Automatic Detection of Clone Websites Using Combined Clustering.

Sachin Bhagwat, Nikhil Dagade, Mangesh Bhosale, Rushikesh Hingase

Abstract— To achieve success, Hackers try to find a way to scale their scams. They duplicate content on new websites, often staying one step forward to defenders that finish off past schemes. For a few scams, like phishing and counterfeit product retailers, the duplicated content remains nearly identical. In others, like advanced-fee fraud and online Ponzi schemes, the hackers should alter content that seems completely different, so to pretend detection by victims and application. So, similarities typically stay, in terms of the web site structure or content, since creating really distinctive copies doesn't scale well. In this paper we present machine-controlled ways to extract key web site options, HTML structure, file structure, and screenshots. We describe a method to mechanically establish the simplest combination of such attributes to most accurately cluster similar websites. To demonstrate the method, we have a liability to assess its performance against 2 collected datasets of scam websites: pretend written agreement services and high-yield investment programs (HYIPs). We have a liability to show that our technique a lot of accurately teams similar websites along than those existing general-purpose agreement agglomeration ways.

Index Terms— Clustering, Escrow fraud, Hierarchical agglomerative clustering; HTML feature extraction, HYIP fraud, Ponzi schemes.

I. INTRODUCTION

Cybercriminals have adopted 2 well-known methods for defrauding customers online: large-scale and targeted attacks. Several roaring scams square measure designed for enormous scale. Phishing scams impersonate banks and on-line service providers by the thousand, blasting out a lot of spam emails to lure a awfully little fraction of users to faux websites beneath criminal management [1,2]. Miscreants vend counterfeit product and prescription drugs, succeeding despite terribly low conversion rates [3]. The criminals profit because they will simply replicate content across domains, despite efforts to quickly take down content hosted on compromised

websites [1]. Defenders have responded by using machine learning techniques to mechanically classify malicious web sites [4] and to cluster website copies together [5-8]
Frauds may produce bank websites for non-existent banks, complete with on-line banking services wherever the victim will log in to examine their 'deposits'. On the surface, the written agreement websites look totally different. However they usually share similarities in page text or HTML structure. Yet another example is on-line Ponzi schemes known as high-yield investment programs (HYIPs) [10]. The programs supply outlandish interest rates to draw investors, which imply they inevitably collapse once new deposits dry up. Hence, the criminals create a lot of united effort to distinguish their new copies from the recent ones. While in theory the criminals might begin everywhere from scratch with every new scam, in follow, it's costly to recreate entirely new content repeatedly. Hence, things that may be modified simply area unit (e.g., service name, domain name, registration information). Web site structure (if returning from a kit) or the text on a page (if the criminal's English or writing composition skills area unit weak) are a lot of expensive to alter, thus solely minor changes area unit frequently created.
The purpose of this paper is to style, implement, and evaluate a way for cluster these 'logical copies' of scam websites. In Section II we can see high-level overview of the combined-clustering method. In Section III, we tend to describe two sources of knowledge on scam websites used for evaluation: fake written agreement websites and HYIPs. Next, Section IV details however individual web site options like HTML tags, web site text, file structure, and image screenshots area unit extracted to form pairwise distance matrices scrutiny the similarity between websites. In Section V, we outline two optimized combined-clustering methods that take all website options into thought so as to link disparate websites along. We tend to describe a unique technique of combining distance matrices by choosing the minimum pair wise distance. We review related work in Section VI,VII and conclude in Section VIII.

Manuscript received Jan, 2016.
Sachin Bhagwat, *Computer Engineering, Pune University/ PGMCOE Pune, India, 8888959552*
Nikhil Dagade*, Computer Engineering, Pune University/ PGMCOE Pune, India,.9860727225*
Mangesh Bhosale*, Computer Engineering, Pune University/ PGMCOE Pune, India,9096898563*
Rushikesh Hingase, *Computer Engineering, Pune University/ PGMCOE Pune, India, 8600251301*

II. PROCESS FOR DISTINCTIVE REPLICATED CRIMINAL WEBSITES

This paper describes all-purpose technique for distinctive replicated websites. Figure 1 provides a high-level overview

ISSN: 2278 – 1323

that is currently concisely represented before every step is mentioned in bigger detail within the following sections.

