IPSJ Journal Vol. 49 No. 9 3136–3145 (Sep. 2008)

Regular Paper

Estimation of Increase of Scanners Based on ISDAS Distributed Sensors

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

Given independent multiple access logs, we develop a mathematical model to identify the number of malicious hosts in the current Internet. In our model, the number of malicious hosts is formalized as a function taking two inputs, namely the duration of observation and the number of sensors. Assuming that malicious hosts with statically assigned global addresses perform random port scans to independent sensors uniformly distributed over the address space, our model gives the asymptotic number of malicious source addresses in two ways. Firstly, it gives the cumulative number of unique source addresses in terms of the duration of observation. Secondly, it estimates the cumulative number of unique source addresses in terms of the number of sensors. To evaluate the proposed method, we apply the mathematical model to actual data packets observed by ISDAS distributed sensors over a one-year duration from September 2004, and check the accuracy of identification of the number of malicious hosts[⋆1].

1. Introduction

Malicious hosts routinely perform port scans of IP addresses to find vulnerable hosts to compromise. According to Ref. 1), the *Sasser* worm performs scans to fully randomly determined destinations with a probability of 0.52, and to partially random destinations that have the highest two octets identical and one octet, with probabilities of 0.25 and 0.23, respectively. Many major worms have well-engineered algorithms for performing port scans and for choosing random destinations, e.g., Slammer [6], Witty [7], and Code Red [8]. In the Internet, the mixture of these complicated behaviors is a significant source of complexity, which prevents prediction of the exact impact of worms and distributed attacks, even though new malicious codes now appear daily.

One of the effective countermeasures against the dynamic behavior of malicious hosts is the Network *telescope* [9], which records packets sent to unused blocks of the address space, the so-called "*dark net*", and uses the logs of worm activity to infer aggregate properties, such as the worm's infection rate, the total scanning rate, and the evolution of these quantities over time. Kumar, et al. carefully analyze the telescope observations of the Witty worm, and succeed in revealing information about the host, such as access bandwidth, uptime, and the number of physical drivers. They finally identify patient Zero, the host that the worm's author used to release the worm.

However, the Network telescope requires large unused address blocks. The greater the block size, the more accurate is the estimation; but, at the same time, malicious hosts have a greater chance of discovering the telescope. Instead, we use *small but orthogonally distributed* sensors with unused IP addresses and combine the logs to calculate the behavior of the target set of malicious hosts. Our estimation is based on a mathematical model of the cumulative distribution of unique hosts observed by the sensors with respect to the number of sensors and the duration of observation. A mathematical model allows us to identify the population of the malicious hosts, and the frequency that port scans are performed, from arbitrary given logs. In addition, fitting our model to any subset of logs provides useful characteristics of the subset, which can be seen as a degree of risk. For example, the differences of port scans for destination ports, source/destination addresses, duration of observation, and time of the year can be clarified from the result of fitting. Even if an attempt of fitting fails, it immediately implies that there is an ordinary event happening in the Internet scale.

Hence, our model is useful for many applications including

- a prediction of malicious scanning behaviors,
- a risk assessment of target subnet,
- a detection of significant changes in port scanning behavior, and
- a detection of largely coordinated attempts for DoS.

In this paper, we use the Internet Scan Data Acquisition System (ISDAS)

†1 School of Information Technology, Tokai University
†2 Hitachi Incident Response Team (HIRT), Hitachi, Ltd.
†3 Facility of Science and Engineering, Chuo University
⋆1 The original version of this paper was presented at the 2nd International Symposium on Information Security (IS'07).

distributed sensor [4], under the operation of JPCERT/CC, to estimate the scale of current malicious events and their performance. The idea was first presented in Ref. 2), using a very limited number of sensors for just 18 weeks, and therefore the accuracy was not fully evaluated.

Our Contribution

There are some significant aspects to our contributions.

- We present a new mathematical model for a cumulative unique host's number in terms of a number of potential scanners, a duration for observation and a frequency of port scanning attempts. As in the case of duration, a model in terms of a number of sensors is also presented.
- Based on the actual packets log data observed by the ISDAS distributed sensors for several years, we evaluate our model to clarify the performance and the accuracy in several environments.
- Our experiment shows the estimated number of malicious hosts and the estimated performance of port scans at the time of the observation.

The rest of our paper is organized as follows. We first provide some fundamental definitions and show the characteristics of ISDAS data. Then, we present a mathematical model for cumulative unique host's numbers. To evaluate the proposed model, we use the actual logs of ISDAS, we evaluate the validity of the model and examine each of the properties with respect to an actual set of malicious hosts in Internet-scale events.

