Call Behavioral Analysis to Thwart SPIT Attacks on VoIP Networks

Hemant Sengar[1], Xinyuan Wang[2], and Arthur Nichols[1]

[1] Technology Development Dept., Windstream Communications, Greenville, SC 29601
{Hemant.Sengar,Arthur.Nichols}@windstream.com
[2] Dept. of Computer Science, George Mason University, Fairfax, VA 22030
xwangc@gmu.edu

Abstract. The threat of voice spam, commonly known as Spam over Internet Telephony (SPIT) is a real and contemporary problem. If the problem remains unchecked then it may become as potent as email spam today. In this paper, we present two approaches to detect and prevent SPITting over the Internet. Both of our approaches are based on the anomaly detection of the distributions of selected call features (i.e., day and time of calling, call durations etc.). The first approach uses *Mahalanobis Distance* as a summarization tool and it is able to reliably detect individual spam VoIP calls at a microscopic level. The second approach is designed to detect groups of (potentially collaborating) VoIP spam calls at a macroscopic level. By computing *entropy* of call durations of groups of calls, we are able to build profile of normal calls and reliably detect the deviation from normal human call behavior that are caused by bulk spam calls. We empirically validate our VoIP spam call detection approaches with real VoIP call traces obtained from a VoIP service provider network. Our experimental results show that call feature distributions can be used to build a fairly general and effective anomalous call behavior detection framework.

Keywords: Voice Spam, SPIT, VoIP, Behavioral Analysis.

1 Introduction

In Japan where the VoIP market is more mature than USA has witnessed some recent voice spam attacks. The SoftbankBB, a VoIP service provider with 4.6 million users, has reported 3 incidents of spam attacks within its own network [9]. Similarly, Columbia University at New York experienced voice spam attack, with someone accessing the SIP proxy server and "war dialing" a lot of IP phone extensions [10]. Technically, it is easier for the spammer to generate unsolicited bulk VoIP calls and target multiple VoIP subscribers than generating spam calls over PSTN. As the number of VoIP subscribers hits a critical mass, it is expected that VoIP spam will emerge as a potentially serious threat. If the SPIT problem is not effectively addressed, it may become as rampant as email spam today and hinder the deployment of IP telephony.

M. Rajarajan et al. (Eds.): SecureComm 2011, LNICST 96, pp. 501–510, 2012.

The *Internet Engineering Task Force* (IETF)'s RFC [7] analyzed the voice spam problem in SIP environment and examined various potential solutions for solving the email spam problem. Unfortunately, many of the anti-spam solutions that have been proposed or deployed are either heavily influenced by or directly inherited from the email spam world. For example, the anti-spam solutions based on *computational puzzles* [7] try to frustrate the VoIP spam call generator by requiring it to solve some computational puzzles. While such methods require modification of the underlying signaling protocol, they are not effective against distributed VoIP spam call generation where multiple powerful PCs are compromised into zombies and used for generating bulk spam calls. The *Turing tests* [7, 11] based approaches, on the other hand, require manual and active involvement of callers, which is not intuitive and may scare away many potential users. The solutions relying on *social network* [2, 14] and *caller's reputation value* [8, 5, 1] require infrastructure changes and modifications of SIP UAs, yet they are susceptible to malicious reputation poisoning. The anti-spam solutions based on a *trusted third party* [4] are not scalable. Similarly, it is hard to apply the *content based filtering* [3] to voice spam since the real-time voice content analysis and is exceedingly difficult. Recently, Wu et al. [13] proposed a spam detection approach involving user-feedback and semi-supervised clustering technique to differentiate between spam and legitimate calls. However, the current generation of telephone sets do not provide an option to give feedback of a call to service provider's system. In summary, voice spam problem can not be effectively addressed by simple adaption of existing email spam solutions or asking for overhauling of network infrastructure and signaling protocols.

