Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Malware and Phishing Website Detection Using Ensemble Clustering

Rashmi Karnik
Student, Department of Computer Engg.
BSIOTR, Savitribai Phule University,
Pune, Maharashtra, India

Prof. Gayatri M. Bhandari
Professor, Department of Computer Engg.
BSIOTR, Savitribai Phule University,
Pune, Maharashtra, India

Abstract*— Internet security new challenges are Malware and phishing website detection that has become prime interest now a days. In last few years, many clustering techniques have been working for to detect the fraud website usually through automatic malware and phishing website detection. The Malware and phishing website detection techniques, the process is generally divided into two steps: The first step is feature extraction, where representative features are extracted to capture the characteristics of the sample files or the websites; and The second step is categorization, where intelligent techniques are used to automatically group the samples of file or website into different classes based on computational analysis of the feature representations. The automatic categorization system, automatically club phishing websites and malware samples using a cluster ensemble by aggregating the clustering solutions that are generated by different types of base clustering algorithms. The principled cluster ensemble framework is used to integrate various partitions based on the individual clustering, which can be applied for both malware categorization and phishing website clustering.

Keywords*—Cluster ensemble, malware categorization, phishing website detection.

I. INTRODUCTION

Malware such as different types of viruses like worms, Trojan Horses, spyware, backdoors, and root kits has presented a serious threat for the computer systems to be secure. Currently, the most important line of security against malware is Internet security software products, which mainly used for client is a signature-based method to recognize the threats. Given sample is a collection of malware, these peddler first categorize the samples into families so that samples in the Same family shares some common traits, and generates the common string(s) to detect variants of a family of malware samples. Compared with malware attack, Phishing Website Detection is a relatively new Internet crime. Phishing is a online fraud, whereby perpetrators adopt social engineering schemes by sending e-mails, instant messages, or online advertising to allure users to phishing websites that imitate trustworthy websites in order to trick individuals into revealing their sensitive information (e.g., passwords, financial accounts, and personal identification numbers) which can be used for profit [16]. To secure against phishing websites, security software products generally use blacklisting to filter against known websites. However, there is always a delay between blacklist updating and website reporting. Indeed, as lifetimes of phishing websites are hold to hours from days, this method might be ineffective. The number of new phishing websites that are collected by the Kingsoft Antivirus Laboratory is usually larger than 20000 per day, and the number of new malware samples that are collected usually larger than 10000 per day. So there is an urgent need of effective and efficient methods for automatic detection of these threats. Though the phishing websites and the samples of malware evolve constantly, most of their essence or the inherent structure is relatively stable. For example, a malware family samples typically exhibit similar behaviour profiles [5]. It has also been shown that phishing websites are not abandoned from their targets but have strong relationships with them, which can be used as clues to cluster them into families and generate the signature for detection.

II. SYSTEM ARCHITECTURE

The architecture of Automatic Categorization System(ACS) is shown inn Fig. 1. Its each component is described below.:

1) *Term-frequency feature extractor*: Phishing website is categorized in ACS by using the term-frequency feature extractor to extract terms from webpages of the collected phishing websites, and then transforms the data into term-frequency feature vectors. These vectors are stored in the database and the transaction data can be easily converted to relational data if necessary.
2) *Extractor feature of Instruction-frequency*: Malware is categorized in ACS by using the instruction-frequency to extract the feature extractor of function-based instructions from the collected Portable Executable (PE) malware samples, the instructions are converted into a group of 32-bit global IDs as the features of the data collection, and stores in the signature database of signature features. These integer vectors are then transformed to instruction frequencies and stored in the database and the transaction data can be easily converted to relational data if necessary.
3) *Base clustering algorithms:* Base clustering is generated by applying different clustering algorithms that are

based on the feature representations. The KM partitioned approach and HC algorithm are applied on the Term-frequency vectors or instruction-frequency vectors with the TF and TF-IDF weighting schemes [17], which are widely used for representation document in IR (information retrieval).