2. Model of Cumulative Unique Source Addresses

2.1 Fundamental Definitions

We give the fundamental definitions necessary to discuss the characteristics of worms.

Definition 2.1 (**Scanner and Sensor**) A *scanner* is a host which performs port scan to other hosts looking for the target to be attacked. A *sensor* is a host that passively observes all packets sent from scanners.

Typically, a scanner is a host that has some vulnerability, and thereby is controlled by a 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 assignment provided via Dynamic Host Control Protocol (DHCP) or Network Address Translation (NAT).

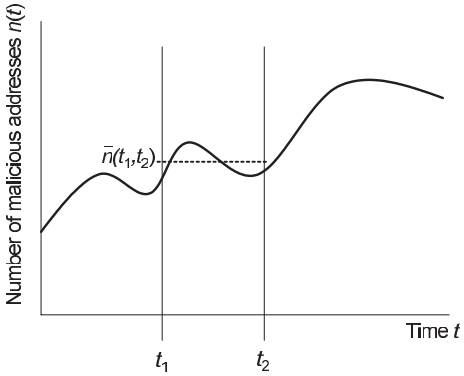

Fig. 1 Number of malicious hosts over time.

Definition 2.2 (**Population**) Let n_0 be the number of active global IP addresses. Consider the set of active addresses in the whole 32-bit address space[⋆1]. Let n and x be the numbers of scanners and sensors, respectively. Clearly, $n, x \ll n_0$.

The number of scanners varies hourly. For example, it increases with a new virus infection and decreases with the extermination of the virus. However, over a long duration, e.g., monthly, a stationary population of scanners can be assumed. **Figure 1** illustrates the stationary behavior of the number of scanners, where $\overline{n}(t_1, t_2)$ indicates the average population during t_1 to t_2, in satisfying

$$\overline{n}(t_1, t_2) \doteq n(t_1) \doteq n(t_2)$$

over the duration.

The frequency of scans depends on the scanners. In our analysis, we focus on the increase in distinct source addresses observed by sensors, which are defined as *unique hosts*.

⋆1 Because of unassigned address blocks and private addresses, the number of active addresses is smaller than 2^{32}. According to Ref. 5), which estimated host counts by pinging a sample of all hosts, the total number of active addresses in July 2005 was reported as being 353, 284, 184.

Definition 2.3 (**unique hosts**) Let $h(x,t)$ be the cumulative number of unique hosts observed at x independent sensors within the time interval $[0,t]$, where t is a monthly, weekly or hourly unit of time.

Putting distributed log files together provides useful knowledge about the set of scanners. For example, from the log files we obtain the average number of scans observed by a sensor per hour, a list of frequently observed scanners, some common patterns among port scans, the correlation among sensors, the relationship between scans and classes of sensors, and scanning variations per hour, week, or month, etc. Formally, our objective is to identify the total number of scanners, n, given the unique hosts $h(x,t)$ observed by distributed sensors.

Definition 2.4 (**static address**) A scanner is *static* if it is assigned a static global address and does not use a spoofed address.

In practice, a host may have multiple addresses assigned by a DHCP server. Alternatively, one proxy address can be shared by multiple hosts inside a firewall under NAT. We begin with the simplest version of our model with static address and will generalize the condition toward the reality.

Definition 2.5 (**uniform destination**) A scan is *uniform* if the destination address is randomly determined, and is uniformly distributed over the set of active sensors.

Note that the actual distribution of destination addresses is not uniform over the address space because there are some worms that perform local port scans. However, by filtering the local scans we can minimize the effect of local scans and approximate the destinations of scans as a uniform distribution over the set of sensors.

If a scan is uniform, the probability of a certain sensor being chosen is $p_0 = 1/n_0$. Since there are n scanners, which can be considered as Bernoulli trials, we have an expected value for the number of scans as the mean of a binominal distribution, i.e.,

$$E[h(1,t)] = np_0 = n/n_0.$$

Definition 2.6 (**stationariness**) A set of scanners is *stationary* if there exists a duration T for which a population of scanners reaches a stationary state.

Over the long duration, an increase of population of scanners asymptotically equals to the decrease at some point.

Definition 2.7 (**average scans**) A scanner performs c scans in a time interval $[0,t]$, on average.

Ideally, we formalize a performance of scans with a single parameter of the mean of scans though the performance depends on the kind of worms and on the performance of the infected hosts. From a macroscopic viewpoint, the number of scans can be approximated by c. To simplify, we represent the average number of unique hosts observed by a sensor as

$$a = c\frac{n}{n_0}. \tag{1}$$

Given multiple observations, as the number of sensors increases, the unique hosts number increases as well. The number of unique hosts is, however, not linear with the number of sensors x.

