Automatic Identification of Replicated Criminal Websites Using Combined Clustering

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

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

***Abstract*—To be successful, cybercriminals must figure out how to scale their scams. They duplicate content on new websites, often staying one step ahead of defenders that shut down past schemes. For some scams, such as phishing and counterfeit-goods shops, the duplicated content remains nearly identical. In others, such as advanced-fee fraud and online Ponzi schemes, the criminal must alter content so that it appears different in order to evade detection by victims and law enforcement. Nevertheless, similarities often remain, in terms of the website structure or content, since making truly unique copies does not scale well. In this paper, we present a novel combined clustering method that links together replicated scam websites, even when the criminal has taken steps to hide connections. We evaluate its performance against two collected datasets of scam websites: fake-escrow services and high-yield investment programs (HYIPs). We find that our method more accurately groups similar websites together than does existing general-purpose consensus clustering methods.**

I. Introduction and Background

Cybercriminals have adopted two well-known strategies for defrauding consumers online: large-scale and targeted attacks. Many successful scams are designed for massive scale. Phishing scams impersonate banks and online service providers by the thousand, blasting out millions of spam emails to lure a very small fraction of users to fake websites under criminal control [8], [20]. Miscreants peddle counterfeit goods and pharmaceuticals, succeeding despite very low conversion rates [12]. The criminals profit because they can easily replicate content across domains, despite efforts to quickly take down content hosted on compromised websites [20]. Defenders have responded by using machine learning techniques to automatically classify malicious websites [23] and to cluster website copies together [4], [16], [18], [27].

Given the available countermeasures to untargeted large-scale attacks, some cybercriminals have instead focused on creating individualized attacks suited to their target. Such attacks are much more difficult to detect using automated methods, since the criminal typically crafts bespoke communications. One key advantage of such methods for criminals is that they are much harder to detect until after the attack has already succeeded.

Yet these two approaches represent extremes among available strategies to cybercriminals. In fact, many miscreants operate somewhere in between, carefully replicating the logic of scams without completely copying all material from prior iterations of the attack. For example, criminals engaged in advanced-fee frauds may create bank websites for non-existent banks, complete with online banking services where the victim can log in to inspect their "deposits". When one fake bank is shut down, the criminals create a new one that has been tweaked from the former website. Similarly, criminals establish fake escrow services as part of a larger advanced-fee fraud [21]. On the surface, the escrow websites look different, but they often share similarities in page text or HTML structure. Yet another example is online Ponzi schemes called high-yield investment programs (HYIPs) [22]. The programs offer outlandish interest rates to draw investors, which means they inevitably collapse when new deposits dry up. The perpetrators behind the scenes then create new programs that often share similarities with earlier versions.

The designers of these scams have a strong incentive to keep their new copies distinct from the old ones. Prospective victims may be scared away if they realize that an older version of this website has been reported as fraudulent. Hence, the criminals make a more concerted effort to distinguish their new copies from the old ones.

While in principle the criminals could start over from scratch with each new scam, in practice it is expensive to recreate entirely new content repeatedly. Hence, things that can be changed easily are (e.g., service name, domain name, registration information). Website structure (if coming from a kit) or the text on a page (if the criminal's English or writing composition skills are weak) are more costly to change, so only minor changes are frequently made.

The purpose of this paper is to design, implement, and evaluate a method for clustering these "logical copies" of scam websites. In Section II we describe two sources of data on scam websites that we will use for evaluation: fake-escrow websites and HYIPs. In Section III we outline a combined clustering method that weighs HTML tags, website text, and file structure in order to link disparate websites together. We then evaluate the method compared to other approaches in the consensus clustering literature and cybercrime literature to demonstrate its improved accuracy in Section IV. In Section V we apply the method to the entire fake-escrow and HYIP datasets and analyze the findings. We review related work in Section VI and conclude in Section VII.

DOI 10.1109/SPW.2014.26

II. Data Collection Methodology

In order to demonstrate the generality of our clustering approach, we collect datasets on two very different forms of cybercrime: online Ponzi schemes known as High-Yield Investment Programs (HYIPs) and fake-escrow websites. In both cases, we fetch the HTML using `wget`. We followed links to a depth of 1, while duplicating the website's directory structure. All communications were run through the anonymizing service Tor [6].

Data Source 1: Online Ponzi schemes We use the HYIP websites identified by Moore et al. in [22]. HYIPs peddle dubious financial products that promise unrealistically high returns on customer deposits in the range of 1–2% interest, compounded *daily*. HYIPs can afford to pay such generous returns by paying out existing depositors with funds obtained from new customers. Thus, they meet the classic definition of a Ponzi scheme. Because HYIPs routinely fail, a number of ethically questionable entrepreneurs have spotted an opportunity to track HYIPs and alert investors to when they should withdraw money from schemes prior to collapse. Moore et al. repeatedly crawled the websites listed by these HYIP aggregators, such as `hyip.com`, who monitor for new HYIP websites as well as track those that have failed. In all, we have identified 4 191 HYIP websites operational between November 7, 2010 and September 27, 2012.

