IPSJ Journal Vol. 51 No. 9 1659–1669 (Sep. 2010)

Regular Paper

Principal Component Analysis of Port-scans for Reduction of Distributed Sensors

Hiroaki Kikuchi[†1] and Masato Terada[†2]

There are many studies aimed at using port-scan traffic data for fast and accurate detection of rapidly spreading worms. This paper proposes two new methods for reducing the traffic data to a simplified form comprising of significant components of smaller dimensionality. (1) 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. (2) One of the significant applications of PCA is to reduce the number of sensors without losing the accuracy of estimation. Our proposed method based on PCA allows redundant sensors to be discarded and the number of packets estimated even when half of the sensors are unavailable with accuracy of less than 3% of the total number of packets. 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].

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 are also malicious software that uses particular ports to provide a "back door" to companies. The number of packets targeting the destination port used 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. For instance, Ref. 3) published the alert indicating that the number of scans destined to TCP 5168 are rapidly increasing. Port 5168 is not commonly used but should be considered carefully because it is used by a particular anti-virus service.

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 semi-passive sensor that pretends to be a vulnerable host in faked communications with intruders or worms [4]. Some sensors are *passive* in the sense that captured packets are sent to an unused IP address without any interaction. The Network Telescope [5], Internet Storm Center [6], DShield [7], and ISDAS [8] are examples of passive sensors.

There are many studies aimed at using port-scan traffic data for the fast and accurate detection of rapidly spreading worms. Kumar used the characteristics of the pseudorandom number generation algorithm used by the Witty worm to reconstruct the spread of infected hosts [9]. Ishiguro, et al. proposed Wavelet coefficients as metrics for anomaly detection [10]. Jung, et al. presented an algorithm to detect malicious packets, called Sequential Hypothesis Testing based on Threshold of Random Walk (TRW) [11]. Dunlop, et al. presented a simple statistical scheme called the Simple Worm Detection Scheme (SWorD) [12], where the number of connection attempts is tested with threshold values.

The accuracy of detection, however, depends on the assumption that *the set of sensors is independently distributed over the address space*. Since the locality of destination addresses in port-scans has been well studied [9),13),14)], it is known that when sensors are distributed too closely, they may observe packets from common source addresses with high probability. Moreover, the installation of sensors is limited to unused address blocks, and hence it is not easy to ensure truly independent sensor distribution. Since any distortion of the address distribution could cause false detection and a misdetection, independence of sensor distribution is one of the issues we should consider. Nevertheless, it is not trivial to evaluate the distribution of sensors in terms of its independence because the

†1 School of Information and Network Engineering, Tokai University
†2 Hitachi, Ltd. Hitachi Incident Response Team (HIRT)

⋆1 Parts of this work have been published in Refs. 1) and 2).

traffic data comprise ports and addresses that are correlated in high-dimensional domains.

Our 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 PCA**. Our proposal is based on an orthogonal linear transformation, which is widely used as a tool in exploratory data analysis. PCA enables the estimation of how independently 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.

(2) **Reduction of the set of sensors without sacrificing the accuracy in estimation**. Our proposed method based on PCA allows us to identify the principal components of sensors, discard the redundancy of sensors and finally estimate the number of packets when only a part of the sensors are available. This is especially useful because the unused IP addresses are assigned under the constraint of the routing and the lack of address space. Some sensors may be distributed closely and redundantly. Our experiments show that one third of the sensors is needed to estimate the number of packets with accuracy of less than 3% of the total number of packets.

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

The remainder of the paper is organized as follows. After we define some fundamental notations, the idea of PCA in our model is covered in Section 2, and experimental results are given in Section 3, where the scatter plots of port-scanning packets in the principal components are provided. Section 4 gives some concluding remarks.

2. Preliminary

2.1 Port-Address Matrices

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 a 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 ignore the effect from the dynamic behavior 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*, as part of 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 possible 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 of 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 over a time period 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

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

which are referred to as the *port vector* and the *address vector*. All packets observed by n independent sensors are characterized by the $m \times n$ 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.

