Principal Components of Port-Address Matrices in Port-Scan Analysis

Hiroaki Kikuchi[1], Naoya Fukuno[1], Masato Terada[2], and Norihisa Doi[3]

[1] School of Information Technology, Tokai University, 1117 Kitakaname, Hiratsuka, Kangawa, 259-1292, Japan
[2] Hitachi, Ltd. Hitachi Incident Response Team (HIRT), 890 Kashimada, Kawasaki, Kanagawa, 212-8567, Japan
[3] Dept. of Info. and System Engineering, Facility of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo, Tokyo, 112-8551, Japan

Abstract. There are many studies aiming at using port-scan traffic data for the fast and accurate detection of rapidly spreading worms. This paper proposes two new methods for reducing the traffic data to a simplified form comprising significant components of smaller dimensionality. (1) Dimension reduction via Term Frequency – Inverse Document Frequency (TF-IDF) values, a technique used in information retrieval, is used to choose significant ports and addresses in terms of their "importance" for classification. (2) Dimension reduction via Principal Component Analysis (PCA), widely used as a tool in exploratory data analysis, enables estimation of how uniformly the sensors are distributed over the reduced coordinate system. PCA gives a scatter plot for the sensors, which helps to detect abnormal behavior in both the source address space and the destination port space. In addition to our proposals, we report on experiments that use the Internet Scan Data Acquisition System (ISDAS) distributed observation data from the Japan Computer Emergency Response Team (JPCERT).

1 Introduction

The Internet backbone contains port-scanning packets that are routinely generated by malicious hosts, e.g., worms and botnets, looking for vulnerable targets. These attempts are usually made on a specific destination port for which services with known vulnerable software are available. Ports 135, 138, and 445 are frequently scanned. There is also malicious software that uses particular ports to provide a "back door" to companies. The number of packets targeting the destination port for the back door is not large, but the statistics for these ports are sometimes helpful for detecting a new type of attack, a coordinated attack made by a botnet, or targeted attacks.

Related Works. There have been several attempts to identify attacks via changes in the traffic data observed by sensors distributed across the Internet. A honeypot is a semipassive sensor that pretends to be a vulnerable host in faked communications with intruders or worms [10]. Some sensors are *passive*

R. Meersman and Z. Tari (Eds.): OTM 2008, Part II, LNCS 5332, pp. 956–968, 2008.

in the sense that capture packets are sent to an unused IP address without any interaction. The Network Telescope [8], Internet Storm Center [11], DShield [12], and ISDAS [3] are examples of passive sensors.

There are many studies aiming at using port-scan traffic data for the fast and accurate detection of rapidly spreading worms. Kumar uses the characteristics of the pseudorandom number generation algorithm used in the Witty worm to reconstruct the spread of infected hosts [13]. Ishiguro et al. propose the Wavelet coefficients used as metrics for anomaly detection [14]. Jung et al. present an algorithm to detect malicious packets, called Sequential Hypothesis Testing based on Threshold of Random Walk (TRW) [2]. Dunlop et al. present a simple statistical scheme called the Simple Worm Detection Scheme (SWorD) [15], where the number of connection attempts is tested with threshold values.

The accuracy of detection, however, depends on an assumption that *the set of sensors is distributed uniformly over the address space*. Because the installation of sensors is limited to unused address blocks, it is not easy to ensure uniform sensor distribution. Any distortion of the address distribution could cause false detection and a misdetection, and therefore uniformity of sensor distribution is one of the issues we should consider. Nevertheless, it is not trivial to evaluate a distribution of sensors in terms of its uniformity because the traffic data comprise ports and addresses that are correlated in high-dimensional domains.

Contribution. This paper proposes a new method for reducing the traffic data to a simplified form comprising significant components of smaller dimensionality. Our contribution is twofold:

1. **Dimension reduction via TF-IDF values.** We apply a technique used in information retrieval and text mining, called the *TF-IDF weight*, given that there are similarities between our problem and the information retrieval problem. Both deal with high-dimensional data, defined sets of words (ports or addresses), and documents (sensors). Both sets are discrete. Most elements are empty.
2. **Dimension reduction via PCA.** Our second proposal is based on an orthogonal linear transformation, which is widely used as a tool in exploratory data analysis. PCA enables estimation of how uniformly the sensors are distributed over the reduced coordinate system. The results of PCA give a scatter plot of sensors, which helps to detect abnormal behavior in both the source address space and the destination port space.