There may be a small number of scanners observed by both sensors. Therefore, we have

$$h(2,t) \leq 2h(1,t).$$

In general, the difference $\Delta h(x,t) = h(x,t) - h(x-1,t)$ decreases as x increases, and asymptotically disappears. In addition, we note that Δh depends on the total number of scanners n, because n is the dominating factor in the probability of *collision*, i.e., two sensors observing a common scanner. Therefore, we can estimate the total number of scanners from the reduction in the increase of unique hosts with respect to the number of sensors.

Similarly, we have a relationship between the number of unique hosts and the duration of observation, namely

$$h(x,2) \leq 2h(x,1).$$

This analogy between the number of sensors x and the duration of observation t provides dual estimation paths. If the two estimates from the increase of x and t are close, we can be highly confident of our estimate of n.

2.2 Estimation Model of n using Duration t

First, we try to estimate the number of scanners by varying the duration of observation. In the next section, we will estimate it in an alternative way.

From Eq. (1), we begin with $h(1,1) = a$, namely an increase by a in every time interval. The probability that a new address has already been observed is $p = h(1,1)/n$, so we can regard a observations as a Bernoulli trials with

probability p. It follows that $ap = a\,h(1,1)/n = a^2/n$ addresses are duplicates, on average. More precisely, the probability that k addresses have been observed in a newly observed addresses is given by the binomial distribution specified by the probability density function

$$P(k,a) = \left(\begin{array}{c} a \\ k \end{array} \right) p^k (1-p)^{a-k}.$$

Taking the mean of k, we have

$h(1,2) = h(1,1) + a - a\,h(1,1)/n = 2a - a^2/n$,

which follows

$h(x,t+1) = h(x,t)(1-a/n) + a$.

Taking the difference $\Delta h(x,t) = h(x,t+1) - h(x,t)$ gives the differential equation of the unique host function $h()$

$$\frac{dh}{dt} = -\frac{a}{n} h(x,t) + a, \tag{2}$$

which follows the general form

$$h(x,t) = C \cdot e^{-\frac{a}{n}t} + e^{-\frac{a}{n}t} \int e^{\frac{a}{n}t} \cdot a \; dt = C\, e^{-\frac{a}{n}t} + n.$$

With the initial condition $h(0,0) = C\,e^0 + n = 0$, we have $C = -n$. Therefore, we have the following theorem.

Theorem 2.1 (**Unique hosts with respect to duration** t) If a set of scanners is static, uniform and stationary, then a cumulative unique hosts number is

$$h(x,t) = n(1 - e^{-\frac{a}{n}t}), \tag{3}$$

where n is the total number of potential scanners, x is a number of sensors and a is the average number of unique hosts observed by a single sensor in the time interval.

2.3 Estimation Model of n using Number of Sensors x

Recall the analogy between the duration of observation t and the number of sensors x. By replacing t with x in Eq. (3), we have the following.

Theorem 2.2 (**Unique hosts with respect to sensors** x) If a set of scanners is static, uniform and stationary, then a cumulative unique hosts number observed by x sensors is

$$h(x,t) = n(1 - e^{-\frac{a}{n}x}), \tag{4}$$

where a is the average increase in unique hosts for one sensor in certain duration.

These dual functions will be investigated by experiment.

Note that the variation between sensors is greater than that for durations. Although we have assumed uniform scans, an actual port scan is not performed globally over the address space. There are some worms and viruses that scan multiple destinations by incrementing the fourth octet of the IP address. Therefore, we should carefully choose the location of hosts for sensing, and, to minimize the difference between sensors, we should take the average for unique hosts from all possible combinations of x sensors. For example, if we have three sensors, s_1, s_2 and s_3, then $h(2,t)$ is defined as the average for pairs $(s_1,s_2),(s_1,s_3),(s_2,s_3)$.

3. Experiments

This section evaluates our model using experimental log data observed in the Internet to estimate the number of scanners in the Internet.

3.1 ISDAS Observation Data

The ISDAS comprises multiple passive sensors distributed among multiple Internet service providers in Japan [4]. The ISDAS provides daily statistics of packets observed by the distributed sensors for each of the major ports, 13, 80, 135, 139, 445, and 1026.

In our analysis, we have a set of orthogonal log data, observed by 12 independent sensors chosen from more than 40 sensors, from September 1, 2004 through September 30, 2005.