Definition 3 (accuracy) An approximation of B is an $\ell \times n$ matrix of the number of packets, denoted by

$$B' = \begin{pmatrix} b'_{11} & \cdots & b'_{1n} \\ \vdots & \ddots & \vdots \\ b'_{\ell 1} & \cdots & b'_{\ell n} \end{pmatrix},$$

estimated from the subset of sensor $S' \subset S$. Similarly, an approximation of C is an $m \times n$ matrix C' estimated from $S' \subset S$. The accuracy of the approximation is evaluated by the Mean Square Error (MSE) of B', i.e.,

$$MSE(B') = \sum_{i}^{\ell} \sum_{j}^{n} (b_{ij} - b'_{ij})^2.$$

The accuracy of the approximation of C' is defined similarly for B'. We often refer to the number of sensors used for the estimate as $n = |S|$ and $n' = |S'|$. The steps to estimate the number of packets will be given in Section 2.3.

2.2 Principal Component Analysis

PCA is a well-known technique, which is used to reduce multidimensional data to a lower dimension where lower-order principal components that contributes most to its variance are kept, while higher-order components are ignored.

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 $i = 1, \ldots, n$,

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

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 $\boldsymbol{U}$ comprising the eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$, sorted in order of decreasing eigenvalue $\lambda_1 > \cdots > \lambda_m$, and the *covariance matrix* which is defined as

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

From the 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

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

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, while the second component y_{i2} contributes the second most significant effect on the 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 transformation not only to the matrix $\boldsymbol{C}$ defined over the port number and the sensors ($m \times n$) but also to the matrix $\boldsymbol{B}$ of the address spaces and the sensors ($\ell \times n$), 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 $\boldsymbol{B}$.

The matrix $\boldsymbol{C}$ is often too large to apply PCA due to the large computational

power needed for the size of the matrix. In order to make the PCA possible for the large matrix, we apply a technique used in information retrieval and data mining, called *TF-IDF weighting*. The TF-IDF weight gives the degree of importance of a word in a collection of documents. TF-IDF is properly defined in Appendix A.1.

2.3 Estimation of Port-Scan Packets

One of the significant applications of PCA is to reduce the set of sensors without losing the accuracy of estimation. This is especially useful when a limited number of sensors are available over the set of IP addresses space because of the lack of unused IP addresses and the constrained assignment of addresses coming from the routing requirements. Note that the distribution of sensors is not ideally uniform and some sensors are distributed closely and redundantly in the reduced coordinate spaces. The redundancy of sensors can be discarded by using the orthogonal property of principle components basis as follows.

Recall that the $m \times n$ port-sensor metrics $\boldsymbol{C}$

$$\boldsymbol{C}' = \boldsymbol{C} - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \begin{pmatrix} \overline{c_1} \\ \vdots \\ \overline{c_m} \end{pmatrix}^{\top}$$

is estimated by the principal component basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and the PCA coefficient matrix is

$$Y = (\boldsymbol{y}_1 \cdots \boldsymbol{y}_n) = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mn} \end{pmatrix} = \boldsymbol{C}' \cdot \boldsymbol{U}.$$

The first order approximation of $\boldsymbol{C}'$ is given by

$$\boldsymbol{C}' \approx \boldsymbol{y}_1 \cdot \boldsymbol{u}_1$$

where $\boldsymbol{u}_1$ is the first eigenvector $(1 \times n)$, shown by Table 3. In the same way, we have the k-th order approximation of $\boldsymbol{C}'$ as

$$\boldsymbol{C}' \approx \sum_{i=1}^{k} \boldsymbol{y}_i \cdot \boldsymbol{u}_i.$$

2.4 Estimation from Limited Sensors

The PCA transformation provides us an efficient way of reducing the number of redundant sensors. Since the principal component basis are constant vectors representing the correlation over a set of sensors, we can estimate the number of packets given a fraction of the sensor set.