Data Source 2: Fake-escrow websites A long-running form of advanced-fee fraud is for criminals to set up fraudulent escrow services [21] and dupe consumers with attractively priced high-value items such as cars and boats that cannot be paid for using credit cards. After the sale the fraudster directs the buyer to use an escrow service chosen by the criminal, which is in fact a sham website. A number of volunteer groups track these websites and attempt to shut the websites down by notifying hosting providers and domain name registrars. We identified reports from two leading sources of fake-escrow websites, `aa419.org` and `escrow-fraud.com`. We used automated scripts to check for new reports daily. When new websites are reported, we collect the relevant HTML. In all, we have identified 1 216 fake-escrow websites reported between January 07, 2013 and June 06, 2013.

For both data sources, we expect that the criminals behind the schemes are frequently repeat offenders. As earlier schemes collapse or are shut down, new websites emerge. However, while there is usually an attempt to hide evidence of any link between the scam websites, it may be possible to identify hidden similarities by inspecting the structure of the HTML code and website content. We next describe a process for identifying such similarities.

III. Method for Identifying Replicating Criminal Websites

We now describe a general-purpose method for identifying replicated websites. Figure 1 provides a high-level overview. We now briefly describe each step before detailing each in greater detail below.

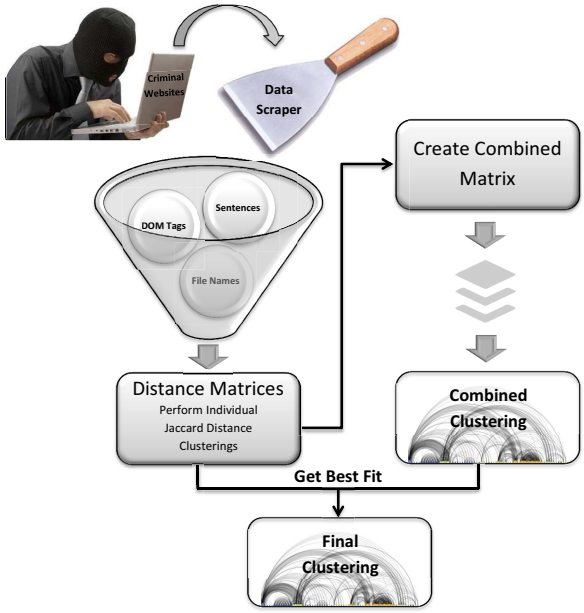


Fig. 1: High-level diagram explaining how the method works.

1) **Data scraper:** Raw information on websites is gathered (as described in Section II).
2) **Input attributes:** Complementary attributes such as website text and HTML tags are extracted from the raw data on each website.
3) **Distance matrices:** Pairwise similarities between websites for each attribute are computed using Jaccard distance metrics.
4) **Clustering stage 1:** Hierarchical, agglomerative clustering methods are calculated using each distance matrix, rendering distinct clusterings for each input attribute.
5) **Combined matrix:** A single distance matrix is calculated by combining the individual distance matrices.
6) **Clustering stage 2:** Hierarchical, agglomerative clustering methods are calculated using the combined distance matrix to arrive at the final clustering.

Extracting Website Input Attributes We identified three input attributes of websites as potential indicators of similarity: website text sentences, HTML tags and website file names.

To identify the text that renders on a given webpage, we used a custom "headless" browser adapted from the Watin package for C#[1]. We extracted text from all pages associated with a given website, then split the text into sentences using the OpenNLP sentence breaker for C#.

We extracted all HTML tags in the website's HTML files, while noting how many times each tag occurs. We then constructed a compound tag with the tag name and its frequency. For example, if the "
" tag occurs 12 times within the targeted HTML files, the extracted key would be "
12".

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

Fig. 2: Examples of replicated website content and file structures for the HYIP dataset.

Finally, we examined the directory structure and file names for each website since these could betray structural similarity, even when the other content has changed. However, some subtleties must be accounted for during the extraction of this attribute. First, the directory structure is incorporated into the filename (e.g., `admin/home.html`). Second, since most websites include a home or main page given the same name, such as `index.htm`, `index.html`, or `Default.aspx`, websites comprised of only one file may in fact be quite different. Consequently, we exclude this input attribute from consideration for such websites.