3.2 Methods of Evaluation

Let $h(S,T,P)$ be the number of unique source addresses observed using a set of sensors $S = \{s_1, \ldots, s_{12}\}$, for duration T, and for destination port P. For example, $h(\{s_9\}, [2004/5 - 2004/7], 135)$ denotes the cumulative unique source addresses observed by a single sensor s_9, from May 2004 for three months, for destination port 135.

We specify parameters n and a in our model by fitting the model to the actual observed log data in the following ways.

(1) Identification of the total number of malicious hosts, estimated for several observation durations $t = 1, \ldots, 360$ (days).
For each sensor s in S, we perform a fitting of Eq. (3) to the observed data $h(s, t, p)$, where the destination port is one of 135, 139, or 445. The optimized parameters for our model give the estimated number of total hosts n.

(2) Identification of the total number of malicious hosts, estimated from the number of distributed sensors, x.
For $x = 1, \ldots, 12$, we perform a fitting of Eq. (4) to $h(S_x, T, p)$, where $|S_x| = x$ and T is a constant. Because the number of packets varies considerably among sensors, we choose the two extreme sets S_{x*} and S_x^* and take the average of the minimal and maximal sets for each x.

We examine the difference in estimates for several observation durations and check the stability of the fitting accuracy for several observation intervals. We show the correlation between the number of unique source addresses and the number of packets sent to a given sensor.

The strategy for performing port scans to random addresses depends on the hosts and the malicious codes. We investigate the uniformity of addresses scanned by showing the statistics for the number of sensors observing a given source address.

3.3 Estimation of Scanners Based on Duration of Observation

Table 1 shows the total number of packets and the unique source addresses $h(1, T_1, 445)$, observed during T_1 (September 1, 2004 through September 30, 2005), for each of sensors $s_1, \ldots, s_{12}$. $\Delta h(s, T_1, 445)$ denotes the average increase in unique source addresses per day. Sensor s_9 observed the least number of packets among all sensors, which is about 1/25 of that for sensor s_1.

In **Fig. 2**, we illustrate the daily average increase in unique source addresses ($\Delta h(s, T1, 445)$), which decreases during the period of observation. In other words, the set of unique source addresses becomes saturated, and asymptotically reaches a fixed size.

With reference to Eq. (3), we visually demonstrate our fitting accuracy in **Fig. 3**, where sensors $s = s_1, s_3, s_{11}$ are used in the estimation. For other sensors, **Table 2** shows the estimated total number of malicious hosts and the scanning ratio, in conjunction with the *asymptotic standard error*$^{\star 1}$ for each sensor. We also show the average number of port scanning packets per second, c, in the rightmost column of the table. The average number of packets is $\bar{c} = 16.75$, which seems to imply automatic generation.

Table 1 Statistics of packets for sensors.

sensor ID	total packets	unique host $h(x)$	$\Delta h(x)$ [per day]
s_1	268024	97102	245.8
s_2	153310	63198	160.0
s_3	154126	60755	153.8
s_4	137848	40315	102.1
s_5	168191	62881	159.2
s_6	173566	47809	121.0
s_7	17167	10066	25.5
s_8	164078	54865	138.9
s_9	10667	9046	22.9
s_{10}	170417	24394	61.8
s_{11}	30898	13200	33.4
s_{12}	143725	53716	136.0

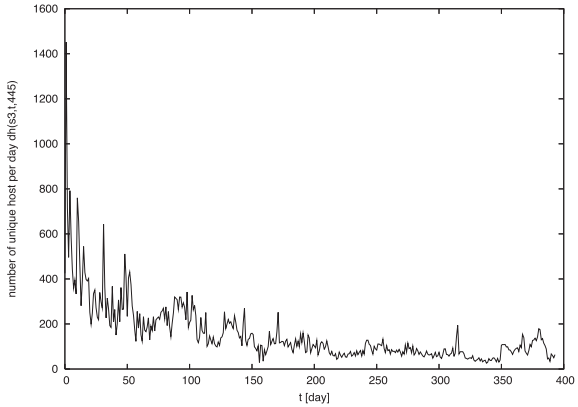

Fig. 2 Average increase of unique source addresses per day $\Delta h(s_3, \Delta t, 445)$.

Note that not all estimations are successful. For example, sensors s_7 and s_9 are implausibly shown as having more than 10^9 source addresses, which approaches

$\star 1$ It provides a degree of accuracy for the qualitative purpose, though it is obtained via the *variance-covariance matrix* after the final iteration for fitting. Note it does not determine the confidence interval.