Letting n' be $n' < n$, we replace from the n'-th and the n-th rows by 0 vectors, resulting in the partial matrix $(m \times n)$

$$\boldsymbol{C}_{(n')} = \begin{pmatrix} c_{11} & \cdots & c_{1n'} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & 0 & \cdots & 0 \\ c_{m1} & \cdots & c_{mn'} & 0 & \cdots & 0 \end{pmatrix},$$

and then estimate the number of packets from the remaining n' sensors as,

$$\boldsymbol{C}'_{(n')} \approx \sum_{i=1}^{k} \boldsymbol{y}_{(n')i} \cdot \boldsymbol{u}_i.$$

where

$$Y_{(n')} = (\boldsymbol{y}_{(n')1} \cdots \boldsymbol{y}_{(n')n}) = \boldsymbol{C}'_{(n')} \cdot \boldsymbol{U}.$$

3. Analysis

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

3.1 Experimental Data

3.1.1 ISDAS Distributed Sensors

ISDAS is a distributed set of sensors [8], 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, where we denote by $h(x)$ a unique IP address observed by sensor x. 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.

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

3.2 Principal Component Basis

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.

Table 2 shows the experimental results for the first two orthogonal vectors of the principal component basis $\boldsymbol{u}_1(C), \boldsymbol{u}_2(C), \ldots$ for the ports-and-sensors matrix $\boldsymbol{C}$ and the 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 attacks*. The second principal component has about half the effect of the projected values because eigenvalue λ_1 is almost double in value compared to 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 3**, 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 $\boldsymbol{u}_2(B)$.

3.3 Major and Minor Port Numbers

There are many backdoors on P2P Botnet and a Trojan code that uses minor port numbers other than the major ones such as 445, 135, 137, 1434, 80, and ICMP. Hence, the proposed PCA-based method may have a risk to fail to detect small changes happen on minor ports that the malware often uses. In order to minimize the risk of false detection, we chose significant ports in terms of a *TF-IDF* measure mentioned in Appendix A.1. The significance of a port is evaluated by means of the *TF-IDF* measure so that a high weight is attributed not only to frequently observed port but also to minor ports that are not too frequently observed. For instance, **Table 4** shows *TF-IDF* measures for port numbers chosen as the principal component basis in Table 2. Minor ports have small document frequencies (DFs) which increase the *TF-IDF* measure and hence ports are likely to be chosen. Indeed, it is clear that both major and minor port numbers are used for the estimate the number of packets.

The use of the *TF-IDF* measure to chose port numbers before PCA is involved has its pros and cons. The advantage of the *TF-IDF* measure when used with PCA is to allow a scalable analysis in terms of large domains, e.g., full address space and port number, with reduced computational overhead. The estimation takes into consideration of both major and minor port numbers. On the other hand, unknown minor ports newly used after *TF-IDF* can not be taken into account here. We should note this disadvantage of combining *TF-IDF* with

Table 2 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	$\boldsymbol{u}_1(C)$	$\boldsymbol{u}_2(C)$	a_k	$\boldsymbol{u}_1(B)$	$\boldsymbol{u}_2(B)$
445	**−0.37**	0.01	221.188	**−0.54**	0.20
135	**−0.36**	0.01	222.148	**−0.54**	0.20
137	−0.34	−0.07	219.114	−0.53	0.20
1433	−0.33	0.17	219.165	−0.28	**−0.52**
4899	−0.30	0.27	221.208	−0.17	−0.41
1434	−0.30	0.16	220.221	−0.14	**−0.59**
1026	−0.28	−0.27	58.93	−0.01	−0.20
1025	−0.28	−0.01	222.13	0.00	−0.09
1027	−0.25	−0.28	222.159	0.01	−0.06
22	−0.23	0.08	61.199	0.03	0.03
32656	−0.13	−0.27	219.111	0.03	0.02
12592	−0.13	−0.27	220.109	0.03	0.03
139	−0.10	0.18	61.205	0.03	0.03
23310	−0.09	−0.03	221.16	0.03	0.03
80	−0.02	**0.45**	61.252	0.03	0.04
ICMP	−0.02	**0.44**	203.174	0.03	0.04
113	0.00	0.25	61.193	0.03	0.04
4795	0.00	0.25	203.205	0.04	0.04
631	0.05	−0.04	219.2	0.06	0.14
1352	0.09	−0.08	218.255	0.06	0.14
eigenvalue λ_i	6.19	2.49	eigenvalue λ_i	3.16	2.29