We give experimental results for our method using the JPCERT/ISDAS distributed observation data.

2 Proposed Methods

2.1 Preliminary

We give the fundamental definitions necessary for discussion about the characteristics of worms.

Definition 1. *A* scanner *is a host that performs port-scans on other hosts, looking for targets to be attacked.*

A sensor *is a host that can passively observe all packets sent from scanners. Let S be a set of sensors $\{s_1, s_2, \ldots, s_n\}$, where n is the number of sensors.*

Typically, a scanner is a host that has some vulnerability and thereby is controlled by malicious code such as a worm or virus. Some scanners may be human operated, but we do not distinguish between malicious codes and malicious operators. Sensors have always-on static IP addresses, i.e., we will omit the dynamic behavior effects of address assignments provided via Dynamic Host Control Protocol (DHCP) or Network Address Translation (NAT).

An IP packet, referred to as a "datagram", specifies a *source address* and a *destination address*, in conjunction with a *source port number* and a *destination port number*, specified in the TCP header.

Definition 2. *Let P be a set of ports $\{p_1, p_2, \ldots, p_m\}$, where m is the number of possible port numbers. Let A be a set of addresses $\{a_1, a_2, \ldots, a_\ell\}$, where ℓ is the number of all IP addresses.*

In IP version 4, possible values for m and ℓ are 2^{16} and 2^{32}, respectively. Because not all address blocks are assigned as yet, the numbers of addresses and ports observed by the set of sensors are typically limited, i.e., $m \ll 2^{16}$, $\ell \ll 2^{32}$. To handle reduced address set sizes, we distinguish addresses with respect to the two highest octets. For example, address $a = 221.10$ contains the range of addresses from 221.10.0.0 through 221.10.255.255.

Let c_{ij} be the number of packets whose destination port is p_j that are captured by sensor s_i in duration T. Let b_{ik} be the number of packets that are observed by sensor s_i and sent from source address a_k. An *observation* of sensor s_i is characterized by two vectors

$$\boldsymbol{c}_i = \begin{pmatrix} c_{i1} \\ \vdots \\ c_{im} \end{pmatrix} \quad \text{and} \quad \boldsymbol{b}_i = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{im} \end{pmatrix},$$

which are referred to as the *port vector* and the *address vector*. All packets observed by n independent sensors are characterized by the $n \times m$ matrix $\boldsymbol{C}$ and $\ell \times n$ matrix $\boldsymbol{B}$ specified by $\boldsymbol{C} = (\boldsymbol{c}_1 \cdots \boldsymbol{c}_n)$ and $\boldsymbol{B} = (\boldsymbol{b}_1 \cdots \boldsymbol{b}_n)$. Matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ will usually contain many unexpected packets caused by possible misconfigurations or by a small number of unusual worms, which we wish to ignore to reduce the quantity of observation data.

2.2 Reduced Matrix Via TF-IDF Values

Observation by a limited number of sensors shows an incomplete and small fragment of the Internet traffic of unauthorized packets. Therefore, the observation matrices P and A are "thinly populated", i.e., most elements are empty. To

reduce the dimension of the matrices to a subset of the matrix comprising significant elements from the given P and A, we try to apply a technique used in information retrieval and text mining, called the *TF-IDF weight*.

The TF-IDF weight gives the degree of importance of a word in a collection of documents. The importance increases if the word is frequently used in the set of documents (TF) but decreases if it is used by too many documents (IDF). The *term frequency* in the given set of documents is the number of times the term appears in the document sets. In our study, we use the term frequency to evaluate how important a specific destination port p_j is to a given set of packets $C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$ observed by n sensors, and defined as the average number of packets for the port p_j, i.e.,

$$TF(p_j) = \frac{1}{n} \sum_{i=1}^{n} c_{ij}.$$

The *document frequency* of destination port p_j is defined by

$$DF(p_j) = \big|\{\boldsymbol{c}_i \in C | c_{ij} > 0, i \in \{1, \ldots, n\}\}\big|,$$

which gives the degree of "uselessness", because a destination port with the highest $DF(p_j) \approx n$ implies that the port is always specified by any sensor, and therefore we would regard the port p_j as unable to distinguish between sensors. By taking the logarithm of the inverse of the document frequency, we obtain a *TF-IDF* for a given port p_j as

$$\text{TF-IDF}(p_j) = TF(p_j) \cdot \log_2\left(\frac{n}{DF(p_j)} + 1\right),$$

where the constant 1 is used to avoid the *TF-IDF* of a port with $DF(p_j) = n$ from being zero.