4) *Cluster ensemble with constraints:* Cluster ensemble is applied to different base clustering algorithm in order to combine them. The cluster ensemble is also able to apply the domain knowledge in the form of website-level/sample-level constraints.
5) *Domain knowledge:* ACS is a user-friendly mechanism to incorporate the expert knowledge and expertise of human experts. Internet security experts can manually generate website-level/sample-level constraints which can be used to improve categorization performance.

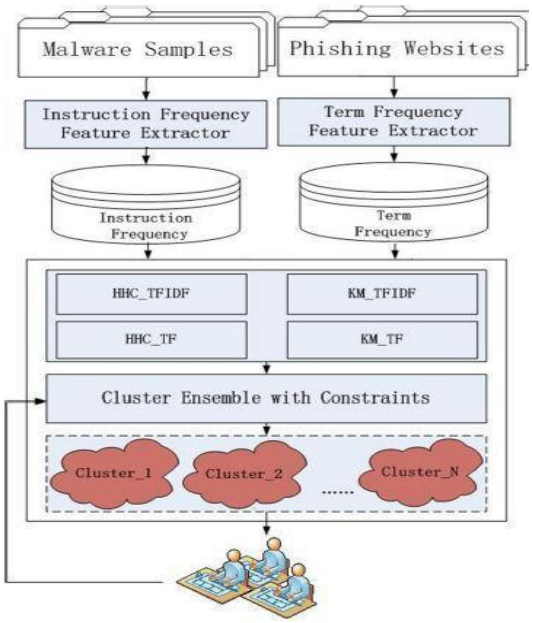

Fig. 1 Example of an unacceptable low-resolution image.

III. MALWARE CATEGORIZATION AND PHISHING WEBSITE DETECTION

A. Malware Feature Extraction

The behaviours of a program under analysis is characterize by features. These features are used as input to data mining algorithms and can be derived from different levels of abstractions, including instruction level, cross-module level and API level. The most used three categories of feature extraction methods are: static, dynamic and hybrid. Dynamic analysis techniques observe the execution of the malware to derive features. The execution can be on a virtual processor or real. Well-known techniques include debugging and profiling. One advantage of dynamic feature extraction is that the environment- or configuration-dependent information has been resolved in the extraction, e.g., a variable whose value depends on the hardware, system configuration, or program input. Its disadvantage is its limited coverage. Static analysis techniques analyse the malware without running it. The analysis target can be binary or source code. Static analysis has the advantage that it explore all possible execution paths in the malware; therefore, it can be exhaustive in detecting malicious logic. One static analysis disadvantage is its inability to address certain situations due to undesirability, e.g., transfer through the indirect control to the function pointers [18]. Hybrid analysis is an approach that combines static and dynamic analysis to achieve the both benefits. ACS use the instruction-frequency feature extractor to extract the instructions of function-based from the collected PE samples.

B. Malware Categorization

Different classification approach including association classifiers, support vector machines, and Naive Bayes have been applied in malware detection. Malware families detects by HOLMES [11] combines frequent sub graph mining and concept analysis to synthesize selective specifications. Research efforts have been described on combining different classification methods using different methods of learning with possible different feature representations from malware detection. For building classification model it require to frame a large number of training samples. In recent years, there have been several actions in automatic malware categorization using clustering techniques [7]. Using hierarchal clustering and locality sensitive hashing helps to efficiently group large datasets of malware samples into clusters. Mody and Lee [6] adopted KM clustering approach to categorize the malware samples. Several efforts have also been reported the similarities on computing between different malware samples using edit distance (ED) measure [15] or statistical tests.