Table 3 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_i	$\boldsymbol{u}_1(\boldsymbol{B}^\top)$	$\boldsymbol{u}_2(\boldsymbol{B}^\top)$
s_7	−0.04	0.34	s_{12}	**−0.34**	0.16
s_{20}	−0.03	0.30	s_{18}	**−0.34**	0.18
s_8	−0.03	**0.42**	s_6	**−0.34**	0.18
s_{22}	−0.01	**0.42**	s_{20}	**−0.34**	0.02
s_{26}	−0.01	0.25	s_{22}	**−0.34**	0.18
s_{30}	0.03	−0.12	s_{13}	−0.32	0.21
s_{28}	0.05	−0.19	s_{17}	−0.32	0.01
s_{12}	0.06	**0.37**	s_{29}	−0.28	−0.20
s_{15}	0.06	−0.16	s_{28}	−0.21	**−0.35**
s_{29}	0.07	−0.22	s_{27}	−0.20	−0.11
s_{25}	0.17	−0.01	s_4	−0.17	−0.27
s_{23}	0.18	−0.08	s_{23}	−0.10	**−0.33**
s_6	0.18	0.24	s_1	−0.05	−0.30
s_{24}	0.19	0.04	s_3	−0.05	−0.21
s_5	0.21	0.02	s_5	−0.03	−0.03
s_4	0.22	0.08	s_{11}	−0.01	0.03
s_{17}	0.22	−0.12	s_{10}	0.00	−0.15
s_{16}	0.22	−0.09	s_{14}	0.01	−0.08
s_{21}	0.22	−0.02	s_{26}	0.01	−0.05
s_{27}	0.23	−0.06	s_9	0.01	0.07
s_{13}	0.23	0.03	s_2	0.01	0.06
s_{14}	**0.24**	−0.02	s_{15}	0.02	−0.11
s_{18}	**0.24**	0.10	s_{30}	0.02	−0.07
s_{11}	**0.24**	0.07	s_{16}	0.03	−0.00
s_{19}	**0.24**	0.01	s_{19}	0.03	0.12
s_3	**0.24**	0.05	s_{24}	0.04	0.15
s_1	**0.24**	0.03	s_8	0.04	0.13
s_2	**0.24**	0.01	s_{25}	0.04	**0.32**
s_{10}	**0.24**	−0.02	s_{21}	0.06	**0.31**
s_9	**0.24**	0.03	s_7	0.07	0.18
eigenvalue λ_i	16.64	3.73	eigenvalue λ_i	7.81	2.66

Table 4 *TF-IDF* measures in major and minor ports.

	port	*TF*	*DF*	*TF-IDF*
major	445	762283	43	17661
	135	1011078	45	22447
	137	49600	43	1149
minor	32656	4168	3	333
	12592	2774	1	286
	23310	23687	2	2095

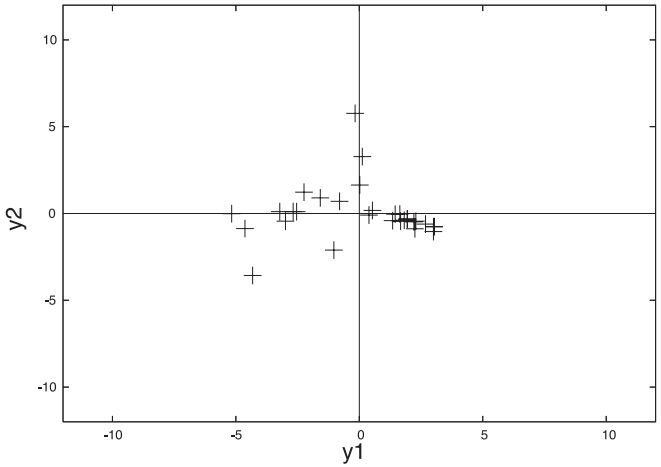

Fig. 1 Scatter plot for *ISDAS sensors* S of a dataset with $n = 30$, displaying the coefficients of the first two principal components in terms of *ports*.

PCA. To avoid this risk, we also use the full PCA without any *TF-IDF* filtering at a large computational cost. Consequently, there is a tradeoff between accuracy and efficiency.

3.4 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 an independent set of sensors is distributed over the reduced coordinate system. If we wish to identify the abnormal behavior of source addresses, applying PCA 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 standard behaviors.