In this paper, we propose two approaches for detecting VoIP spam calls. Both approaches build normal call behavior upon distribution of selected call characteristics (e.g., day and time of the call, call duration) and neither of them requires callee's feedback or modification of the underlying signaling protocol. Compared with existing VoIP spam defenses, our proposed approaches have the following advantages:

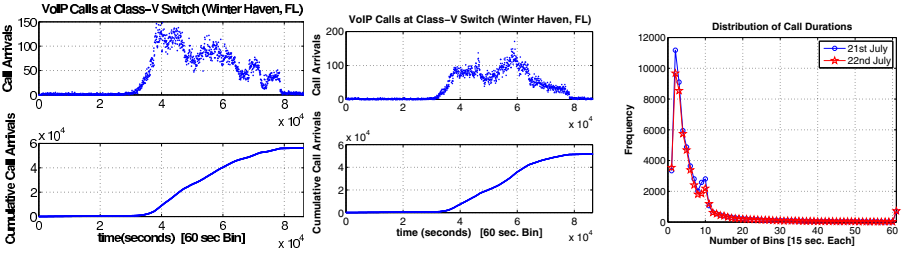
- They are transparent to end users, and they do not require any explicit feedback from the end users or modification of the underlying signaling protocols or UAs.
- They are designed to detect both sporadic and bulk VoIP spam calls. The proposed approach is able to suppress VoIP spam calls from local, authenticated callers.

We empirically evaluated our VoIP spam detection approaches using real VoIP call traces, and our results show that our approaches are effective in detecting both individual and bulk VoIP spam calls.

The remainder of the paper is structured as follows. In section 2, we establish the baseline of normal VoIP call behavior. In section 3, we present our first approach to detect individual local misbehaving callers. In Section 4, we discuss how to distinguish normal human generated calls from bulk machine generated spam calls based on entropy measurement of call duration. Section 5 concludes the paper.

2 Baseline of Normal VoIP Call Behavior

In this section, we establish the baseline of normal VoIP call behavior. Specifically, we used the call logs collected from a VoIP network of NuVox Communications, a voice service provider in Southeast and Midwest regions of the USA [6]. The seven days (July 21 - 25, July 28, and August 04' 2009) call logs were collected from a Class-V switch located at Winter Haven, Florida. The call logs correspond to VoIP calls made by subscribers of Orlando and Tampa cities in Florida. Figure 1 shows the call arrivals and the distribution of call duration characteristics of two days (21^{st}-22^{nd} July'09). Each of the call logs are of 24 hours duration starting at the midnight. The logs of 21^{st} and 22^{nd} July contain 56259 and 51625 successfully completed calls, respectively.



(a.) Call Arrivals (21st July) (b.) Call Arrivals (22nd July) (c.) Distribution of Call Durations

Fig. 1. Call Arrivals and Distribution of Call Durations

Call Duration Probability Distribution. The call logs for VoIP traffic traces are analyzed to obtain call duration distribution. As shown in Figure 1 (c.), we observe that $\approx 50\%$ of the calls complete within a minute. The measured call durations are used to calculate the mean μ and standard deviation σ. The mean and standard deviation pair (μ, σ) [in seconds] for the 21^{st}-22^{nd} July VoIP traces are found to be $(111.87, 264.04)$, and $(115.83, 283.58)$, respectively.

3 Detecting Individual Misbehaving Subscriber

In this section, we focus on detecting individual misbehaving VoIP subscribers who are local and authenticated to the protected VoIP network. A VoIP caller can be classified as local or external subscriber based on the following attributes: 1) the source IP and the status of `REGISTER` message – the successfully completed `REGISTER` transaction lets us know that this particular subscriber is local (i.e., subscriber account is maintained by the service provider) and also from where to expect next outbound call request; 2) the SIP URI and the source IP of `INVITE` call requests that do not have corresponding `REGISTER` messages – these inbound call requests represent external unauthenticated subscribers and the source IP determines whether the request is from one of the peering partners or known business SIP trunking customers.

Discriminant Analysis Based on Mahalanobis Distance: The spam detection module (collocated with the session border controller) detects abnormal call behavior of individual local subscribers in the collection of past calling data points, going through a process consisting of two phases: the *training phase* and the *testing phase*. During the training phase, for each of the local subscribers we collect *day*, *time of calling*, and *call duration* for successfully completed calls. Since each subscribers calling behavior is quite different, we need a common base to make comparison and find out how individual subscribers deviate from the base. This common base is known as a *reference pattern*.