C. Phishing Website Detection

Phishing technique target semantic attack to the user not the computer. It is a very large security threat for internet users as compared to malware. Many detection methods like say support vector machine or say naïve bayes for detection of phishing websites [1]. But the reality is that today there exist only few methods which efficiently detect phishing website detection Using clustering approach. Method starts by exploring of associated web pages with the give page,

then mark link. Ranking relationship, Webpage text, webpage layout similarity between given webpage and related webpage and final step is to apply DBSCAN [3] clustering algorithm to decide if any cluster exists around given web page. In case any webpage finder then is treated as phishing webpage or either as genuine page.

Layton [4] proposed the following framework for phishing website clustering: It first extracts the bag-of-word representation from the source of the websites and then principal component analysis (PCA) for feature selection, and, finally, uses certain clustering algorithms (such as k-means, DBSCAN) for detection. For example, the experiments of were performed based on 8745 phishing web pages and 1000 legitimate web pages, while Layton evaluated their proposed methods based on a dataset containing 24 403 websites. We believe that the further progress can be made in clustering particular sets of malware samples or sets of phishing websites. In particular, existing clustering methods usually apply a specific clustering method on a feature representation. Different clustering methods have their own advantages and limitations in malware detection. It propose a principled cluster ensemble framework to integrate different clustering solutions.

IV. FEATURE REPRESENTATION

A. Instruction Frequencies Of Malware Samples

There are mainly two types for feature extraction in malware analysis: static extraction and dynamic extraction. Dynamic feature extraction helps to present the behaviours of the malware files and analysing packed malware [5][6]. However, it has coverage of limited. Only Runnable files can be executed or simulated. Actually, for Kingsoft Internet Security Laboratory to the daily data collection, more than 60% of malware samples are dynamic link library files, which cannot be analysed dynamically. Also dynamic feature extraction is time consuming. It uses the informer K32Dasm which was developed by the Kingsoft Internet Security Laboratory to dissemble the PE code and output the file of unpacked or decrypt format as the input for feature extraction.

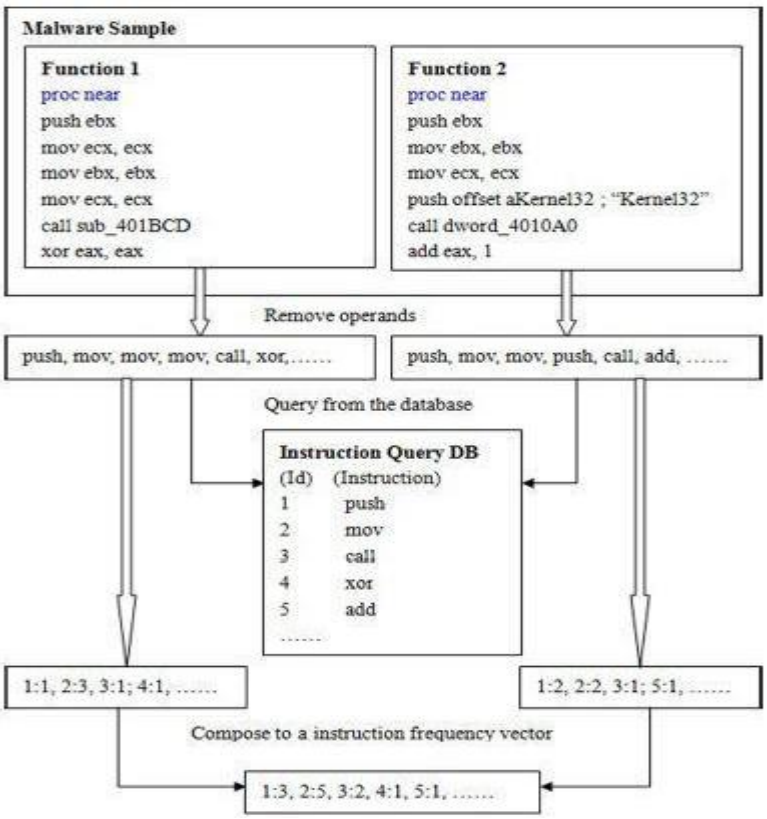

Fig. 2 Malware feature extraction and Transformation processes of the ACS

The instruction frequencies are used for malware representation. Malware feature extraction and transformation processes of ACS are shown in Fig. 2. Comparing with other static features that extraction and transformation processes uses, such as construction phylogeny tree, Windows API calls, control flow graph or arbitrary binaries, the instruction frequencies and function-based instruction sequences for malware representation have extended ability to represent variants of a malware family, high coverage rate of good semantic implications, malware samples, and high efficiency for feature extraction [8].