For these purposes, we show the experimental results of ISDAS observation data, in **Figs. 1** and **2**, corresponding to matrices $\boldsymbol{C}$, and $\boldsymbol{B}$, respectively.

The set of ISDAS sensors is independently distributed in Fig. 1, but the distribution is skewed by some irregular sensors in Fig. 2, 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.

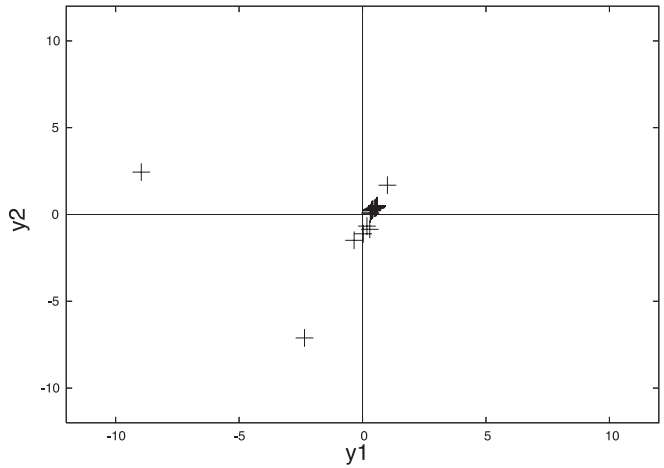

Fig. 2 Scatter plot for *ISDAS sensors* S of a dataset with $n = 30$, displaying the coefficients of the first two principal components in terms of *addresses*.

3.5 PCA Evaluation

Figure 3 demonstrates the estimated number of packets observed by each sensor. The original distribution of packets are well approximated by a 2nd order approximation using just $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ out of $n = 30$ orthogonal vectors. We see that even the first one alone is a good approximation of the number of packets, except for sensor s_{26}, which can be fixed by the 2nd order approximation.

In order to visually understand the accuracy, we show the approximation with respect to port numbers in **Fig. 4**. The failure of estimation at port 1026 and ICMP in the first order approximation are brought by the fact that the first principal basis is independent from these port numbers. The difference between the estimate and the original number of packets is reduced as the order of approximation increases.

Our proposed method applies various kinds of statistical values other than port numbers. **Figure 5** demonstrates the approximation of number of packets for the attacker's source IP address. The accuracy of the 1st estimation at address space 222.148 improves after the 4th approximation in this experiment. Since the distribution of the number of packets are distorted in comparison to the port

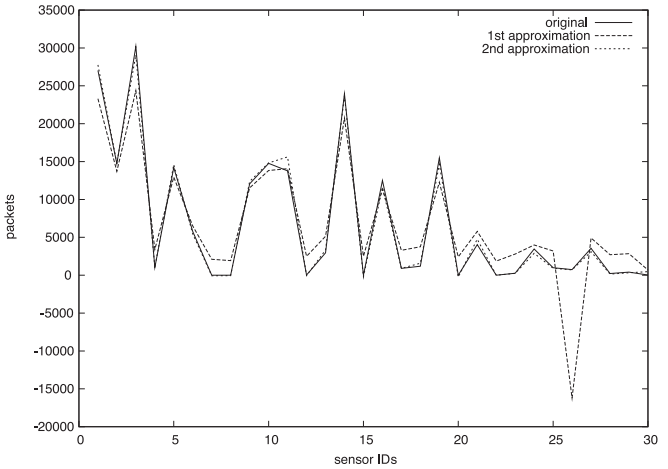

Fig. 3 Approximation of number of packets observed by each sensor ID.