1. *Computer address crawler*: Raw data on websites is gathered.
2. *Computer address feature extraction*: Complementary attributes such as web site text and HTML tags, area unit extracted from the data for every computer address provided.
3. *Input attribute feature files*: Extracted options for each web site area unit saved into individual feature files for efficient pairwise distance calculation.
4. *Distance matrices*: Pairwise distances between websites for every attribute area unit computed victimization the Jacquard distance metrics.
5. *Individual clustering*: Ranked, clustered clustering strategies area unit calculated victimization every distance matrix, rendering distinct clusterings for each input attribute. 6. *Combined matrices*: Combined distance matrices area unit calculated victimization numerous individual distance matrix combinations.
7. *Ground truth selection*: Criminal websites area unit manually divided into replication clusters and used as a supply of ground truth.
8. *Cut height optimization*: Ground truth clusters area unit used in combination with the Rand index to spot the optimum cluster cut height for every input attribute.
9. *Combined clustering*: Ranked, clustered clustering strategies area unit calculated victimization every combined distance matrix to make any variety of multi-feature clusterings.
10. *Prime entertainer selection*: The Rand index is calculated for all clusterings against the bottom truth to identify the highest acting individual feature or combined feature set.

III. KNOWLEDGE COLLECTION METHODOLOGY

In order to demonstrate the generality of our clump approach, we have a tendency to collect datasets on 2 terribly completely different forms of cybercrime: on-line Ponzi themes referred to as HYIPs and fake written agreement websites. In each case, we have a liability to fetch the hypertext markup language using wget(). We have a liability to followed links to a depth of one, while duplicating the website's directory structure. All communications were run through the anonymizing service Tor [11].

3.1 *Knowledge Supply One:*

On-line Ponzi schemes we use the HYIP websites known by Moore et al. in [10]. HYIPs hawk dubious monetary product that promise unrealistically high returns on client deposits in vary of one hundred and twenty fifth to twenty interest, combined daily. HYIPs will 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. As a result of HYIPs habitually fail, a number of ethically questionable entrepreneurs have noticed associate opportunity to trace HYIPs and alert investors to once they should withdraw cash from schemes before collapse. Moore et al. repeatedly crawled the websites listed by these HYIP aggregators, like hyip.com, who monitor for new HYIP websites likewise as track those who have unsuccessful. In all, we've known four, 191 HYIP websites operational between seven November 2010 and twenty seven Gregorian calendar month 2012.

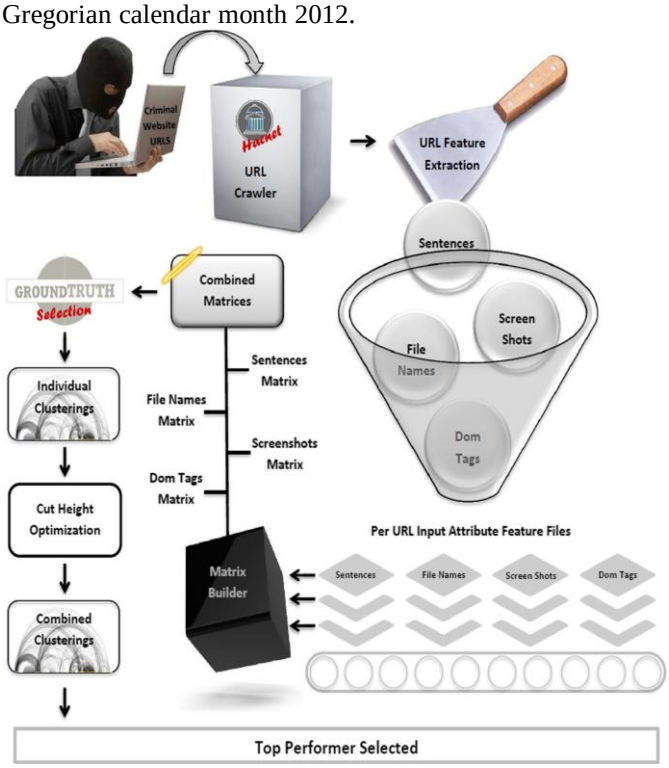

Figure 1 High-level diagram explaining how the method works.

3.2 *Knowledge Supply Two:*

Faux written agreement websites A long-running style of advanced-fee fraud is for criminals to set up dishonorable written agreement services [9] and dupe consumers with beautifully priced high-value things such as cars and boats that can't be obtained exploitation credit cards. once the sale, the fraudster directs the client to use an written agreement service chosen by the criminal, that is in reality a same website. Variety of volunteer teams track these websites and plan to shut the websites down by notifying hosting suppliers and name registrars. We identified reports from two leading sources of faux written agreement websites, aa419.org and escrow-fraud.com. We used automatic scripts to ascertain for brand new reports daily. When new websites are according, we have a tendency to collect the relevant HTML. In all, we've known one, 216 faux written agreement websites according between 07 January 2013 and 06 June 2013. For each knowledge sources, we have a tendency to expect that the criminals behind the schemes are often repeat offenders. As earlier schemes collapse or are clean up, new websites emerge. However, whereas there's sometimes associate attempt to hide proof of any link between the scam websites, it should be attainable to spot hidden similarities by inspecting the structure of the HTML code and web site content. We have a tendency to next describe a method for characteristic such similarities.