Similarly to the destination port, we define the *TF-IDF* weight of source address a_k as $\textit{TF-IDF}(a_k) = TF(a_k) \cdot \log_2(\frac{n}{DF(a_k)} + 1)$, where

$$TF(a_k) = \frac{1}{n} \sum_{i=1}^{n} c_{ik},$$
$$DF(a_k) = \big|\{\boldsymbol{c}_i \in B | b_{ik} > 0, i \in \{1, \ldots, n\}\}\big|.$$

Note that a high value for *TF-IDF* is reached by a high term (port/address) frequency and a low document (sensor) frequency for the port among the whole set of packets, thereby working to filter out common ports. Based on the order of *TF-IDF* values, we can choose the most important destination ports within the 2^{16} possible values, from the perspective of frequencies of sets of packets.

2.3 Reduced Matrix Via PCA

PCA is a well-known technique, which is used to reduce multidimensional data to a smaller set that contributes most to its variance by keeping lower-order principal components and ignoring higher-order components.

Our goal is to transform a given matrix $\boldsymbol{C} = (\boldsymbol{c}_1 \cdots \boldsymbol{c}_m)$ of m dimensions (observations) to an alternative matrix Y of smaller dimensionality as follows.

Given a matrix of packets

$$\boldsymbol{C} = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix},$$

where c_{ij} is the number of packets such that the destination port is p_j, captured by sensor s_i, we subtract the mean for every port to obtain $\boldsymbol{C}' = (\boldsymbol{c}'_1 \cdots \boldsymbol{c}'_m)$, where

$$\boldsymbol{c}'_i = \begin{pmatrix} c_{i1} - \overline{c_1} \\ \vdots \\ c_{im} - \overline{c_m} \end{pmatrix}$$

and $\overline{c_j}$ is the average number of packets at the j-th port, i.e., $\overline{c_j} = 1/n \sum_{i=1}^{n} c_{ij}$.

PCA transforms $\boldsymbol{C}'$ to $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ such that, for $i = 1, \ldots, n$,

$$\boldsymbol{c}'_i = U\boldsymbol{y}_i = y_{i1}\boldsymbol{u}_1 + \cdots + y_{im}\boldsymbol{u}_m,$$

where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ are m unit vectors, called the *principal component basis*, which minimizes the mean square error of the data approximation. The principal component basis is given by a matrix U comprising the eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$, sorted in order of decreasing eigenvalue $\lambda_1 > \cdots > \lambda_n$, of the *covariance matrix* that is defined by

$$V = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{c}_i \boldsymbol{c}_i^{\top}.$$

From a fundamental property of eigenvectors, the elements of the principal component basis are orthogonal, i.e., $\boldsymbol{u}_i \cdot \boldsymbol{u}_j = 0$ for any $i \neq j \in \{1, \ldots, m\}$. This gives the matrix $Y = (\boldsymbol{y}_1 \cdots \boldsymbol{y}_m)$, where

$$\boldsymbol{y}_i = U^{\top} \boldsymbol{c}'_i = (y_{i1} \cdots y_{im})^{\top}, \tag{1}$$

which maximizes the variance for each element and gives a zero average, for $i = 1, \ldots, m$.

The first principal component, namely y_{i1}, contains the most significant aspect of the observation data, and the second component y_{i2} contributes the second most significant effect on its variance. These "lower-frequency" components give a first impression of the port-scanning pattern, even though the "higher-frequency" ones are ignored.

We apply the PCA transform not only to the matrix $\boldsymbol{C}$ defined over the sensor and port spaces ($n \times m$) but also to the matrix $\boldsymbol{B}$ of the sensor and the address spaces ($n \times m$), and to the transposed matrices $\boldsymbol{C}^{\top}$ and $\boldsymbol{B}^{\top}$. We use the notation $\boldsymbol{u}(C)$ and $\boldsymbol{u}(B)$ if we need to distinguish between matrices $\boldsymbol{C}$ and $\boldsymbol{B}$.

