppy-scams/news-story/fd87b9be1d7982660a47cbdc2a10e8ef

[17] B. Collier, "Boredom, routine activities, and cybercrime during the pandemic," Cambridge Cybercrime Centre, COVID Briefing Paper 4.

[18] T.-C. Chen, S. Dick, and J. Miller, "Detecting visually similar web pages: Application to phishing detection," *ACM Transactions on Internet Technology (TOIT)*, vol. 10, no. 2, pp. 1–38, 2010.