IV. EXTRACTING WEBSITE OPTIONS

We known four primary options of internet sites as potential indicators of similarity: displayed text, HTML tags, directory file names, and image screenshots. These are described in

ISSN: 2278 – 1323

Section four.1. In Section 4.2, we explain how the options area unit computed in a very pairwise distance matrix.

4.1 *Web Site Options*

4.1.1 *Web Site Text*

To identify the text that renders on a given website, we used a custom 'headless' browser custom-made from the WatiN package for C# [12]. We tend to extracted text from all pages related to a given web site, then split the text into sentences exploitation the OpenNLP sentence breaker for C#. Further lower level text options were conjointly extracted such as character n-grams, word n-grams, and individual words for similarity benchmarking. All text options were placed into individual baggage by web site. Baggages for every website were then compared to make pairwise distance matrices for agglomeration.

4.1.2 *Hypertext Markup Language Content*

Given that cybercriminals of times accept kits with similar underlying hypertext markup language structure, it's necessary to check the underlying hypertext markup language files additionally to the rendered text on the page. variety of selections exist, ranging from comparison the document object model (DOM) tree structure to treating tags on a page as a group of values. From experimentation, we tend to found that DOM trees were too specific, in order that even slight variations in otherwise similar pages yielded completely different trees.

4.1.3 *File Structure*

We examined the directory structure and file names for each web site since these might betray structural similarity, even once the opposite content has modified. However, some subtleties should be accounted for throughout the extraction of this attribute. First, the directory structure is incorporated into the file name (e.g., admin/home.html). Second, since most websites embrace a home or main page given the same name, like index.htm, index.html, or Default.aspx, websites comprised of just one file may in fact be quite completely different.

4.1.4 Web Site Screenshot Pictures

Finally, screenshots were taken for every websites exploitation the Selenium automatic applications program for C# [14]. Images were resized to one, 000 × 1, 000 pixels. We tend to calculate each vertical and horizontal luminousness histograms for every image. Image luminousness options and similarity measures were determined exploitation the EyeOpen image library for C# [15]. Throughout image feature extraction, the red, green, and blue channels for every image constituent were isolated to estimate relative brightness level, and these values were then aggregated by every vertical and horizontal image constituent row to calculate 2 luminousness histograms for every image.

4.2 *Constructing Distance Matrices*

For each input attribute, excluding pictures, we tend to calculate both the Jaccard and cos distances between all pairs of internet sites making pairwise distance matrices for each input attribute and distance live. Throughout analysis, it was determined that the Jaccard distance was the most correct metric for with success characteristic criminal website replications.

The Jaccard distance between 2 sets S and T is defined as one − J(S, T), where J(S, T) = |S ∩ T| |S ∪ T| Consider comparison web site similarity by sentences. If website A has fifty sentences within the text of its sites and website B has forty sentences, and that they have thirty five sentences in common, then the Jaccard distance is one − J(A, B) = 1 − 35 65 = 0.46. Website screenshot pictures were compared for each vertical and horizontal similarity exploitation luminousness histograms. The luminousness histograms for every matched image combine were compared for similarity by calculative the weighted mean between each the vertical and horizontal histograms. Next, each the common and most similarity values between histograms were through empirical observation evaluated for agglomeration accuracy. Taking the common similarity score between the vertical and horizontal histograms performed best throughout our analysis. Once the common vertical and horizontal similarity score made up our minds, and then the pairwise image distance was calculated as one - the pairwise image similarity. Distance matrices were created in parallel for every input attribute by 'mapping' web site input attributes into pairwise matches, and so at the same time 'reducing' pairwise matches into distances exploitation the suitable distance metric. The pairwise distance matrices were chosen as the output since they're the specified input for the hierarchical collective agglomeration method used throughout optimized agglomeration.

V. OPTIMIZED COMBINED-CLUSTERING METHOD

A. Once we've individual distance matrices for every input attribute as delineated within the previous section, the next step is to create the clusters. We have a tendency to initial describe 2 approaches for mechanically choosing cut heights for agglomerative clustering: dynamic cut height, which is unsupervised, and optimized cut height, that is supervised. Next, we have a tendency to reckon individual clusterings supported each input attribute. Finally, we have a tendency to construct combined distance matrices for mixtures of input attributes and cluster supported the combined matrices.