3 Analysis

We apply the proposed methods to the dataset of packets observed by sensors distributed over the Internet.

3.1 Experimental Data

ISDAS Distributed Sensors. ISDAS is a distributed set of sensors [3], under the operation of the JPCERT Coordination Center (JPCERT/CC), that can estimate the scale of a current malicious event and its performance.

Table 1 shows the statistics for $m = 30$ sensors from April 1, 2006 through March 31, 2007. The most frequently scanned sensor is s_1 with about 451,000 counts, which is 70 times that for the least frequently scanned sensor s_{15}. In this sense, the destination addresses to scan are not uniformly distributed.

Institutional Access Sensors. Table 2 shows a set of sensors installed in institutional LANs and some commercial Internet Service Providers (ISPs). The bandwidth and the method of address assignment are listed for each of the sensors.

3.2 TF-IDF Analysis

We show the results of TF-IDF analysis in Table 3, where the top 20 ports and source addresses (two octets) are listed in order of corresponding TF-IDF values. In the table, destinations 135, 445, ICMP, 139, and 30 are known as frequently scanned ports and are therefore listed at the top, while destination ports 23310 and 631 are listed because of their low DFs, implying their "importance" in classifying sensors. On the other hand, we note that the top 20 source addresses have higher DFs. For example, the third address 203.205 has DF = 16, i.e., the address is found by 16 of the 30 sensors.

Table 1. Statistics for ISDAS distributed sensors

	sensor	count	unique $h(x)$	$\Delta h(x)$[/day]
Average	–	146000	37820	104.9
Standard deviation	–	134900	29310	82.72
Max	s_1	450671	98840	270.79
Min	s_{15}	6475	1539	4.22

Table 2. Specification of sensors from Nov. 30, 2006 through Jan. 12, 2007

	s_{101}	s_{102}	s_{103}	s_{104}	s_{105}	s_{106}	s_{107}	s_{108}
Subnet class	B	C	B				C	C
Bandwidth [Mbps]	100	8	100				12	8
Type	inst. 1	ISP 1	institutional 2				ISP 2	ISP 3
IP assignment	static	dynamic	static				dynamic	dynamic

Table 3. Top 20 ports and addresses, ordered by TF-IDF value (ISDAS)

p_j	TF(p_j)	DF(p_j)	TF-IDF(p_j)	a_k	TF(a_k)	DF(a_k)	TF-IDF(a_k)
135	19499.73	29	20160.80	219.111	4668.60	23	5909.06
445	15326.47	27	16941.27	58.93	4490.57	24	5492.61
ICMP	6537.40	29	6759.03	203.205	2939.13	16	4786.70
139	5778.23	27	6387.03	222.148	3055.33	25	3612.39
80	3865.90	30	3865.90	61.252	2159.63	21	2929.92
1026	3705.97	30	3705.97	61.193	1994.30	21	2705.62
23310	789.57	2	2927.75	61.205	1858.40	21	2521.24
1433	2423.33	30	2423.33	220.221	2035.27	26	2326.52
631	552.17	3	1823.58	61.199	1810.27	25	2140.32
1027	1268.73	30	1268.73	222.13	1504.80	20	2114.94
1434	1130.90	27	1250.05	219.2	561.77	12	1076.51
137	989.53	26	1131.14	218.255	676.33	17	1060.48
4899	1007.90	30	1007.90	222.159	774.90	23	980.79
1025	713.13	29	737.31	220.109	722.17	22	946.15
4795	150.67	1	663.11	221.208	861.07	29	890.26
22	470.47	30	470.47	219.114	750.70	25	887.57
32656	119.17	2	441.88	203.174	408.50	12	782.80
12592	92.47	1	406.96	221.188	600.40	25	709.87
113	174.57	8	405.30	221.16	245.23	6	639.92
1352	108.37	2	401.83	219.165	533.77	25	631.08

Filtering out the less important ports and addresses in terms of TF-IDF values gives reduced matrices of 20 dimensions, which are small enough for the PCA transform to be applied.