Constructing Distance Matrices For each input attribute, we calculated the Jaccard distance between all pairs of websites. The Jaccard distance between two sets S and T is defined as 1- $J(S,T)$, where:

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

Consider comparing website similarity by sentences. If website A has 50 sentences in the text of its web pages and website B has 40 sentences, and they have 35 in common, then the Jaccard distance is $1 - J(A,B) = 1 - \frac{35}{65} = 0.46$.

Clustering Stage 1 We compute clusterings for each input attributes using a hierarchical clustering algorithm [11]. Instead of selecting a static height cutoff for the resulting dendrogram, we employ a dynamic tree cut using the method described in [15]. These individual clusterings are computed because once we evaluate the clusters against ground truth, we may find that one of the individual clusterings work better. If we intend to incorporate all input attributes, this step can be skipped.

Creating a Combined Matrix Combining orthogonal distance measures into a single measure must necessarily be an information-lossy operation. A number of other consensus clustering methods have been proposed [2], [5], [7], [9] , yet as we will demonstrate in the next section, these algorithms do not perform well when linking together replicated scam websites, often yielding less accurate results than clusterings based on individual input attributes.

Consequently, we have developed a simple and more accurate approach to combining the different distance matrices. We define the pairwise distance between two websites a and b as the *minimum* distance across all input attributes. The rationale for doing so is that a website may be very different across one measure but similar according to another. Suppose a criminal manages to change the textual content of many sentences on a website, but uses the same underlying HTML code and file structure. Using the minimum distance ensures that these two websites are viewed as similar. Figure 2 demonstrates examples of both replicated website content and file structures. The highlighted text and file structures for each website displayed are nearly identical.

One could also imagine circumstances in which the average or maximum distance among input attributes was more appropriate. We calculate those measures too, mainly to demonstrate the superiority of the minimum approach.

Clustering Stage 2 We then cluster the websites for a second time, based upon the combined matrix. Once again, hierarchical clustering with dynamic cut tree height is used.

When labeled clusters are available for a sample of websites, the final step is to compare the combined clustering following stage 2 to the individual clusterings based on single input attributes. The more accurate method is selected for subsequent use.

	Tags	Files	Sent.	T&F	T&S	S&F	Combined
Fake-Escrow Services							
Minimum	0.672	0.072	0.100	0.603	0.900	0.097	**0.952**
Average	0.672	0.072	0.100	0.071	0.077	0.086	0.075
Max	0.672	0.072	0.100	0.075	0.076	0.093	0.080
DISTATIS	0.672	0.072	0.100	0.069	0.076	0.071	0.070
Clue SE	0.672	0.072	0.100	0.223	0.207	0.053	0.128
Clue DWH	0.672	0.072	0.100	0.214	0.307	0.081	0.126
Clue GV3	0.672	0.072	0.100	0.665	0.513	0.086	0.562
Clue soft/symdiff	0.672	0.072	0.100	0.132	0.115	0.087	0.095
High-Yield Investment Programs							
Minimum	0.388	0.245	**0.710**	0.276	0.444	0.255	0.301
Average	0.388	0.245	0.710	0.351	0.516	0.402	0.443
Max	0.388	0.245	0.710	0.314	0.668	0.638	0.623
DISTATIS	0.388	0.245	0.710	0.374	0.565	0.542	0.563
Clue SE	0.388	0.245	0.710	0.187	0.307	0.262	0.245
Clue DWH	0.388	0.245	0.710	0.275	0.464	0.389	0.472
Clue GV3	0.388	0.245	0.710	0.281	0.549	0.415	0.508
Clue soft/symdiff	0.388	0.245	0.710	0.259	0.423	0.340	0.401

(a) Adjusted Rand index for different clusterings, varying the number of input attributes considered.

	Escrow	HYIPs
Minimum	0.952	0.301
Average	0.075	0.443
Max	0.080	0.623
Best Min.	**0.952**	**0.710**
DISTATIS	0.070	0.563
Clue SE	0.128	0.245
Clue DWH	0.126	0.472
Clue GV3	0.562	0.508
Clue soft/symdiff	0.095	0.401
Click trajectories [18]	0.022	0.038

(b) Adjusted Rand index for different clusterings.

TABLE I: Table evaluating various consensus and combined clustering methods against ground truth dataset.

IV. Evaluation Against Ground-Truth Data

One of the fundamental challenges to clustering logical copies of criminal websites is the lack of ground-truth data for evaluating the accuracy of automated methods. Some researchers have relied on expert judgment to assess similarity, but most forego any systematic evaluation due to a lack of ground truth (e.g., [17]). We now describe a method for constructing ground truth datasets for samples of fake-escrow services and high-yield investment programs.

We developed a software tool to expedite the evaluation process. This tool enabled pairwise comparison of website screenshots and input attributes (i.e., website text sentences, HTML tag sequences and file structure) by an evaluator.