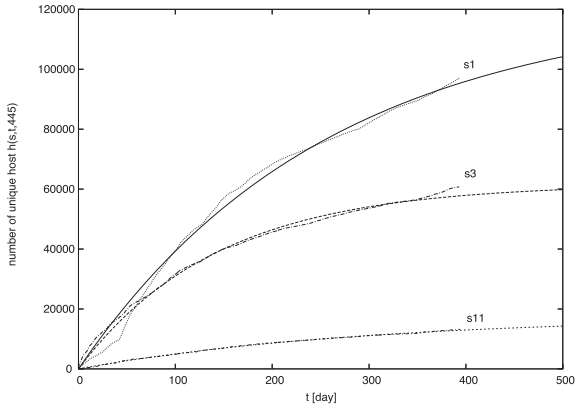

Fig. 3 Cumulative unique source addresses $h(s, t, 445)$.

Table 2 Estimated number of total malicious hosts during one year $n(1,T_1,445)$.

s	n	error [%]	n/a	error [%]	c [pckts/s]
s_1	121000	1.02	256	1.77	16.0
s_2	76900	0.82	245	1.46	17.7
s_3	61900	0.34	146	0.83	28.1
s_4	49100	0.60	250	1.03	16.4
s_5	68200	0.25	171	0.53	23.9
s_6	58300	0.89	242	1.59	16.9
s_7	4.59E+09	8.70E+03	1.87E+08	8.71E+03	0.0
s_8	65000	0.61	239	1.10	17.1
s_9	1.11E+09	6.56E+03	5.51E+07	6.57E+03	0.0
s_{10}	30700	0.80	263	1.37	15.6
s_{11}	17600	0.45	298	0.72	13.7
s_{12}	75300	1.09	330	1.69	12.4
average $\overline{n_1}$	62400	–	–	–	17.8
SD σ_1	28100	–	–	–	4.72

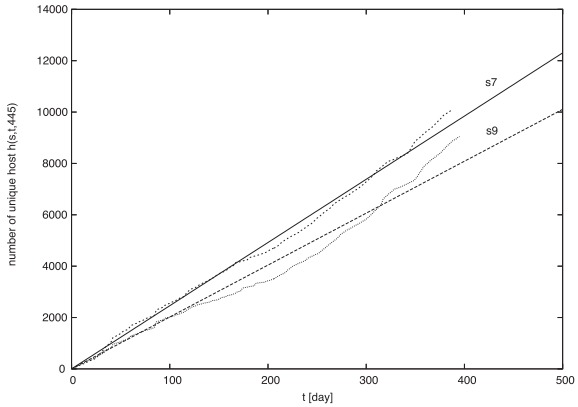

Fig. 4 Failure of estimation.

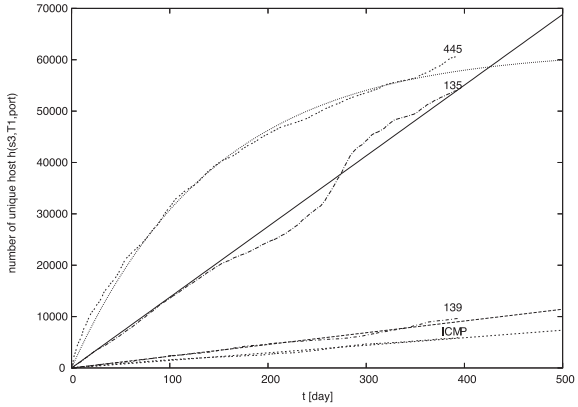

Fig. 5 Estimation of $h(s_3, T_1, p)$ w.r.t. destination port $p = 135, 139, 445$, and ICMP.

the total number of all IP addresses. To investigate the reason for this failure of estimation, details for the two sensors are shown in **Fig. 4**. In the middle of duration ($t = 200$), the number of unique source addresses increases sharply for some reason, possibly a malicious code's local impact or a sudden change of network topology. Excluding this failure of estimation, which is indicated by an n estimate of over 10^6, we show the probability density function of the estimated number of malicious hosts, n, in **Fig. 6**. Here, the most likely value of n can be seen at the first peak and the average is $\overline{n_1} = 62400$ with the confidence interval being 95% of $\pm 2\sigma_1$.