3.3 PCA

We have performed PCA for each of the matrices $\boldsymbol{C}, \boldsymbol{B}, \boldsymbol{C}^\top$, and $\boldsymbol{B}^\top$, namely the ports-and-sensors, addresses-and-sensors, sensors-and-ports, and sensors-and-ports matrices, respectively.

Principal Component Basis. Table 4 shows the experimental results for the first two orthogonal vectors of principal component basis $\boldsymbol{u}_1(C), \boldsymbol{u}_2(C), \ldots$ for the ports-and-sensors matrix $\boldsymbol{C}$ and basis $\boldsymbol{u}_1(B), \boldsymbol{u}_2(B), \ldots$ for the addresses-and-sensors matrix $\boldsymbol{B}$. The elements indicated in boldface are the dominant elements of each basis. For example, the ports 445 and 135, having the largest (in absolute value) elements -0.37 and -0.36 in $\boldsymbol{u}_1(C)$, are the primary elements determining the value of the first principal component y_1. Informally, we regard the first coordinate as the *degree of well-scanned ports* because 445 and 135 are likely to be vulnerable. In the same way, the second principal component basis $\boldsymbol{u}_2(C)$ indicates attacks on web servers ($p = 80$) and ICMP, and we may therefore refer to y_2 as the *degree of http attack*. The second principal component has about half the effect of the projected values because eigenvalue λ_1 is almost double λ_2.

The addresses-and-sensors matrix $\boldsymbol{B}$ provides the principal component vectors indicating the degree of importance in source address set A, as shown in

Table 4. The first two vectors of principal component basis $\boldsymbol{u}_1(C), \boldsymbol{u}_2(C), \ldots$ for port matrix $\boldsymbol{C}$ and basis $\boldsymbol{u}_1(B), \boldsymbol{u}_2(B), \ldots$ for address matrix $\boldsymbol{B}$

p_j	$\boldsymbol{u}_1(C)$	$\boldsymbol{u}_2(C)$
445	**-0.37**	0.01
135	**-0.36**	0.01
137	-0.34	-0.07
1433	-0.33	0.17
4899	-0.30	0.27
1434	-0.30	0.16
1026	-0.28	-0.27
1025	-0.28	-0.01
1027	-0.25	-0.28
22	-0.23	0.08
32656	-0.13	-0.27
12592	-0.13	-0.27
139	-0.10	0.18
23310	-0.09	-0.03
80	-0.02	**0.45**
ICMP	-0.02	**0.44**
113	0.00	0.25
4795	0.00	0.25
631	0.05	-0.04
1352	0.09	-0.08
eigenvalue λ_i	6.19	2.49

a_k	$\boldsymbol{u}_1(B)$	$\boldsymbol{u}_2(B)$
221.188	**-0.54**	0.20
222.148	**-0.54**	0.20
219.114	-0.53	0.20
219.165	-0.28	**-0.52**
221.208	-0.17	-0.41
220.221	-0.14	**-0.59**
58.93	-0.01	-0.20
222.13	0.00	-0.09
222.159	0.01	-0.06
61.199	0.03	0.03
219.111	0.03	0.02
220.109	0.03	0.03
61.205	0.03	0.03
221.16	0.03	0.03
61.252	0.03	0.04
203.174	0.03	0.04
61.193	0.03	0.04
203.205	0.04	0.04
219.2	0.06	0.14
218.255	0.06	0.14
eigenvalue λ_i	3.16	2.29

Table 5, as well as in matrix C. In these results, we find that $\boldsymbol{u}_1(B)$ has dominant addresses that are disjoint from those of $\boldsymbol{u}_2(B)$.

Scatter Plot for Sensors in Reduced Coordinate System. In Fig. 1, we illustrate how the observed data are projected into the new coordinate system defined by the first two principal components y_1 and y_2 as the X-axis and Y-axis of the scatter plot for the sensors. The sensors $s_{101}, \ldots, s_{108}$, specified in Table 2, are indicated at the coordinate (y_{i1}, y_{i2}), computed by Eq. (1). The plot shows that there are three clusters: (1) sensors in institutional LANs, $\{s_{101}, s_{103}, \ldots, s_{106}\}$, (2) commercial ISPs, $\{s_{107}, s_{108}\}$, and (3) ISP 3, $\{s_{102}\}$. ISP 3 uses a cable modem, whereas the access network for ISP 1 and 2 is ADSL. We see that the two-dimensional principal components successfully retain the characteristics of each cluster of sensors. In other words, the 20-dimensional data for the ports are reduced to just two dimensions.