A. Performing Manual Ground Truth Clusterings

After the individual clusterings were calculated for each input attribute, websites could be sorted to identify manual clustering candidates which were placed in the exact same clusters for each individual input attribute's automated clustering. Populations of websites placed into the same clusters for all three input attributes were used as a starting point in the identification of the manual ground truth clusterings. These websites were then analyzed using the comparison tool in order to make a final assessment of whether the website belonged to a cluster. Multiple passes through the website populations were performed in order to place them into the correct manual ground truth clusters. When websites were identified which did not belong in their original assigned cluster, these sites were placed into the unassigned website population for further review and other potential clustering opportunities.

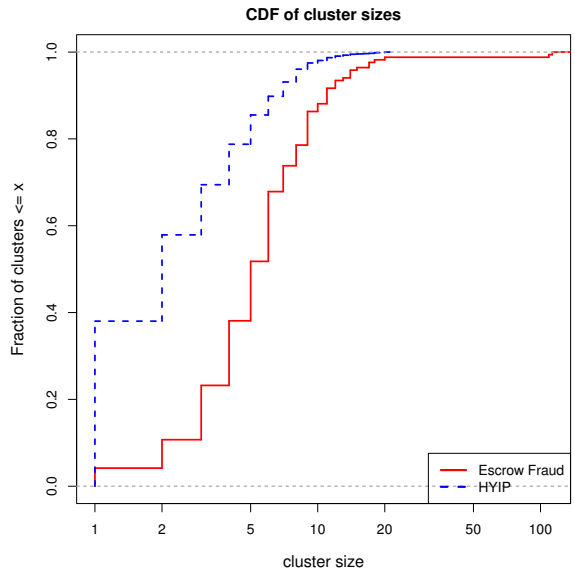
Deciding when to group together similar websites into the same cluster is inherently subjective. We adopted a broad definition of similarity, in which sites were grouped together if they shared most, but not all of their input attributes in common. Furthermore, the similarity threshold only had to be met for one input attribute. For instance, HYIP websites are typically quite verbose. Many such websites contain 3 or 4 identical paragraphs of text, along with perhaps one or two additional paragraphs of completely unique text. For the ground-truth evaluation, we deemed such websites to be in the same cluster. Likewise, fake-escrow service websites might appear visually identical in basic structure for most of the site. However, a few of the websites assigned to the same cluster might contain extra web pages not present in the others.

We note that while our approach does rely on individual input attribute clusterings as a starting point for evaluation, we do not consider the final combined clustering in the evaluation. This is to maintain a degree of detachment from the combined clustering method ultimately used on the datasets. We believe the manual clusterings identify a majority of clusters with greater than two members. Although the manual clusterings contain some clusters including only two members, manual clustering efforts were ended when no more clusters of greater than two members were being identified.

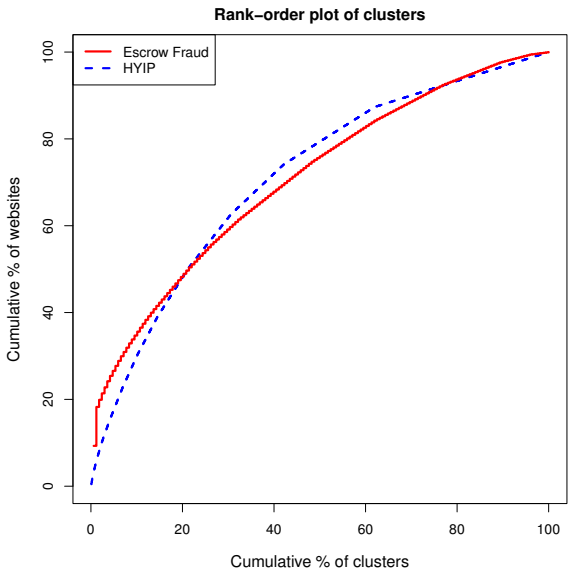
B. Results

In total, we manually clustered 687 of the 4 191 HYIP websites and 684 of the 1 221 fake-escrow websites. We computed an adjusted Rand index [24] to evaluate the combined clustering method described in Section III against the constructed ground-truth datasets described in Section V. We also evaluated other consensus clustering methods for comparison. Rand index ranges from 0 to 1, where a score of 1 indicates a perfect match between distinct clusterings.

Table Ia shows the adjusted Rand index for both datasets for all combinations of input attributes, combined, and consensus clustering methods. The first three columns show the Rand index for each individual clustering. For instance, for fake-escrow services, clustering based on tags alone yielded a Rand index of 0.672. Thus, clustering based on sentences alone is much more accurate than by file structure or website sentences alone (Rand indices of 0.072 and 0.10 respectively). When combining these input attributes, however, we see further im-



(a) Cumulative distribution function of cluster size.



(b) Rank order plot of cluster sizes.