B. Term Frequencies Of Phishing Websites

There are many ways of feature extraction methods for phishing website representation: URL of the website, user interface associated web pages of the website, layout, webpage block, and overall style, terms of given webpage with the TF-IDF scores, etc. Considering the explanation capability of the website and the complexity for the categorization inputs, in this paper, it extract the phrase frequencies from the web pages of their corresponding websites. It extract the

terms from the "Title," "Alt" "Copyright," "Description," and "Keywords," of the web pages. The description of the extraction is illustrated as follows.

1) *Title:* extracting the content from the title tag of the webpage, i.e., the content between "<TITLE> . . . </TITLE>."
2) *Keywords:* extracting the keyword information of the website from the meta tag of the webpage, i.e., the content between "<META name=description content =. . .>."
3) *Description:* extracting the description information of the website from the meta tag of the webpage, i.e., the content between "<META name=keywords content =. . .>."
4) *Copyright:* extracting the copyright information of the website from the meta tag of the webpage, i.e., the content between "<META name=copyright content=. . .>."
5) *Alt:* extracting the text from the Alt tag of the webpage, i.e., the content between "<IMG alt=. . .>."

C. Characteristics of the Feature Representation

The phishing websites represented by the term frequencies of the webpage content share similar characteristics with malware samples defined with instruction frequencies. First, the feature representation is representative and can well group the instances of the same cluster. It has been observed malware samples that shares similar shapes of instruction frequency pattern are derived from same families or from same source of code, and they differ if they are derived from different malware families.

Second, the term frequencies of the webpage content and the instruction frequencies of file samples have similar distribution patterns. Fig. 3 shows the distribution of term frequency on a set of 2004 phishing websites with 3038 dimensions as well as instruction frequency on a sample dataset with 1434 malware samples with 1222 dimensions. These two features with TFIDF scheme have been extracted, and PCA is performed to extract the first two important dimensions for visualization As shown in Fig. 3, the distributions of phishing websites and malware samples are typically skewed, irregular, and of varied densities.

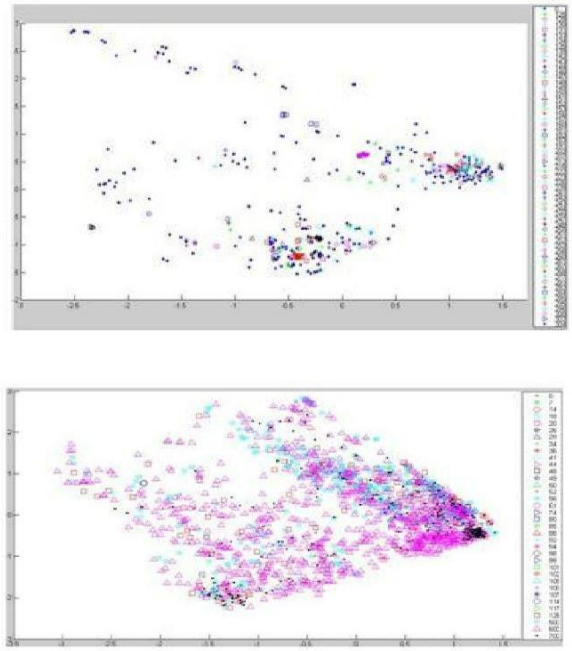

Fig. 3 Feature distributions after PCA transformation.