The resulting clusters depend on the given matrix. The same set of sensors are classified differently into the three clusters shown in Fig. 2 if we begin with the matrix $\boldsymbol{B}$. It is interesting that sensors s_{107} and s_{108} are distributed quite differently, even though they were close in Fig. 1.

Analysis from Several Perspectives. PCA can be applied to arbitrary matrices prepared from different perspectives. If we are interested in the independence of sensors, PCA enables us to show how uniformly the set of sensors is

Table 5. The principal component basis $\boldsymbol{u}_1(\boldsymbol{C}^\top), \boldsymbol{u}_2(\boldsymbol{C}^\top), \ldots$ for sensor-port matrix $\boldsymbol{C}^\top$ and basis $\boldsymbol{u}_1(\boldsymbol{B}^\top), \boldsymbol{u}_2(\boldsymbol{B}^\top), \ldots$ for sensor-address matrix $\boldsymbol{B}^\top$

s_i	$\boldsymbol{u}_1(\boldsymbol{C}^\top)$	$\boldsymbol{u}_2(\boldsymbol{C}^\top)$
s_7	-0.04	0.34
s_{20}	-0.03	0.30
s_8	-0.03	**0.42**
s_{22}	-0.01	**0.42**
s_{26}	-0.01	0.25
s_{30}	0.03	-0.12
s_{28}	0.05	-0.19
s_{12}	0.06	**0.37**
s_{15}	0.06	-0.16
s_{29}	0.07	-0.22
s_{25}	0.17	-0.01
s_{23}	0.18	-0.08
s_6	0.18	0.24
s_{24}	0.19	0.04
s_5	0.21	0.02
s_4	0.22	0.08
s_{17}	0.22	-0.12
s_{16}	0.22	-0.09
s_{21}	0.22	-0.02
s_{27}	0.23	-0.06
s_{13}	0.23	0.03
s_{14}	**0.24**	-0.02
s_{18}	**0.24**	0.10
s_{11}	**0.24**	0.07
s_{19}	**0.24**	0.01
s_3	**0.24**	0.05
s_1	**0.24**	0.03
s_2	**0.24**	0.01
s_{10}	**0.24**	-0.02
s_9	**0.24**	0.03
eigenvalue λ_i	16.64	3.73

s_i	$\boldsymbol{u}_1(\boldsymbol{B}^\top)$	$\boldsymbol{u}_2(\boldsymbol{B}^\top)$
s_{12}	**-0.34**	0.16
s_{18}	**-0.34**	0.18
s_6	**-0.34**	0.18
s_{20}	**-0.34**	0.02
s_{22}	**-0.34**	0.18
s_{13}	-0.32	0.21
s_{17}	-0.32	0.01
s_{29}	-0.28	-0.20
s_{28}	-0.21	**-0.35**
s_{27}	-0.20	-0.11
s_4	-0.17	-0.27
s_{23}	-0.10	**-0.33**
s_1	-0.05	-0.30
s_3	-0.05	-0.21
s_5	-0.03	-0.03
s_{11}	-0.01	0.03
s_{10}	0.00	-0.15
s_{14}	0.01	-0.08
s_{26}	0.01	-0.05
s_9	0.01	0.07
s_2	0.01	0.06
s_{15}	0.02	-0.11
s_{30}	0.02	-0.07
s_{16}	0.03	-0.00
s_{19}	0.03	0.12
s_{24}	0.04	0.15
s_8	0.04	0.13
s_{25}	0.04	**0.32**
s_{21}	0.06	**0.31**
s_7	0.07	0.18
eigenvalue λ_i	7.81	2.66

distributed over the reduced coordinate system. If we wish to identify abnormal behavior of source addresses, PCA with respect to a sensors-and-address matrix $\boldsymbol{B}^\top$ gives a scatter plot of addresses in which particular addresses stand out from the cluster of the standard behaviors.

For these purposes, we show the experimental results of ISDAS observation data, in Figs. 3, 4, 5, and 6, corresponding to matrices $\boldsymbol{C}$, $\boldsymbol{B}$, $\boldsymbol{C}^\top$, and $\boldsymbol{B}^\top$, respectively.