Fig. 3: Evaluating the distribution of cluster size in the escrow fraud and HYIP datasets.

provement. Clustering based on taking the minimum distance between websites according to HTML tags and sentences yield a Rand index of 0.9, while taking the minimum of all three input attributes yields an adjusted Rand index of 0.952. This combined score far exceeds the Rand indices for any of the other comparisons.

Because cybercriminals act differently when creating logical copies of website for different types of scams, the input attributes that are most similar can change. For example, for HYIPs, we can see that clustering by website sentences yields the most accurate Rand index, instead of HTML tags as is the case for fake-escrow services. We can also see that for some scams, combining input attributes does not yield a more accurate clustering. Clustering based on the minimum distance of all three attributes yields a Rand index of 0.301, far worse than clustering based on website sentences alone. This underscores the importance of evaluating the individual distance scores against the combined scores, since in some circumstances an individual input attribute or a combination of a subset of the attributes may fare better.

We used several general-purpose consensus clustering methods from R's Clue package [10] as benchmarks against the our "best minimum" approach:

1) **"SE"** - Implements "a fixed-point algorithm for obtaining soft least squares Euclidean consensus partitions " by minimizing using Euclidean dissimilarity [5], [10].
2) **"DWH"** - Uses an extension of the greedy algorithm to implement soft least squares Euclidean consensus partitions [5], [10].
3) **"GV3"** - Utilizes an SUMT algorithm which is equivalent to finding the membership matrix m for which the sum of the squared differences between $C(m) = mm'$ and the weighted average co-membership matrix $\sum_b w_b C(_m b)$ of the partitions is minimal [9], [10].
4) **"soft/symdiff"** - Given a maximal number of classes, uses an SUMT approach to minimize using Manhattan dissimilarity of the co-membership matrices coinciding with symdiff partition dissimilarity in the case of hard partitions [7], [10].

Table Ib summarizes the best-performing measures for the different combined and consensus clustering approaches. We can see that our "best minimum" approach performs best. It yields more accurate results than other general-purpose consensus clustering methods, as well as the custom clustering method used to group spam-advertised websites by the authors of [18].

V. Examining the Clustered Criminal Websites

We now apply the best-performing clustering methods identified in the prior section to the entire fake-escrow and HYIP datasets. The 4 191 HYIP websites formed 864 clusters of at least size two, plus an additional 530 singletons. The 1 216 fake-escrow websites observed between January and June 2013 formed 161 clusters of at least size two, plus seven singletons.

A. Evaluating Cluster Size

We first study the distribution of cluster size in the two datasets. Figure 3a plots a CDF of the cluster size (note the logarithmic scale on the x-axis). We can see from the blue dashed line that the HYIPs tend to have smaller clusters. In addition to the 530 singletons (40% of the total clusters), 662 clusters (47% of the total) include between 2 and 5 websites.

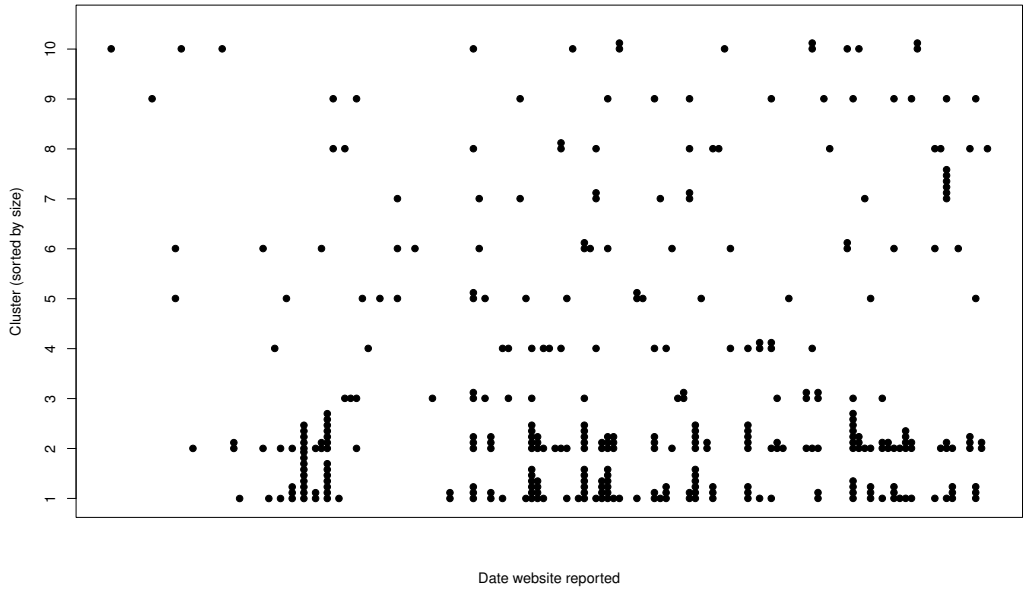


Fig. 4: Top 10 largest clusters in the fake-escrow dataset by date the websites are identified.