B. Cluster Cut Height Choice

We use a hierarchical agglomerated agglomeration algorithmic rule to cluster the websites supported the gap matrices. During HAC, a cut height parameter is needed to determine the difference threshold at that clusters are allowed to be incorporating along. This parameter greatly influences the agglomeration accuracy, as measured by the Rand index, of the ultimate clusters made.

C. Individual agglomeration

Because completely different classes of criminal activity might betray their likenesses in numerous ways in which, we

ISSN: 2278 – 1323

want a general process that may choose the simplest combination of input attributes for every dataset. we have a tendency to cannot grasp, a priori, which input attributes are most informative in revealing logical copies. Hence, we have a tendency to begin by agglomeration on every individual attribute severally, before combining the input attributes as delineated below. It's so quite plausible that one attribute higher identifies clusters than does a mix..

D. Best min combined agglomeration

While individual options will typically yield extremely correct clustering results, completely different individual options or maybe different mixtures of multiple options might perform better across completely different populations of criminal websites as our results can show. Combining multiple distance matrices into one 'merged' matrix may well be helpful once different input attributes are vital. However, combining orthogonal distance measures into a single measure must essentially be AN information-lossy operation. variety of alternative consensus-clustering ways have been projected, nevertheless as we'll demonstrate in the next section, these algorithms don't perform well once linking along replicated scam websites, often yielding less correct results than clusterings supported individual input attributes.

VI. ANALYSIS AGAINST GROUND TRUTH INFORMATION

One of the basic challenges of cluster logical copies of criminal websites are that the lack of ground truth data for evaluating the accuracy of automatic ways. Some researchers have relied on knowledgeable judgment to assess similarity, however most antedate any systematic analysis due to a scarcity of ground truth. We developed a software package tool to expedite the analysis process. This tool enabled pairwise comparison of website screenshots and input attributes (i.e., web site text sentences, HTML tag sequences, associated file structure) by an evaluator.

6.1 *Playacting Manual Ground Truth Clusterings*

After the individual clusterings were calculated for every input attribute, websites can be sorted to spot manual clustering candidates that were placed within the actual same clusters for every individual input attribute's automatic clustering. Populations of internet sites placed into the same clusters for all four input attributes were used as a start line within the identification of the manual ground truth clusterings. These websites were then analyzed using the comparison tool so as to create a final assessment of whether or not the web site belonged to a cluster. Multiple passes through the web site populations were performed so as to put them into the right manual ground truth clusters. Once websites were known but didn't belong in their original appointed cluster, these sites were placed into the unassigned web site population for more review and different potential cluster opportunities. Deciding once to cluster along similar websites into the same cluster is inherently subjective. For instance, HYIP websites area unit is generally quite windy. Many such websites contain 3 or four identical paragraphs of text, beside maybe one or 2 extra paragraphs of fully distinctive text. We note that whereas our approach will have faith in individual input attribute clusterings as a start line for analysis, we don't think about the ultimate combined cluster in the analysis. This can be to keep up a degree of detachment from the combined-clustering technique ultimately used on the datasets.

VII. EXAMINING THE CLUSTERED CRIMINAL WEBSITES

We currently apply the dynamic cut-height clustering methods presented earlier to the complete pretend written agreement (considering sentences, DOM tags, and file structure) and HYIP datasets (considering sentences alone). We conclude that duplication is employed more typically as a criminal maneuver within the pretend written agreement websites than for the HYIPs. Another way to appear at the distribution of cluster sizes is to look at the rank-order plot in Figure 4(right panel). Again, we are able to observe variations within the structure of the two datasets. Rank-order plots type the clusters by size and show the chances of internet sites that are lined by the smallest variety of clusters. For example, we can see from the red solid line the result of the 2 giant clusters in the pretend written agreement dataset. These 2 clusters account for nearly 2 hundredth of the entire pretend escrow websites. After that, the next biggest clusters build a far smaller contribution in identifying additional websites. With all, the progressive contributions of the HYIP clusters (shown within the broken blue line) are quite tiny. This relative dispersion of clusters differs from the concentration found in different cybercrime datasets wherever there's large-scale replication of content.