The set of ISDAS sensors is independently distributed in Fig. 3, but the distribution is skewed by some irregular sensors in Fig. 4, where the horizontal axis has more elements with source addresses in class C. As a consequence, the distribution of ISDAS sensors may be distorted in terms of differences between source addresses.

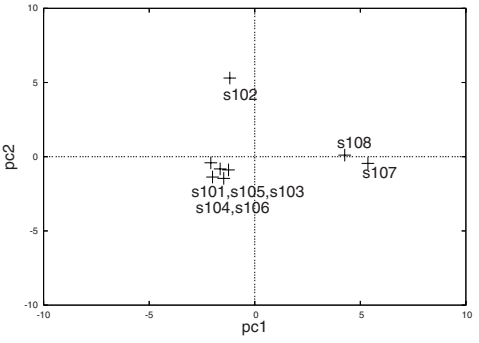

Fig. 1. Scatter plot for institutional sensors of a dataset with $n = 8$, indicating the coefficients of the first (X-axis) and second (Y-axis) principal components, $y_1(C)$ and $y_2(C)$, respectively

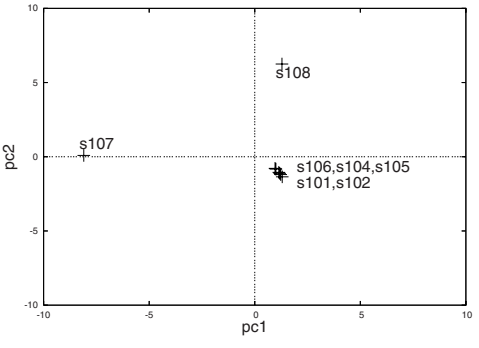

Fig. 2. Scatter plot for institutional sensors of a dataset with $n = 8$, indicating the coefficients of the first (X-axis) and the second (Y-axis) principal components, $y_1(B)$ and $y_2(B)$, respectively

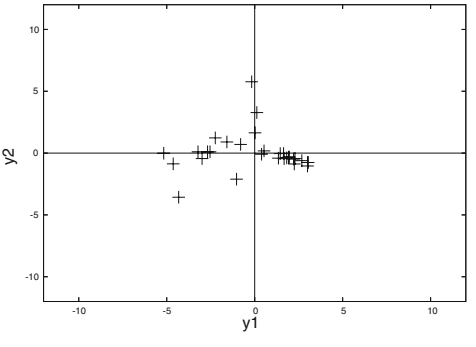

Fig. 3. Scatter plot for *ISDAS sensors* S of a dataset with $n = 30$, displaying the coefficients of the first two principal components in terms of *ports*

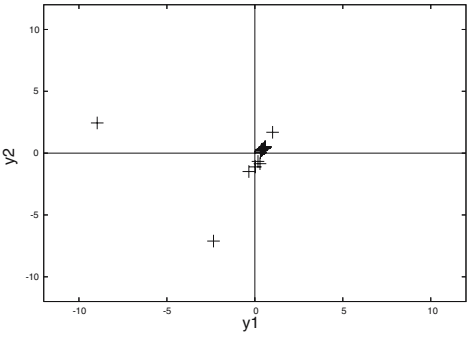

Fig. 4. Scatter plot for *ISDAS sensors* S of a dataset with $n = 30$, displaying the coefficients of the first two principal components in terms of *addresses*

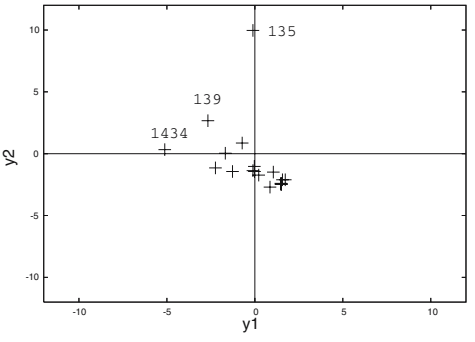

Fig. 5. Scatter plot for destination ports P displaying the coefficients of the first two principal components in terms of *ISDAS sensors*

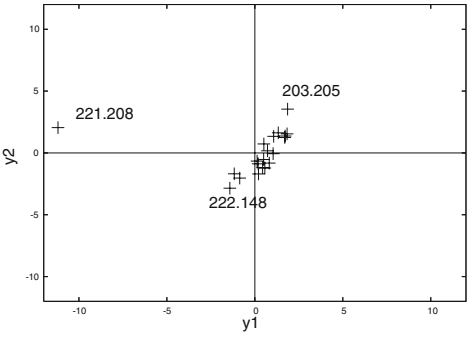

Fig. 6. Scatter plot for source addresses A displaying the coefficients of the first two principal components in terms of *ISDAS sensors*

In Fig. 5, the set of ports P is centrally distributed, with exceptions such as ports 135 and 139 at the top of the plot.