V. BASE CLUSTERING

In ACS ,a collection of phishing websites / malware is represented as a cluster i.e. group of components sharing some common properties and thus dissimilar components are placed at different cluster. Here it uses some basic approaches of clustering one of them is hierarchical and other is partitioning [9]. The HC approach produces irregular datasets more robustly, and during partitioning the cluster like in KM that produces tight cluster usually if clusters are of globular type. Which algorithm to choose for clustering will depend on factor of feature distributions. So HC and KM algorithm are used to form base clustering.

A. Hierarchical Clustering Algorithm

Hierarchical approach is grouped into two sub categories say agglomerative and divisive techniques. Out of this agglomerative gives lower computation cost. The algorithm starts as frame with *N* singleton clusters and then successfully merges two nearest clusters until only one remains. This technique is suitable for both phishing and malware detection or categorization .It cosine similarity to measure similarity [9] between two data points, because of its independent data length. The cosine similarity is described as below,

$$Dij = \cos\alpha = \frac{x^T_i \, x_j}{|x_i||x_j|}$$

Here x_i or x_j represents vector of two data points. There are many solution to problem of finding cosine similarity from C to all clusters: say complete, single and average linkage. Complete linkage produce clusters with rough equal diameter. Single linkage gives out to performance for recovery of compact clusters and return detection of elongated and irregular clusters. Average linkage uses characteristics of term frequency and instruction frequency feature .

The quality of clustering result is measure *by Fukuyama-Sugeno index(FS)* [10]. It is defined as

$$FS = \sum_{i=1}^{N} \sum_{j=1}^{nc} \sum^{m} uij(||x_i - v_j||_A^2 - ||v_j - v||_A^2)$$

where v_j is centre of cluster $C_{j,v}$[10]
V:is centre of data collection
A:n*n positive definite symmetric matrix (n= feature dimension)

B. K-Medoids Clustering Approach

K-mediods approach is a squared error based partioning algorithm. It takes a set of data points into clusters using a recursive relocation approach. A cluster is represented as real data point called mediods or data points called centriod .This algorithm shows effective results thus used in scientific and industrial applications. Assuming that phishing website and malware are of irregular density thus we use this algorithm.

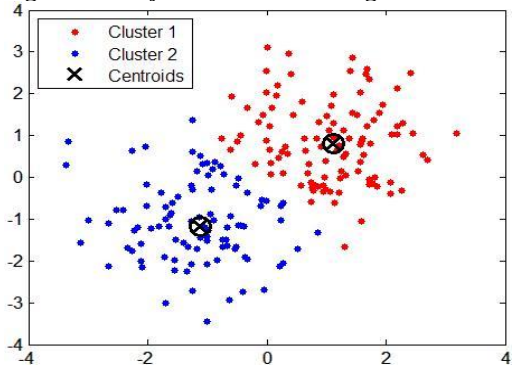

Fig. 4 k-Medoid Clustering Approach

VI. CLUSTER ENSEMBLE

A. Introduction

Clustering algorithms are very indispensable tools used for phishing website and malware categorization. But clustering is by default a difficult problem due to unavailability of supervision information. This is because during multiple run of different algorithm same algorithm would give different result due to random initializations [2]. Here it uses clustering basic algorithms to aggregate the solutions obtained by different algorithm. That is why we have used here hierarchical approach of clustering for providing internet security including categorization of phishing websites and malware detection.

B. Formulation

Let X = { $x_1, x_2, \ldots, x_n$} be a set of n data points (which can be phishing website or malware samples). Assume that we are given a set of T clustering i.e. let P = {$P^1, P^2, \ldots, P^T$ } be data point in X. Each partition P^t (t = 1, . . . , T) include a set of clusters $C^t = \{C_1^t, C_2^t, \ldots\ldots C_{Kt}^t\}$, where K_t = no of clusters for partition P^t and $X = \bigcup_{i=1}^{K} C_i^t$. Taking into consideration the fact that number of clusters K could be different for different Clustering.