VIII. 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 sites to mechanically classify reported URLs as phishing [7]. Layton et al. cluster phishing web pages along employing a combination of k-means and agglomerate bunch [8]. Several researchers have classified and clustered internet spam pages. Urvoy et al. use hypertext markup language structure to classify web pages, and that they develop a bunch methodology mistreatment locality-sensitive hashing to cluster similar spam pages together. Sculptor uses hypertext markup language tag multiset to classify cloaked sites. Lin's technique is employed byWang et al. to observe once the cached hypertext markup language is incredibly completely different from what's bestowed to the user. Finally, Anderson et al. use image shingling to cluster screenshots of internet sites advertised in email spam [5]. Similarly, Levchenko et al. use a custom bunch heuristic methodology to cluster similar spam-advertised sites [6].

IX. CONCLUSION

When coming up with scams, Hackers face trade-offs between scale and victim condition and between scale and ambiguity from application. Large-scale scams cast a wider

ISSN: 2278 – 1323

web, however this comes at the expense of lower victim yield and quicker defender response. Extremely targeted attacks square measure far more seemingly to figure, however they're a lot of expensive to craft. Some frauds lie at the center, where the hackers replicate scams, however not while not taking care to give the looks that every attack is distinct. In this paper, we have a tendency to propose and measure a combined clustering method to mechanically link along such semi-automated scams. we've shown it to be more correct than general consensus-clustering approaches, moreover as approaches designed for large-scale scams like phishing that use a lot of in depth repeating of content. Above all, we have liability to applied the strategy to 2 classes of scams: HYIPs and pretend written agreement websites. The method may prove valuable to enforcement, as it helps tackle cybercrimes that on an individual basis square measure too minor to analyze however put together could cross a threshold of significance. for example, our methodology identifies two distinct clusters of over a hundred pretend written agreement websites each. what is more, our methodology may considerably reduce the employment for investigators as they range which criminals to analyze.

REFERENCES

[1] T Moore, R Clayton, in Second APWG eCrime Researchers Summit. eCrime*'07*. Examining the impact of website take-down on phishing (ACMPittsburgh, 2007)

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

[3] N Provos, P Mavrommatis, M Rajab, F Monrose, in *17th USENIX Security Symposium*. All your iFrames point to us, (2008).

[4] DS Anderson, C Fleizach, S Savage, GM Voelker, in *Proceedings of 16th USENIX Security Symposium*. Spamscatter: Characterizing Internet scam hosting infrastructure (USENIX Association Berkeley, 2007), pp. 10–11014. http://dl.acm.org/citation.cfm?id=1362903.1362913

[5] 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, GM Voelker, S Savage, in *Proceedings of the 2011 IEEE Symposiumon Security and Privacy. SP '11*. Click trajectories: end-to-end analysis of the spam value chain (IEEE Computer Society Washington, DC, 2011), pp. 431–446. doi:10.1109/SP.2011.24. http://dx.doi.org/10.1109/SP.2011.24

[6] B Wardman, G Warner, in *eCrime Researchers Summit, 2008*. Automating phishing website identification through deep MD5 matching (IEEE, 2008), pp. 1–7

[7] R Layton, P Watters, R Dazeley, in eCrime Researchers Summit (eCrime), *2010*. Automatically determining phishing campaigns using the uscap methodology, (2010), pp. 1–8. doi:10.1109/ecrime.2010.5706698

[8] T Moore, R Clayton, The Impact of Incentives on Notice and Take-down. (ME Johnson, ed.) (Springer, 2008), pp. 139–223

[9] T Moore, J Han, R Clayton, in *Financial Cryptography*. Lecture Notes in Computer Science, vol. 7397, ed. by Keromytis A D. The postmodern Ponzi scheme: Empirical analysis of high-yield investment programs (Springer, 2012), pp. 41–56. http://lyle.smu.edu/~tylerm/fc12.pdf

[10] R Dingledine, N Mathewson, P Syverson, in *13th USENIX SecuritySymposium*. Tor: The second-generation onion router, (2004)

[11] WatiN: Web application Testing in.Net. http://www.watin.org Accessed October 16, 2014

[12] D Florencio, C Herley, in Second APWG eCrime Researchers Summit. eCrime*'07*. Evaluating a trial deployment of password re-use for phishing prevention (ACM New York, 2007), pp. 26–36. doi:10.1145/1299015.1299018. http://doi.acm.org/10.1145/1299015.1299018.

Sachin Bhagwat
Computer Engineering,
Pune University/ PGMCOE
Pune, India, Mob 8888959552

Nikhil Dagade
Computer Engineering,
Pune University/ PGMCOE
Pune, India, Mob 9860727225

Mangesh Bhosale
Computer Engineering,
Pune University/ PGMCOE
Pune, India, Mob 9096898563

Rushikesh Hingase
Computer Engineering,
Pune University/ PGMCOE
Pune, India, Mob 8600251301

ISSN: 2278 – 1323