175 clusters (13%) are sized between 6 and 10 websites, with 27 clusters including more than 10 websites. The biggest cluster included 20 HYIP websites. These results indicate that duplication in HYIPs, while frequent, does not occur on the same scale as many other forms of cybercrime.

There is more overt copying in the escrow-fraud dataset. Only 7 of the 1 216 escrow websites could not be clustered with another website. 80 clusters (28% of the total) include between 2 and 5 websites, but another 79 clusters are sized between 6 and 20. Furthermore, two large clusters (including 113 and 109 websites respectively) can be found. We conclude that duplication is used more often as a criminal tactic in the fake-escrow websites than for the HYIPs.

Another way to look at the distribution of cluster sizes is to examine the rank-order plot in Figure 3b. Again, we can observe differences in the structure of the two datasets. Rank-order plots sort the clusters by size and show the percentage of websites that are covered by the smallest number of clusters. For instance, we can see from the red solid line the effect of the two large clusters in the escrow-fraud dataset. These two clusters account for nearly 20% of the total escrow-fraud websites. After that, the next-biggest clusters make a much smaller contribution in identifying more websites. Nonetheless, the incremental contributions of the HYIP clusters (shown in the dashed blue line) are also quite small. This relative dispersion of clusters differs from the concentration found in other cybercrime datasets where there is large-scale replication of content.

B. Evaluating Cluster Persistence

We now study how frequently the replicated criminal websites are re-used over time. One strategy available to criminals is to create multiple copies of the website in parallel, thereby reaching more victims more quickly. The alternative is to re-use copies in a serial fashion, introducing new copies only after time has passed or the prior instances have collapsed. We investigate both datasets to empirically answer the question of which strategy is preferred.

Figure 4 groups the 10 largest clusters from the fake-escrow dataset and plots the date at which each website in the cluster first appears. We can see that for the two largest clusters there are spikes where multiple website copies are spawned on the same day. For the smaller clusters, however, we see that websites are introduced sequentially. Moreover, for all of the biggest clusters, new copies are introduced throughout the observation period. From this we can conclude that criminals are likely to use the same template repeatedly until stopped.

Next, we examine the observed persistence of the clusters. We define the "lifetime" of a cluster as the difference in days between the first and last appearance of a website in the cluster. For instance, the first-reported website in one cluster of 18 fake-escrow websites appeared on February 2, 2013, while the last occurred on May 7, 2013. Hence, the lifetime of the cluster is 92 days. Longer-lived clusters indicate that cybercriminals can create website copies for long periods of time with impunity.

We use a survival probability plot to examine the distribution of cluster lifetimes. A survival function $S(t)$ measures the probability that a cluster's lifetime is greater than time t. Survival analysis takes into account "censored" data points, i.e., when the final website in the cluster is reported near the end of the study. We deem any cluster with a website reported within 14 days of the end of data collection to be censored. We use the Kaplan-Meier estimator [13] to calculate a survival function.