Scanning behavior may depend on malicious codes or viruses. To clarify the differences in behavior, we repeat the same steps for every port $p = 135, 139, 445$, and the Internet Control Message Protocol (ICMP), and show the differences in **Fig. 5**. In comparing destination ports, we notice a difference in the number of packets, but all cases seem to have the same asymptotic value. Therefore, we

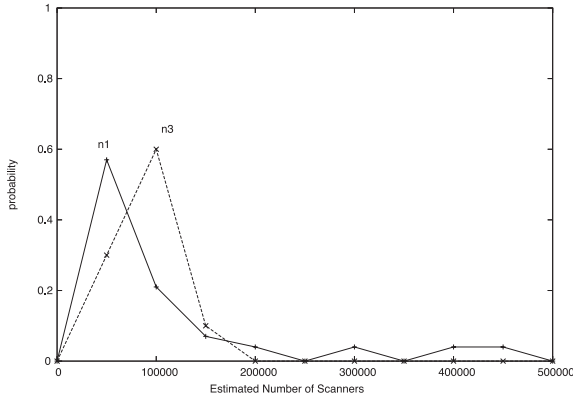

Fig. 6 Probability Density Functions of number of scanners estimated from $h(1, T_1, 445)$ (indicated with n_1) and $h(1, T_3, 445)$ (n_3).

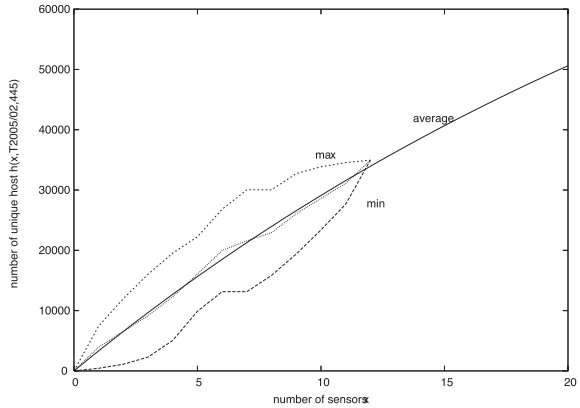

Fig. 7 Cumulative unique source addresses $h(S, T_2, 445)$ with respect to number of sensors $x = |S|$.

can claim that our model is generally appropriate for any destination port, and we will use $p = 445$ as a representative port from now on. A similar discussion is applicable to other ports.

3.4 Estimation of n Using a Number of Sensors x

Figure 7 shows the cumulative unique source addresses $h(x, T_2, 445)$ with respect to a number of sensors x, where the duration is $T_2 = [2005/2/1–2005/2/28]$ and the destination port is $p = 445$. The number of cumulative unique source addresses increases as more sensors are used for distributed observation. Note that the order of sensor choice is a critical factor in the increase of unique addresses because the number of packets varies by a factor of more than ten among sensors. To minimize the effects of this difference, we take an average between the two extreme cases, maximal and minimal, in the choice of sensors, to which we apply Eq. (4) for fitting. The estimated number of malicious hosts $n_2(S, T_{12}, 445)$ is summarized in **Table 3**.

The experimental results show the approximation of n using the number of sensors x as

$$n_2 = 112000,$$

which is consistent with the previous approximation using duration t with a confidence interval,

Table 3 Estimated unique source addresses $\overline{n_2}(S, T_1, 445)$.

duration T_2	n_2	error [%]	n_2/a	error [%]	c [pckts/s]
2005/2/1–2005/2/28	111655	24.96	33.113	28.76	123.48

$$n_1 = 62400 \pm 2\sigma_1 = [6200, 118600]. \quad (5)$$

Therefore, we claim that both models give a similar approximation of the population of scanners. The variance of n_1 is not significant.

3.5 Stability During Observation

The set of malicious hosts may be unstable over too small a period of time. On the other hand, one year may be too long to observe a set of malicious hosts because unexpected events, such as the spread of worms or a flood of packets with spoofed source addresses, would spoil the proposed model.

Therefore, we compare the fitting to a three-month duration of observation data to the estimation results for a one-year duration, as summarized in **Table 4**, where fittings are attempted for each three-month duration T_3 in the period from September 2004 through July 2005. Estimated values of n greater than 10^6 are considered as fitting failures. Successful fittings comprise 28 cases out of 12 sensors $\times 4$ durations $= 48$ pairs, i.e., the fitting success ratio is over 50%. The probability distribution of the number of malicious hosts n_3 is shown in Fig. 6.

Table 4 Number of malicious hosts estimated using three-month observation duration, $n_3(1, T_3, 445)$.

beginning	2004/09	2004/12	2005/03	2005/06	average $\overline{n_3}$
s_1	3.76E+09	95300	43900	112000	83700
s_2	2.44E+08	20700	376000	9.56E+06	198000
s_3	44800	32600	38500	32500	37100
s_4	199000	5.46E+08	25000	1.24E+08	112000
s_5	92500	33000	25000	1.24E+08	50200
s_6	55800	83200	1.30E+09	72100	70300
s_7	7.50E+08	3.42E+07	3.23E+10	2.71E+07	0
s_8	426000	22100	136000	1.47E+08	195000
s_9	2.04E+06	9750	1.69E+08	3.89E+07	9750
s_{10}	13600	1.57E+08	1.13E+08	1.70E+08	13600
s_{11}	31700	28500	39700	7950	27000
s_{12}	2.37E+08	96100	1.40E+10	262000	179000
average $\overline{n_3}$					87700
SD σ_3					106000