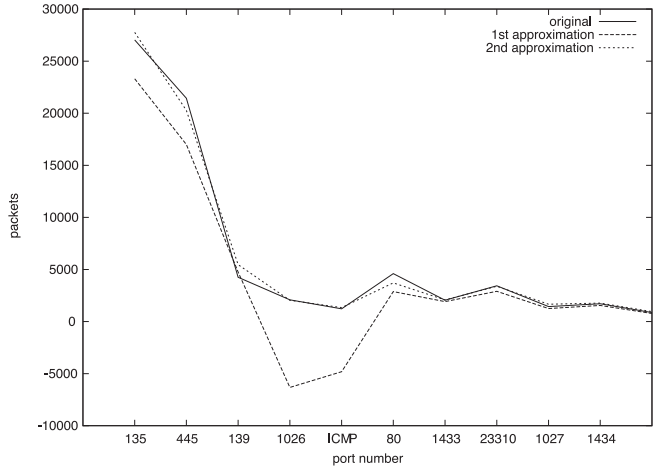

Fig. 4 Approximation of number of packets with respect to port numbers.

number, more degrees of approximation is necessary compared to the case with the port number.

The effect from the accuracy of improvement is shown in **Fig. 6**, where the

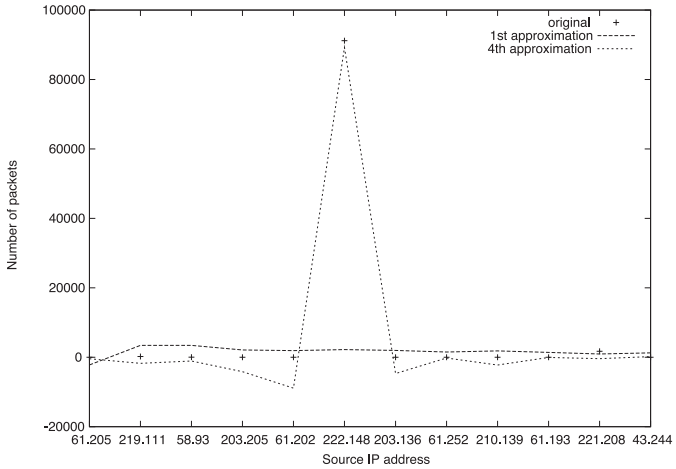

Fig. 5 Approximation of number of packets with respect to source IP address.

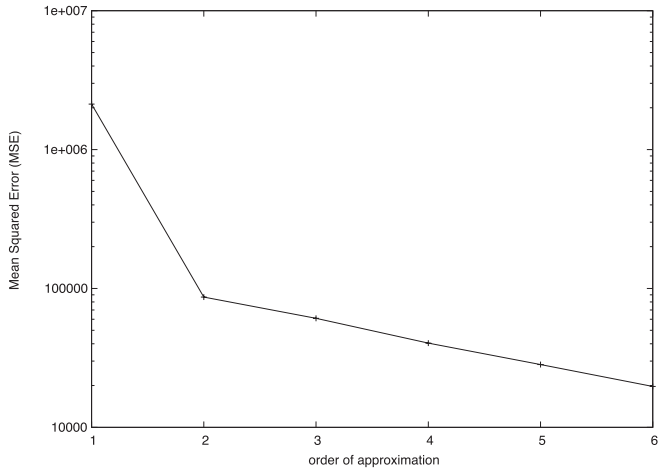

Fig. 6 Mean squared error (MSE) of estimated number of packets for the order of approximation.

overall accuracy for the order of approximation is given by the Mean Squared Error (MSE). Note that the vertical axis uses a logarithmic scale. The size of the error is relatively small for the total number of packets, shown in **Table 5**, where

Table 5 Mean squared error of estimated number of packets.

order	MSE	$\sqrt{\text{MSE}}$	[%]
1	2124658.1	1457.6	21.41
2	86659.6	294.4	4.32
3	60916.4	246.8	3.63
4	40437.4	201.1	2.95
5	28317.8	168.3	2.47
6	19707.2	140.4	2.06

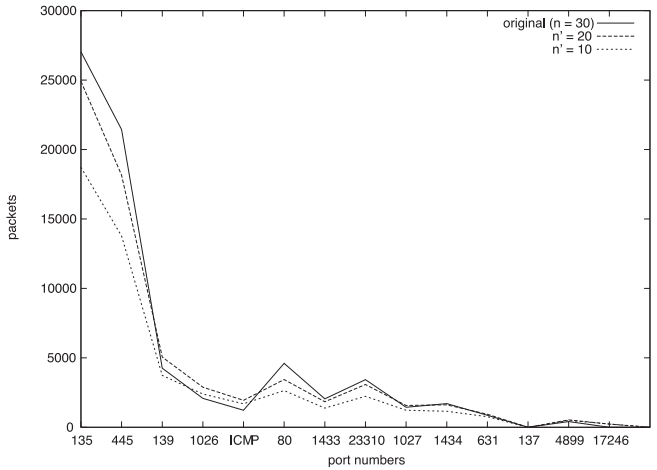

Fig. 7 Estimation of number of packets from several subsets of sensors.

the mean difference of packet counts between the original and the estimation is $\sqrt{MSE}$.