Later, the whole day is divided into small time periods of ΔT (= 15 min.) where individual subscriber's call behavior is compared with common *reference pattern*. The common reference pattern can be assumed to belong to a *virtual user* generating exactly 5 calls within each time window. The call arrivals are assumed to be poisson distributed with mean of 180 sec., and the call durations are exponentially distributed with mean talk time of 60 sec. Within a time window if a subscriber has less than 5 calls, we ignore that time window as this low call-rate cannot be a spam call behavior. Otherwise, using the *Mahalanobis distance*, we measure the distance between two multivariate data sets. In the training phase, the measured distances are used to derive a threshold i.e., an upper bound of distance values considered to be a normal call behavior. In the testing phase, we determine if the measured distance of a time window falls beyond a threshold value raising an alarm.

More formally, now assume that on a particular day of the first week and within a particular time window we have observed n realizations of a $d-$dimensional random variable. From the data set we get a data matrix $\chi(n \times d)$

$$\chi = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}$$

The row $x_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$ denotes the i^{th} observation of a d-dimensional random variable $\chi \in \mathbb{R}^d$. The *center of gravity* of the n observations in $\mathbb{R}^d$ is given by the vector $\overline{x}$ of the means $\overline{x}_j$ of the d variables:

$$\overline{x} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_d \end{pmatrix} = n^{-1}\chi^T 1_n$$

The dispersion of the n observations can be characterized by the covariance matrix of the d variables:

$$S = n^{-1}\chi^T\chi - \overline{x}\,\overline{x}^T$$

This matrix can equivalently be defined by

$$S = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T$$

Now our task is to compare the observed data matrix $\chi_p(n \times d)$ with the reference data matrix $\chi_q(m \times d)$ and find out how calls within a particular time window is correlated with the reference. We use Mahalanobis distance to measure the similarity between two data matrix [12]. The Mahalanobis distance between two populations p and q is defined as:

$$d_{pq} = \{(\overline{x}_p - \overline{x}_q)^T \Sigma^{-1} (\overline{x}_p - \overline{x}_q)\}^{\frac{1}{2}}$$

where Σ is pooled unbiased covariance matrix

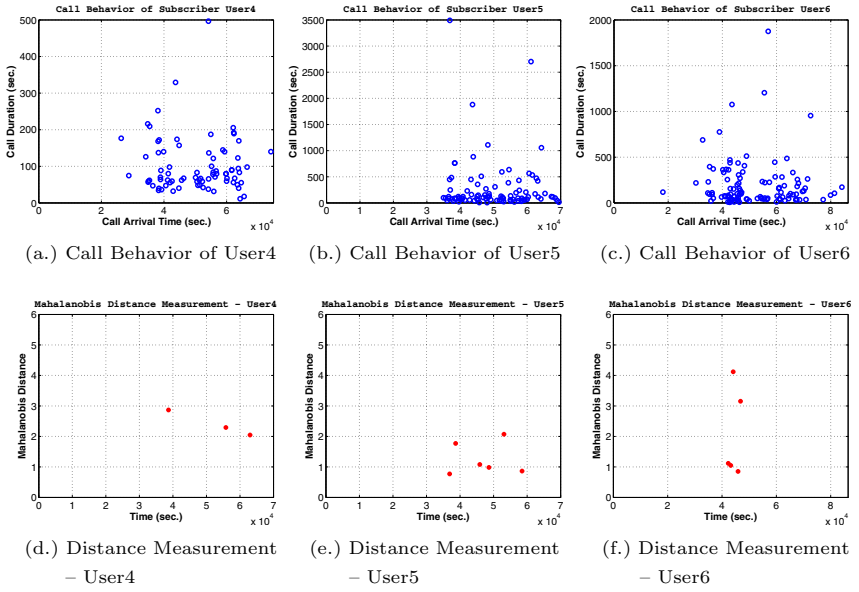
$$\Sigma = [(n-1)S_p + (m-1)S_q]/(n+m-2)$$

Threshold Determination. In the training phase, the distribution of measured Mahalanobis distances are used to calculate the mean μ of all observed distances. To set an upper bound on distance values that may act as a threshold, we use $d_{thresh.} = \mu + n * \mu$, where $n \geqslant 0$. The value of n defines a confidence band where subscriber's calls falling in the region are treated as normal calls. Beyond this normal region, the observed distances are abnormal raising an alarm. The lower value of n governs the detection sensitivity, however at the cost of more false alarms.

White Listing to Suppress VoIP Spam Calls From Local, Authenticated Callers. Based on the normal call profile and the determined threshold, we can determine if an outgoing call from local caller is normal or not. We can further put any active local caller into a dynamic white list if most of its calls are determined normal. This would allow us to suppress VoIP spam calls from those local callers that are not in the dynamic white list. This suppression should only be used when it is determined local callers have issued bulk spam calls.