The three-month average is

$$n_3 = 87700 \pm 2\sigma_3 = [-124300, 299700], \quad (6)$$

which does not conflict with the first approximation, n_1 in Eq. (5), but the interval is too broad. We find two peaks in the distribution of n in Fig. 6, which is considered the source of the error. To discover the reason for the difference between n_1 and n_3, we examine all possible durations from $T = 30$ through 100 by fitting the model to the experimental data, as shown in **Fig. 8**. For example, $T = 30$ divides the one-year duration into $365 - 30 = 335$ sets of fitting data, for which we perform the fitting and investigate the interval between the minimal and the maximal approximations. The experimental results show noncontiguous behavior at $|T| = 30, 44, 52, 60$ in the figure. Possible reasons for this anomaly include the synchronous port scans performed by a *botnet* or the large-scale failure of a backbone.

3.6 Independence of Sensors

To enable uniform sampling of Internet-scale events, the sensors should be distributed uniformly over the address space. However, according to Table 1, the number of packets observed by different sensors varies by a factor of up to ten. To evaluate our model with regard to the independence of sensors, we first show the relationship between the number of packets and the cumulative number of unique source addresses, and then the jointly observed number of source addresses for each pair of the sensors.

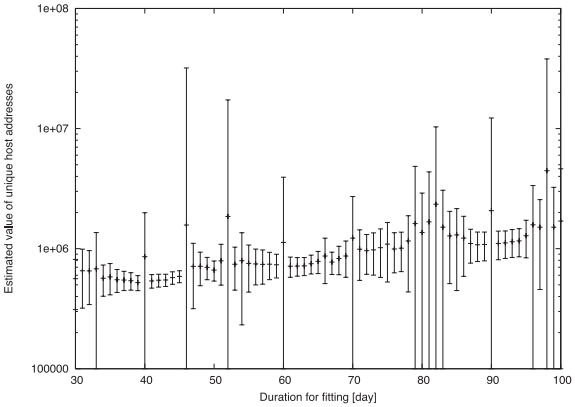

Fig. 8 Estimated number of malicious hosts for duration of observation T.

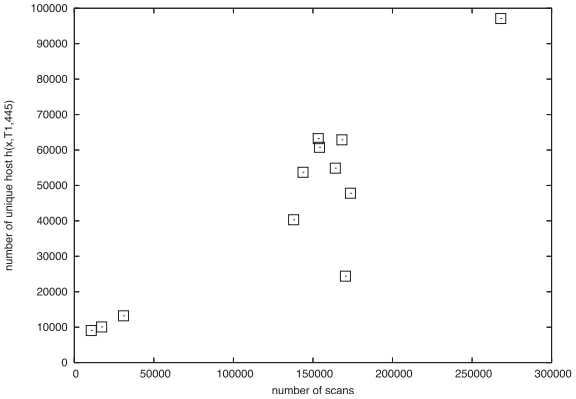

Fig. 9 Scatter diagram of numbers of packets and unique source addresses.

Figure 9 demonstrates a scatter diagram for the number of packets with destination port 445 and the cumulative unique source addresses $h(S, T_1, 445)$. The correlation coefficient is 0.93, which implies a positive correlation between them.

In **Fig. 10**, we show the number of source addresses that are jointly observed by two distinct sensors. The degree of correlation between sensor s_i and s_j is defined by

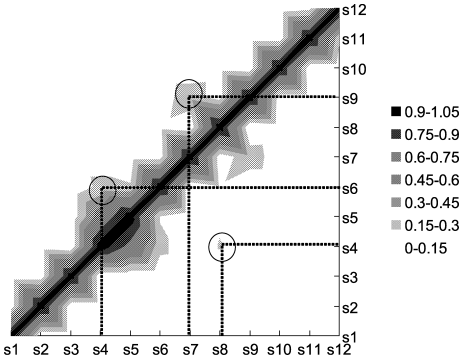

Fig. 10 Correlation between sensors $r_{i,j}$.

$$r_{(i,j)} = \frac{h(\{s_i\},t,p) + h(\{s_j\},t,p) - h(\{s_i,s_j\},t,p)}{h(\{s_i\},t,p)}, \quad (7)$$

where $h(S,t,p)$ is a cumulative unique source address observed by all sensors in S, $t = T_1$ (i.e., one year), and for $p = 445$. Note that pairs (s_4, s_8), (s_4, s_6), and (s_7, s_9) have stronger correlation factors than others. We claim that the correlations are not strong enough to violate our model about sensor independence.