The Connectivity Matrix $M(P^t)$ for partition P^t is as follows,

$$M_{ij} = \begin{cases} 1 & \text{If } x_i \text{ and } x_j \text{ belong to the same cluster } C^t \\ 0 & \text{otherwise} \end{cases}$$

By making use of the connectivity matrix, the distance between two partitions P^a, P^b are defined as follows[12][13],

$$d(P^a, P^b) = \sum_{i,j=1}^{n} d_{ij}(P^a, P^b)$$

$$= \sum_{i,j=1}^{n} |M_{ij}(P^a) - M_{ij}(P^b)|$$

$$= \sum_{i,j=1}^{n} [M_{ij}(P^a) - M_{ij}(P^b)]^2$$

Note that $|M_{ij}(P^a) - M_{ij}(P^b)| = 0$ or 1.

To find a consensus partition P* which is the closest to all the given partitions will be a general way for cluster ensemble:

$$\min_{P^*} J = \frac{1}{T} \sum_{t=1}^{T} d(P^t, P^*)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{i,j=1}^{n}[|M_{ij}(P^t) - M_{ij}(P^*)|]^2$$

Since J is convex in M(P*), by setting $\nabla_{M(P^*)}J = 0$, we can easily show that the partition P* that minimizes is the consensus (average) association: the ijth entry of its connectivity matrix is:

$$\tilde{M}_{ij} = \frac{1}{T}\sum_{t=1}^{T} M_{ij}(P^t) \qquad (1)$$

As per described in section V four base categories are made for the algorithm. 1) Two clustering's are obtained by applying HC on either term frequency or else instruction frequency vectors along with the TF-IDF and TF weighting schemes (denoted by HC_TFIDF and HC_TF); and 2) two clustering's by applying KM on the term-frequency vectors or instruction-frequency vectors with TF-IDF and TF weighting schemes with two different number of clusters: one is generated by HC_TFIDF, while the other is generated by HC_TF. The final clustering is obtained by the consensus association as per equation (1).

C. Incorporating Sample-Level Constraints

The domain knowledge that is in form of website level constraints can easily be embedded into a cluster ensemble. In addition to *t* partitions, given two sets of pair wise constraints:

1) Must link constraints

$$A = \{(x_{i1}, x_{j1}), \ldots, (x_{ia}, x_{ja})\},\ a = |A|$$

Here in this case every pair of points are said to be very close to similarity and hence clustered into the same cluster

2) Cannot link constraints

$$B = \{(x_{p1}, x_{q1}), \ldots, (x_{pb}, x_{pb})\},\ b = |B|$$

Here each and every pair of points are said to be far the way of similarity hence cannot be put in to same group. This constraints are mostly used for *semisupervised clustering* [14] but some research efforts have been reported on incorporating these constraints for process of cluster ensemble.

To combine this constraints in M and C into cluster ensemble ,we require to solve following problem:

$$\min_{p^*} J = \frac{1}{T}\sum_{t=1}^{T}\sum_{i,j=1}^{n}[|M_{ij}(P^t) - M_{ij}(P^*)$$

$$\text{s.t} \quad M_{ij}(P^*) = 1, \text{ if } (x_i, x_j) \in A$$

$$M_{ij}(P^*) = 0, \text{ if } (x_i, x_j) \in B.$$

The above equation is a convex optimization problem with linear constraints. Let $C = A \cup B$ be the set of all constraints; then $c = |C| = |A| + |B|$. We can represent C as $C = \{(x_{i1}, x_{j1}, b_1), \ldots, (x_{ic}, x_{jc}, bc)\}$, where $b_s = 1$ if $(x_{is}, x_{js}) \in A$, and $b_s = 0$ if $(x_{is}, x_{js}) \in B$, $s = 1 \ldots c$.We can then rewrite convex optimization as

$$\min_{p^*} J = \frac{1}{T}\sum_{t=1}^{T}\sum_{i,j=1}^{n}[|M_{ij}(P^t) - M_{ij}(P^*)|]^2$$

$$\text{s.t. } (\mathbf{e}_{is})M(P*)\mathbf{e}_{js} = b_s,\ s = 1,2, \ldots c$$

Where $\mathbf{e}_{is} \in R^{n\times 1}$ is an indicator vector with only the *i*th element being 1 and all other elements being 0.