Empirical Validation: To demonstrate the applicability of the proposed method, we analyzed the call behavior of ≈ 50 subscribers. As a representative sample, from the 21^{st} July call log we randomly selected six local subscribers of varying call rate. The per subscriber data set derived from the successfully completed calls within a particular time window is used to calculate the Mahalanobis distance.

Each individual subscriber is compared with the reference data set to get a whole day's distribution of Mahalanobis distance. This comparison is a part of training phase where we determine as how far a subscriber's legitimate call behavior may deviate from the reference data set as shown in Figure 2. The average of all distance values is found to be 1.21. It is used to derive an upper bound (i.e., $d_{thresh.} = 1.21 + 4 * 1.21 = 6.05$) beyond that calls are assumed to be abnormal. In our experiments we observe that the confidence band of $4 * \mu$ (i.e., $n = 4$) achieves high detection sensitivity with no false alarms. The so obtained threshold value is used to detect misbehavior of callers in the testing phase. The call logs of July 28 and August 04 are used as testing data set. Figure 3 a.), b.) and c.) plot the two whole day's data points for subscribers *User4*, *User5*, and *User6*, respectively. In the testing phse, for each individual time windows where



(a.) Call Behavior of User4 (b.) Call Behavior of User5 (c.) Call Behavior of User6

(d.) Distance Measurement – User4 (e.) Distance Measurement – User5 (f.) Distance Measurement – User6

Fig. 2. Distance Measurement To Determine Threshold Value [Training Phase]

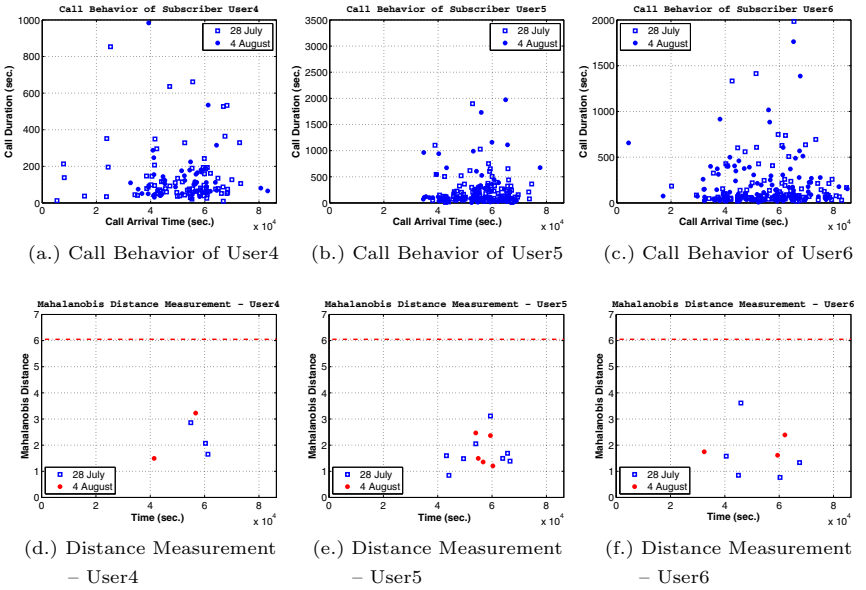
we observe at least 5 calls is compared with the common reference data set to compute a similarity value using Mahalanobis distance as shown in Figure 3 c.), d.) and e.). We observe that for both of these days, the distance values remain well below the threshold value.

Now we mix 20 attack instances (each at an hour apart) within the 28^{th} July call log and each attack instance consists of 20 spam calls. The call arrivals are assumed to be poisson distributed with mean of 30 sec., and the call durations are exponentially distributed with mean talk time of 15 sec. The measured effectiveness of Mahalanobis distance classifier is summarized in Table 1.

4 Detecting Groups of Misbehaving Calls

The proposed scheme in the previous section is to detect abnormal call behavior of authenticated (i.e., local) callers at an individual level. In this section, we develop an entropy-based approach to detect unusual call behavior at an aggregated level irrespective of being local or external subscribers. The basic insight is that if a number of callers misbehave by performing low-rate attacks, it is possible that at an individual level the call behavior may seem benign, however at aggregated level, the entropy-based approach sums up these individual low-rate spam attacks leading to an efficient and easier detection mechanism without maintaining call behavior profiles for unknown and unauthenticated external callers and thus avoiding unnecessary lookups and excessive entries in the database.