3.6 Reduction of Sensors

The accuracy of the estimate used by the limited number of sensors is surprisingly high. We illustrate this observation by showing the packet counts distribution over port numbers in **Fig. 7**, where estimates with the first $n' = 20$ and 10 sensors out of 30 sensors appear to approximate the original distribution ($n = 30$) well. According to the figure, the distribution goes flat when the number of sensors n' is reduced, even with one third of 30 the sensors a rough approximation of the original data is possible. The size of the error can be normalized by taking

Table 6 Mean squared error of estimated number of packets from subset of sensors.

n'	MSE	$\sqrt{\text{MSE}}$	[%]
1	3797064.7	1948.6	3
5	2860140.4	1691.2	2
10	2644375.6	1626.2	2
15	2267337.7	1505.8	2
20	2149148.9	1466.0	2
30	0	0	0

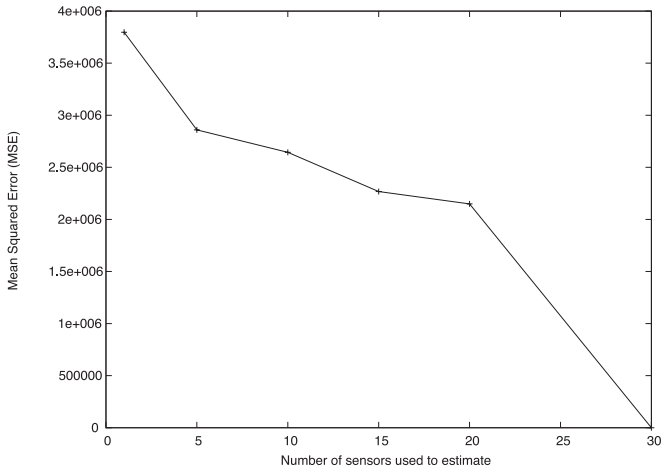

Fig. 8 Mean squared error (MSE) of estimated number of packets with respect to the number of sensors.

the squire root, as shown in **Table 6**. The error is negligibly small even when just one sensor is available.

The results of the experiment are shown in **Fig. 8**, where the accuracy, namely the inverse of the MSE, is shown to improve as the number of sensors increases.

In this experiment, we just choose the first n' sensors arbitrarily, hence it is possible to improve the accuracy if we chose the top n' sensor in terms of the principal component basis. For example, Table 3 indicates that s_{14}, s_{18}, and s_{11} are more significant than s_7, s_{20}, and s_8.

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 PCA. Experimental results demonstrate that both methods correctly reduce a given high-dimension dataset to a 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 in terms of the principal component basis.

Our proposed method based on PCA allows not only the identification the principal components of sensors, but also allows redundant sensors to be discarded so that finally, the number of packets can be estimated when only a portion of sensors are available. Our experiments show that the accuracy of the estimation used by a more limited set of sensors is surprisingly high. A reduction of a third of the sensors successfully provides an estimate of the number of packets with accuracy of less than 3% of the total number of packets. The advantage of our proposed method is to allow us to grasp any change of statistical values by means of fewer principal components without suffering from too many involved factors in the observation matrices.

Acknowledgments We thank Mr. Tomohiro Kobori and Mr. Naoya Fukuno for the discussion, and the JPCERT/CC for the ISDAS distributed data.