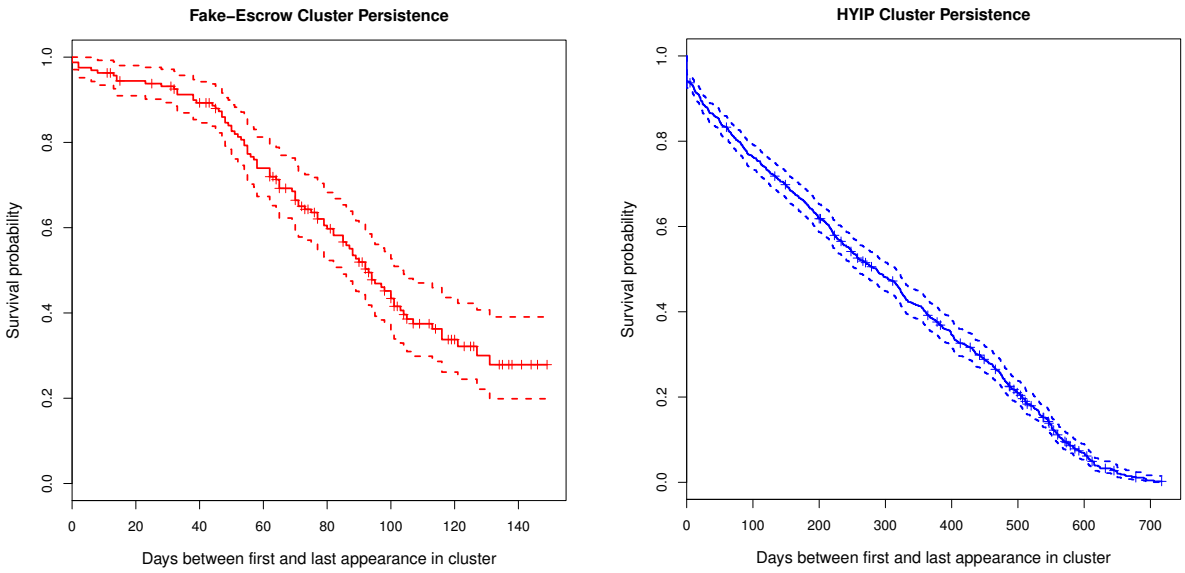


Fig. 5: Survival probability of fake-escrow clusters (left) and HYIP clusters (right).

Figure 5 gives the survival plots for both datasets (solid lines indicate the survival probability, while dashed lines indicate 95% confidence intervals). In the left graph, we can see that around 75% of fake-escrow clusters persist for at least 60 days, and that the median lifetime is 90 days. Note that around 25% of clusters remained active at the end of the 150-day measurement period, so we cannot reason about how long these most-persistent clusters will remain.

Because we tracked HYIPs for a much longer period (Figure 5 (right)), nearly all clusters eventually ceased to be replicated. Consequently, the survival probability for even long-lived clusters can be evaluated. 20% of HYIP clusters persist for more than 500 days, while 25% do not last longer than 100 days. The median lifetime of HYIP clusters is around 250 days. The relatively long persistence of many HYIP clusters should give law enforcement some encouragement: because the criminals reuse content over long periods, tracking them down becomes a more realistic proposition.

VI. Related Work

A number of researchers have applied machine learning methods to cluster websites created by cybercriminals. Wardman et al. examined the file structure and content of suspected phishing webpages to automatically classify reported URLs as phishing [27]. Layton et al. cluster phishing webpages together using a combination of k-means and agglomerative clustering [16].

Several researchers have classified and clustered web spam pages. Urvoy et al. use HTML structure to classify web pages, and they develop a clustering method using locality-sensitive hashing to cluster similar spam pages together [25]. Lin uses HTML tag multisets to classify cloaked webpages [19]. Lin's technique is used by Wang et al. [26] to detect when the cached HTML is very different from what is presented to user. Finally, Anderson et al. use image shingling to cluster screenshots of websites advertised in email spam [4]. Similarly, Levchenko et al. use a custom clustering heuristic method to group similar spam-advertised web pages [18]. We implemented and evaluated this clustering method on the cybercrime datasets in Section IV. Finally, Leontiadis et al. group similar unlicensed online pharmacy inventories [17]. They did not attempt to evaluate against ground truth; instead they used Jaccard distance and agglomerative clustering to find suitable clusters.

Separate to the work on cybercriminal datasets, other researchers have proposed consensus clustering methods for different applications. DISTATIS is an adaptation of the STATIS methodology specifically used for the purposes of integrating distance matrices for different input attributes [3]. DISTATIS can be considered a three-way extension of metric multidimensional scaling [14], which transforms a collection of distance matrices into cross-product matrices used in the cross-product approach to STATIS. Consensus can be performed between two or more distance matrices by using DISTATIS and then converting the cross-product matrix output into into a (squared) Euclidean distance matrix which is the inverse transformation of metric multidimensional scaling [1].

Our work follows in the line of both of the above research thrusts. It differs in that it considers multiple attributes that an attacker may change (site content, HTML structure and file structure), even when she may not modify all attributes. It is also tolerant of greater changes by the cybercriminal than previous approaches. At the same time, though, it is more specific than general consensus clustering methods, which enables the method to achieve higher accuracy in cluster labelings.

VII. Concluding Remarks

When designing scams, cybercriminals face trade-offs between scale and victim susceptibility, and between scale and evasiveness from law enforcement. Large-scale scams cast a wider net, but this comes at the expense of lower victim yield and faster defender response. Highly targeted attacks are much more likely to work, but they are more expensive to craft. Some frauds lie in the middle, where the criminals replicate scams but not without taking care to give the appearance that each attack is distinct.

In this paper, we propose and evaluate a combined clustering method to automatically link together such semi-automated scams. We have shown it to be more accurate than general-purpose consensus clustering approaches, as well as approaches designed for large-scale scams such as phishing that use more extensive copying of content. In particular, we applied the method to two classes of scams: HYIPs and fake-escrow websites.

The method could prove valuable to law enforcement, as it helps tackle cybercrimes that individually are too minor to investigate but collectively may cross a threshold of significance. For instance, our method identifies two distinct clusters of more than 100 fake escrow websites each. Furthermore, our method could substantially reduce the workload for investigators as they prioritize which criminals to investigate.