Few Observations. In the case of spam attacks, the machine generated bulk calls will either be answered by subscribers (i.e., humans) or end up at the



(a.) Call Behavior of User4 (b.) Call Behavior of User5 (c.) Call Behavior of User6

(d.) Distance Measurement – User4 (e.) Distance Measurement – User5 (f.) Distance Measurement – User6

Fig. 3. Distance Measurement To Detect Unusual Calling Behavior [Testing Phase]

voicemail system. If the spam calls are answered by subscribers then the average of call durations is expected to be short compared to other regular calls. Therefore, during the attack, the average of call durations will fall. Further, if the spam calls are answered by a voicemail system, still we are expected to observe unusual behavior. Generally, a voicemail system allows voice recording of only few minutes (a typical value is of $2-3$ minutes). At the expiration of voice recording timer, the voicemail system terminates the call. Hence, many of the calls will be having a constant call duration.

Entropy Classifier. The entropy classifier component makes spam attack detection based on entropy measurement of call durations. The call durations are binned into N contiguous bins (of varying lengths). We can interpret the bins as the states x_i of a discrete random variable X, where $p(X = x_i) = p_i$. The entropy of the random variable X is then

$$H[p] = -\sum_i p(x_i)\ln p(x_i) \tag{1}$$

Distributions $p(x_i)$ that are sharply peaked around a few bins will have a relatively low entropy, whereas those that are spread more evenly across many bins will have higher entropy. For example, if the entropy is low for our selected attribute of *call duration* then it indicates predictable patterns of the abnormal call behavior. It could be due to short call durations are skewed toward few selected lower-side bins or may be constant call durations have filled up one (or few) particular bin(s). However, if the measured entropy is high (i.e., call durations

Table 1. Performance of Mahalanobis Distance Classifier*

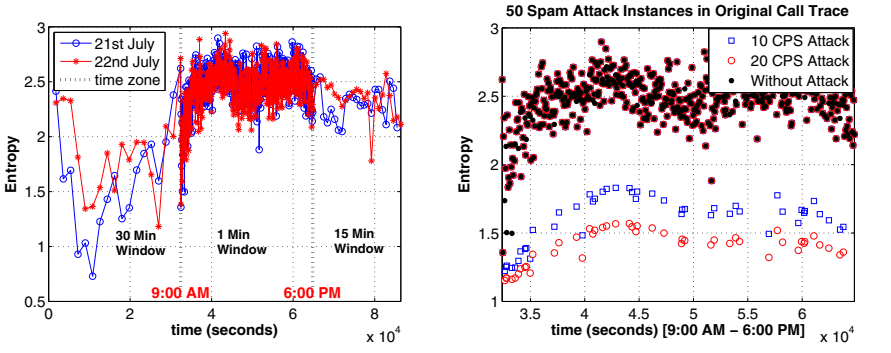
20 Attack Instances Introduced in the Whole Day Traffic of 28^{th} July						
Calls/ΔT	$\mathbb{U}ser1$	$\mathbb{U}ser2$	$\mathbb{U}ser3$	$\mathbb{U}ser4$	$\mathbb{U}ser5$	$\mathbb{U}ser6$
Spam Attack Detection Probability [Poisson arrival mean = 30 sec.]						
20	100%	95%	100%	95%	95%	90%
15	100%	85%	90%	75%	80%	75%
10	100%	80%	85%	70%	75%	70%
Spam Attack Detection Probability [Poisson arrival mean = 20 sec.]						
20	100%	85%	95%	80%	85%	80%
15	100%	80%	90%	80%	75%	80%
10	100%	70%	80%	70%	75%	75%
Spam Attack Detection Probability [Poisson arrival mean = 15 sec.]						
20	100%	80%	90%	80%	80%	80%
15	100%	80%	90%	80%	75%	75%
10	100%	80%	85%	65%	75%	75%

* Without removing the outlier data points.

are distributed across bins), it indicates the irregular or unpredictable behavior of human conversations.