In Fig. 6, there are two clusters: a singleton {221.208} and the remainder. Any botnet-like behavior can be seen from the clustering in the plot. A scatter plot of the principal components provides a useful viewgraph by which any small change is perceptible by human operators.

4 Conclusion

We have proposed a new analysis method for the distributed observation of packets with high-dimensional attributes such as port numbers (2^{16}) and IP addresses (2^{32}). Our methods are based on the TF-IDF value mainly developed for information retrieval, and on PCA. Experimental results demonstrate that both methods correctly reduce a given high-dimension dataset to smaller dimensionality, by at least a factor of two. The principal components of port numbers, in terms of distinguishable sensors, include 445, 135, 137, 1433, 4899, 1434, 80, and ICMP, which enable any sensors to be classified. The source addresses 221.188, 222.148, 219.114, 219.165, 221.208 and 220.221 are specified as dominant on a principal component basis.

Future studies will include the stability of the basis, an accuracy evaluation for a few components, and an application of the orthogonal basis to intrusion detection.

Acknowledgments

We thank Mr. Taichi Sugiyama of Chuo University for the discussion, and the JPCERT/CC for the ISDAS distributed data.

References

1. Terada, M., Takada, S., Doi, N.: Network Worm Analysis System. IPSJ Journal 46(8), 2014–2024 (2005) (in Japanese)
2. Jung, J., Paxson, V., Berger, A.W., Balakrishnan, H.: Fast Portscan Detection Using Sequential Hypothesis Testing. In: Proc. of the 2004 IEEE Symposium on Security and Privacy (S&P 2004) (2004)
3. JPCERT/CC, ISDAS, `http://www.jpcert.or.jp/isdas`
4. Number of Hosts advertised in the DNS, Internet Domain Survey (July 2005), `http://www.isc.org/ops/reports/2005-07`
5. Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., Weaver, N.: Inside the Slammer Worm. IEEE Security & Privacy, 33–39 (July 2003)
6. Shannon, C., Moore, D.: The Spread of the Witty Worm. IEEE Security & Privacy 2(4), 46–50 (2004)
7. Changchun Zou, C., Gong, W., Towsley, D.: Code Red Worm Propagation Modeling and Analysis. In: ACM CCS 2002 (November 2002)
8. Moore, D., Shannon, C., Voelker, G., Savage, S.: Network Telescopes: Technical Report, Cooperative Association for Internet Data Analysis (CAIDA) (July 2004)

9. Kumar, A., Paxson, V., Weaver, N.: Exploiting Underlying Structure for Detailed Reconstruction of an Internet-scale Event. In: ACM Internet Measurement Conference (IMC 2005), pp. 351–364 (2005)
10. The Distributed Honeypot Project: Tools for Honeynets, http://www.lucidic.net
11. SANS Institute: Internet Storm Center, http://isc.sans.org
12. DShield.org, Distributed Intrusion Detection System, http://www.dshield.org
13. Kumar, A., Paxson, V., Weaver, N.: Exploiting Underlying Structure for Detailed Reconstruction of an Internet-scale Event. In: ACM Internet Measurement Conference (2005)
14. Ishiguro, M., Suzuki, H., Murase, I., Shinoda, Y.: Internet Threat Analysis Methods Based on Spatial and Temporal Features. IPSJ Journal 48(9), 3148–3162 (2007)
15. Dunlop, M., Gates, C., Wong, C., Wang, C.: SWorD – A Simple Worm Detection Scheme. In: Meersman, R., Tari, Z. (eds.) OTM 2007, Part II. LNCS, vol. 4804, pp. 1752–1769. Springer, Heidelberg (2007)