There are several promising avenues of future work we would like to pursue. First, the accuracy of the HYIP clustering could be improved. Second, it would be interesting to compare the accuracy of the combined clustering method to other areas where clustering has already been tried, such as in the identification of phishing websites and spam-advertised storefronts. Finally, additional input attributes such as WHOIS registration details and screenshots could be considered.

Acknowledgments

We would like to thank the operators of `escrow-fraud.com` and `aa419.org` for allowing us to use their lists of fake-escrow websites.

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

References

[1] H. Abdi. *Encyclopedia of Measurement and Statistics*, pages 598–605. SAGE Publications, Inc., 2007.

[2] H. Abdi, A. O'Toole, D. Valentin, and B. Edelman. Distatis: The analysis of multiple distance matrices. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 42–42, June 2005.

[3] H. Abdi, L. J. Williams, D. Valentin, and M. Bennani-Dosse. Statis and distatis: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):124–167, 2012.

[4] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker. Spamscatter: Characterizing Internet scam hosting infrastructure. In *Proceedings of 16th USENIX Security Symposium*, pages 10:1–10:14, Berkeley, CA, USA, 2007. USENIX Association.

[5] E. Dimitriadou, A. Weingessel, and K. Hornik. A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(07):901–912, 2002.

[6] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second-generation onion router. In *13th USENIX Security Symposium*, Aug. 2004.

[7] A. V. Fiacco and G. P. McCormick. *Nonlinear programming: sequential unconstrained minimization techniques*. Number 4. Siam, 1990.

[8] D. Florencio and C. Herley. Evaluating a trial deployment of password re-use for phishing prevention. In *Second APWG eCrime Researchers Summit*, eCrime '07, pages 26–36, New York, NY, USA, 2007. ACM.

[9] A. Gordon and M. Vichi. Fuzzy partition models for fitting a set of partitions. *Psychometrika*, 66(2):229–247, 2001.

[10] K. Hornik. A clue for cluster ensembles. 2005.

[11] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[12] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage. Spamalytics: An empirical analysis of spam marketing conversion. In *Conference on Computer and Communications Security (CCS)*, Alexandria, VA, Oct. 2008.

[13] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[14] S. Krolak-Schwerdt. Three-way multidimensional scaling: Formal properties and relationships between scaling methods. In D. Baier, R. Decker, and L. Schmidt-Thieme, editors, *Data Analysis and Decision Support*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 82–90. Springer Berlin Heidelberg, 2005.

[15] P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree. *Bioinformatics*, 24(5):719–720, Mar. 2008.

[16] R. Layton, P. Watters, and R. Dazeley. Automatically determining phishing campaigns using the uscap methodology. In *eCrime Researchers Summit (eCrime), 2010*, pages 1–8, Oct 2010.

[17] 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*, pages 621–638. ACM, 2013.

[18] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage. Click trajectories: End-to-end analysis of the spam value chain. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, SP '11, pages 431–446, Washington, DC, USA, 2011. IEEE Computer Society.

[19] J.-L. Lin. Detection of cloaked web spam by using tag-based methods. *Expert Syst. Appl.*, 36(4):7493–7499, May 2009. Available at http://dx.doi.org/10.1016/j.eswa.2008.09.056.

[20] T. Moore and R. Clayton. Examining the impact of website take-down on phishing. In *Second APWG eCrime Researchers Summit*, eCrime '07, Pittsburgh, PA, Oct. 2007. ACM.

[21] T. Moore and R. Clayton. *The Impact of Incentives on Notice and Take-down*, pages 199–223. Springer, 2008.

[22] T. Moore, J. Han, and R. Clayton. The postmodern Ponzi scheme: Empirical analysis of high-yield investment programs. In A. D. Keromytis, editor, *Financial Cryptography*, volume 7397 of *Lecture Notes in Computer Science*, pages 41–56. Springer, 2012.

[23] N. Provos, P. Mavrommatis, M. Rajab, and F. Monrose. All your iFrames point to us. In *17th USENIX Security Symposium*, Aug. 2008.

[24] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[25] T. Urvoy, E. Chauveau, P. Filoche, and T. Lavergne. Tracking web spam with html style similarities. *ACM Trans. Web*, 2(1):3:1–3:28, Mar. 2008.

[26] D. Y. Wang, S. Savage, and G. M. Voelker. Cloak and dagger: dynamics of web search cloaking. In *Proceedings of the 18th ACM conference on Computer and communications security*, CCS '11, pages 477–490, New York, NY, USA, 2011. ACM. Available at http://doi.acm.org/10.1145/2046707.2046763.

[27] B. Wardman and G. Warner. Automating phishing website identification through deep MD5 matching. In *eCrime Researchers Summit, 2008*, pages 1–7. IEEE, 2008.