References

1) Kikuchi, H., Fukuno, N., Terada, M. and Doi, N.: Principal Components of Port-Address Matrices in Port-Scan Analysis, *On the Move to Meaningful Internet Systems: OTM 2008, LNCS 5332*, pp.956–968, Springer (2008).
2) Kikuchi, H. and Terada, M.: Orthogonal Expansion of Port-Scan – Estimation from Limited Sensors, *2009 Joint Workshop on Information Security* (*JWIS 2009*), 5A-2, pp.1–14 (2009).
3) JPCERT/CC: Increased activity targeting TCP port 5168, JPCERT-AT-2007-0019 (2007). http://www.jpcert.or.jp/at/2007/at070019.txt
4) The Distributed Honeypot Project: Tools for Honeynets. http://www.lucidic.net
5) Moore, D., Shannon, C., Voelker, G. and Savage, S.: Network Telescopes: Technical

Report, *Cooperative Association for Internet Data Analysis* (*CAIDA*) (July 2004).
6) SANS Institute: Internet Storm Center. http://isc.sans.org
7) DShield.org: Distributed Intrusion Detection System. http://www.dshield.org
8) JPCERT/CC: ISDAS. http://www.jpcert.or.jp/isdas
9) Kumar, A., Paxson, V. and Weaver, N.: Exploiting Underlying Structure for Detailed Reconstruction of an Internet-scale Event, *ACM Internet Measurement Conference* (*IMC'05*), pp.351–364 (2005).
10) Ishiguro, M., Suzuki, H., Murase, I. and Shinoda, Y.: Internet Threat Analysis Methods Based on Spatial and Temporal Features, *IPSJ Journal*, Vol.48, No.9, pp.3148–3162 (2007).
11) Jung, J., Paxson, V., Berger, A.W. and Balakrishnan, H.: Fast Portscan Detection Using Sequential Hypothesis Testing, *Proc. 2004 IEEE Symposium on Security and Privacy*, (*S&P'04*) (2004).
12) Dunlop, M., Gates, C., Wong, C. and Wang, C.: SWorD – A Simple Worm Detection Scheme, *OTM Confederated International Conferences: Information Security* (*IS 2007*), *LNCS 4804*, pp.1752–1769 (2007).
13) Terada, M., Takada, S. and Doi, N.: Network Worm Analysis System, *IPSJ Journal*, Vol.46, No.8, pp.2014–2024 (2005) (in Japanese).
14) Ishiguro, M., et al.: Feature Analysis of Illegitimate Packets Monitored on the Internet, *IPSJ Computer Security Symposium* (*CSS 2005*) (2005).

Appendix

A.1 Reduced Matrix via TF-IDF Values

TF-IDF weighting assigns a 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 is 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

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

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 being unable to distinguish between sensors. By taking the logarithm of the inverse of the document frequency, we obtain the *TF-IDF* for a given port p_j as

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

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

Similarly for 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

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

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, with the aim of filtering 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.

(Received December 1, 2009)
(Accepted June 3, 2010)
(Original version of this article can be found in the Journal of Information Processing Vol.18, pp.190–200.)

Hiroaki Kikuchi was born in Japan. He received his B.E., M.E. and Ph.D. degrees from Meiji University in 1988, 1990 and 1994. After working in Fujitsu Laboratories Ltd. from 1990 through 1993, he joined Tokai University in 1994. He is currently a Professor at the Department of Communication and Network Engineering, School of Information and Telecommunication Engineering, Tokai University. He was a Visiting Researcher at the School of Computer Science, Carnegie Mellon University in 1997. His main research interests are fuzzy logic, cryptographical protocol, and network security. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), the Japan Society for Fuzzy Theory and Systems (SOFT), IEEE and ACM. He is a fellow of the Information Processing Society of Japan (IPSJ).

Masato Terada was born in Japan. He received his M.E. in Information and Image Sciences from Chiba University, Japan, in 1986. He joined Hitachi, Ltd. in 1986. He is currently the Chief Researcher at the Security Systems Research Dept., Systems Development Lab., Hitachi. From 2002, he studied at the Graduate School of Science and Technology, Keio University, receiving the Ph.D. in 2006. Since 2004, he has been with the Hitachi Incident Response Team. Also, he is a Visiting Researcher at the Security Center, Information - Technology Promotion Agency, Japan (ipa.go.jp), JVN associate staff at JPCERT/CC (jpcert.or.jp) and a Visiting Researcher at the Research and Development Initiative Chuo University.