4. Conclusions

We have proposed a new mathematical model for the increase of unique source addresses. Using ISDAS distributed sensors and the proposed mathematical model of the increase in unique source addresses, the number of hosts performing port scans of destination port 445 is estimated as: $\overline{n_1} = 62400$ (± 56200) using the one-year observation duration T_1, $\overline{n_3} = 87700$ (± 21000) using the three-month duration T_3, and $\overline{n_2} = 112000$ using $x = 12$ independent sensors, where a confidence interval of $95\%(2\sigma)$ is used. As a result, our experiment shows that the number of malicious hosts averages 80000, and a malicious host performs 16.8 port scans per second on average, during $T_1 = [2004/9, 2005/9]$.

The estimated results are independent neither of the duration of observation nor the starting time for observation. The fitting success ratio implies that a one-year duration is better than a three-month duration for observation. The ISDAS sensors, with a tenfold variation in the observed number of packets, are independently distributed in terms of jointly observed source addresses, but three out of 66 pairs have positive correlations. The malicious hosts are distributed uniformly over the whole address space. The frequency of port scans varies with source addresses, and typical malicious host behavior is observed by an average of 1.2 sensors per year.

Our further studies include an improvement of accuracy of estimation, which can be given significantly through enough number of sensors more than 12 used in this paper.

Acknowledgments We thank JPCERT/CC for providing the ISDAS observation data in order that we could evaluate the proposed mathematical model of scanners. We also thank Mr. Daisuke Kikuchi, Mr. Taichi Sugiyama, and Mr. Takayuki Tanaka for their contribution to the experiment. We thank anonymous reviewers for their useful comments.

References

1) Terada, M., Takada, S. and Doi, N.: Network Worm Analysis System, *IPSJ Journal*, Vol.46, No.8, pp.2014–2024 (2005) (in Japanese).
2) Kikuchi, H. and Terada, M.: How Many Scanners are in the Internet?, *WISA 2006*, Springer LNCS (2006).
3) 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).
4) JPCERT/CC, ISDAS. http://www.jpcert.or.jp/isdas/
5) Number of Hosts advertised in the DNS, Internet Domain Survey (July 2005). http://www.isc.org/ops/reports/2005-07
6) Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S. and Weaver, N.: Inside the Slammer Worm, *IEEE Security & Privacy*, pp.33–39 (July 2003).
7) Shannon, C. and Moore, D.: The spread of the Witty worm, *IEEE Security & Privacy*, Vol.2, No.4, pp.46–50 (Aug. 2004).
8) Changchun, Zou, C., Gong, W. and Towsley, D.: Code Red Worm Propagation Modeling and Analysis, *ACM CCS 2002*, (Nov. 2002).
9) Moore, D., Shannon, C., Voelker, G. and Savage, S.: Network telescopes, Technical Report, Cooperative Association for Internet Data Analysis (CAIDA) (July 2004).
10) 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).

(Received November 25, 2007)
(Accepted March 4, 2008)
(Original version of this article can be found in the Journal of Information Processing Vol.16, pp.100–109.)

Hiroaki Kikuchi was born in Japan. He received 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 in the Department of Information Media Technology, School of Information Science and Technology, Tokai University. He was a visiting researcher of 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 Information Processing Society of Japan (IPSJ), the Japan Society for Fuzzy Theory and Systems (SOFT), IEEE and ACM.

Masato Terada was born in Japan. He received the 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. Since 2002, he has been studying at the Graduate School of Science and Technology, Keio University and received 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 as well.

Naoya Fukuno was born in Japan. He received the B.E. and M.E. from Tokai University in 2005 and 2007, respectively. He joined the Internet Security Systems K.K. in 2007 and is currently a researcher of IBM Japan Ltd.

Norihisa Doi received the B.E., M.E., and Ph.D. in computer science from Keio University, Japan in 1964, 1966, and 1975, respectively. He is currently a professor of the Faculty of Science and Engineering, Chuo University, and also professor emeritus of Keio University. He is a Vice-President of the Science Council of Japan. He is also a member of the Council for Science and Technology (MEXT), a Vice-Chair of the Council for Information and Communication (Ministry of Public Management, Home Affairs, Posts and Telecommunications), President of the Non-Profit Organization Japan Information Security Audit Association, and Chair of the Japan Chapter of the Association for Computing Machinery (ACM). His research interests include software and information security. He is a member of various societies including IPSJ, IEICE, JSSST, IEEE, and ACM.