VII. CONCLUSION

In this paper we have discussed ACS technique which is used for phishing website categorization and malware categorization into clusters or families sharing some common features using different clustering techniques. The method of implementing base clustering is an important aspect of implementation of ACS. In future this work could be extended by exploring various base clustering algorithms, by extending ensemble framework for anomaly detection and by incorporating new domain knowledge methods in detection process.

REFERENCES

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proc. APWG eCrimeRes. Summit*, 2007, pp. 60–69.

[2] A. P. Topchy, A. K. Jain, andW. F. Punch, "Clustering ensembles:Models of consensus and weak partitions," *IEEE Trans. Pattern Anal. Mach.Intell.*, vol. 27, no. 12, pp. 1866–1881, Dec. 2005.

[3] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databasewith noise," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.

[4] R. Layton and P. Watters, "Determining provenance in phishing websites using automated conceptual analysis," in *Proc. eCrime Res. Summit*, 2009, pp. 1–7.

[5] M. Bailey, J. Oberheide, J Andersen, Z. M. Mao, F. Jahanian, and J. Nazario, "Automated classification and analysis of internet malware," in *Recent Advances in Intrusion Detection*, (Lecture Notes in Computer Science vol. 4637). New York: Springer, 2007, pp. 178–197.

[6] T. Lee and J. J. Mody, "Behavioral classification," in *Proc. EICAR*, May 2006.

[7] I. Gurrutxaga, O. Arbelaitz, J. M. Perez, J. Muguerza, J. I. Martin, and I. Perona, "Evaluation of Malware clustering based on its dynamic behaviour.,"in *Proc. 7th Australas. Data Mining Conf.*, 2008, pp. 163–170.

[8] Y. Ye, T. Li, Y. Chen, and Q. Jiang, "Automatic malware categorization using cluster ensemble," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl.Discovery Data Mining*, 2010, pp. 95–104.

[9] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans.Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[10] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program PAM)," *Finding Groups in Data: An Introduction to Cluster Analysis*, 1990.

[11] M. Fredrikson, S. Jha, M. Christodorescu, R. Sailer, and X. Yan, "Synthesizing near-optimal malware specifications from suspicious behaviors," in *Proc. IEEE Symp. Secur. Priv.*, Washington, DC IEEE Computer Society, May 2010, pp. 45–60.

[12] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," in *Proc.21st Int. Conf. Data Eng.*, 2005, pp. 341–352.

[13] T. Li and C. Ding, "Weighted Consensus Clustering," in *Proc. SIAM Int.Conf. Data Mining*, 2008, pp. 798–809.

[14] S. Basu, I. Davidson, and K. L. Wagstaff, Eds., *Constrained Clustering:Advances in algorithms, Theory, and Applications*. Boca Raton, FL:CRC Press, 2008.

[15] M. Gheorghescu, "An automated virus classification system," in *Proc.VIRUS BULLETIN CON.*, Oct. 2005.

[16] R. Layton, S. Brown, and P. Watters, "Using differencing to increase distinctiveness for phishingwebsite clustering," in *Proc. Symp.Workshops Ubiquitous, Autonom. Trusted Comput.*, 2009, pp. 488–492.

[17] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura, "Refinement of TF-IDF schemes for web pages using their hyperlinked neighboring pages," in *Proc. 14th ACM Conf. Hypertext and Hypermedia*, Aug. 2003,pp. 198–207

[18] A. Moser, C. Kruegel, and E. Kirda, "Limits of static analysis for malware detection," in *Proc. 23rd Annu. Computer Secur. Appl. Conf.*, 2007, pp. 421–430.