Entropy Measurement of Call Durations. In our experiments, the binning of call duration data points use 61 contiguous bins. The first 60 bins are of 15 sec. each and the last 61^{th} bin is a default bin to capture all call durations that are longer than 15 minutes. However, it should be noted that the choice of fine granular bins is more accurate in classifying the attacks since it leads to a better estimate of the entropy. In our study of call duration entropy, we divide the whole day in three separate time zones based on the observation of call arrival rate. The first time zone starts at midnight and ends at 9:00 AM. In this time zone the call arrival rate is very low (e.g., see Figure 1).

The entropy estimation is based on 30 minutes time window to make sure that we collect enough data points. As the time increases, the call rate also increases resulting in the growing trend of entropy. The second time zone represents usual



(a.) Entropy Measurement (b.) Entropy Measurement Under Spam Attack

Fig. 4. Entropy Measurement

working hours between 9:00 AM and 6:00 PM where call rate is usually high. In this time zone we use 1 minute time window for entropy estimation. In our analysis, we find that the busy hour entropy remained confined between 2.0 and 3.0 as shown in Figure 4 (a). The third time zone starts at 6:00 PM and ends at midnight. In this time zone, we use 15 minutes time window for entropy measurement that generally varies between 2.0 and 2.5. The off-peak hour entropy is more unpredictable (especially between midnight to 9:00 AM).

Determination of Entropy Cutoff Scores. To use entropy measures for spam attack classification, based on previous collected data during the training period, we build a entropy profile of call durations with respect to time. The measured entropy is used to set a cutoff score and if the test score (during the testing period) is greater than or equal to the cutoff score, the call requests are classified as human generated. If the test score is less than the cutoff score, the call requests are classified as malicious spam calls. The cutoff score and its relation with time is an important parameter in determining the false positive and true positive rates of the entropy classifier. Since in the first time zone the call rate is very low so to avoid detection, most of the attacks are expected to occur during the busy hour of call traffic where malicious calls can easily hide among legitimate call traffic. Our focus is mainly on this time segment. Note that with the proper setting of threshold values, there will be no false alarm (i.e., false positive) under normal conditions. However, to balance both false positives and false negatives, we set our entropy threshold at 1.75. In two day's call log analysis we observed that out of 1082 observations, 4 observations had entropy value below the threshold value of 1.75. Therefore, 0.37% times the entropy value falls below the threshold value and thus giving us false alarms.

Empirical Evaluation of the Entropy Classifier: Now we empirically evaluate the effectiveness of the proposed entropy classifier in terms of its spam detection accuracy. In our experiments, we made the following three assumptions: 1.) during busy-hour spam attack, 95% calls are answered by humans and the remaining 5% by the voicemail system; 2.) for simplistic reason we assume that the human answered call durations are exponentially distributed with mean talk time of 15 sec.; and 3.) the voicemail system's recording time limit is of 2 minutes. After 2 minutes of recording, the voicemail calls are terminated by the voicemail system.

In our experiments, the call logs are used to generate call requests and used as the normal background traffic. Later, this traffic is randomly mixed with the spam traffic of varying call rates. For example, during the busy hour between 9:00 AM to 6:00 PM, we introduce 50 individual spam attack instances of 10, 20, 30, 40 and 50 calls per second. Each of these attack instances lasts for a small time period of 30 seconds and thus introducing 300, 600, 900, 1200, and 1500 spam calls per attack instances. Figure 4 (b) shows 50 individual attack instances (three times two individual attack instances fell within the same time window). These attack instances belong to two different call rates of 10 and 20 CPS. Under spam attack, we could observe as how entropy drops from those representing the

normal call behavior. To measure false negatives, we use detection probability that is defined as the percentage of the successful identified attack instances over the total launched attacks in one set of experiments. The results demonstrate that our proposed entropy classifier is able to reliably detect aggregated ($\geq$ 20 calls per second) VoIP spam calls with no more than 0.37% false positive rate.

5 Conclusion

SPIT is touted as the next biggest spam threat after email spam. To mitigate the potential threat of voice spam, this paper proposed two complementing and yet practical schemes. The first scheme, which is based on Mahalanobis distance, can detect unusual call behavior at the individual subscriber level. The second approach utilized entropy of call durations to detect spam attack at an aggregated level. It can detect spam attacks when a group of subscribers misbehave. The empirical results of our study show that it is feasible for a VoIP service provider to detect VoIP spam attacks irrespective of whether it is launched from within an enterprise network, peering partners or from subscribers.

References

1. Balasubramaniyan, V., Ahamad, M., Park, H.: CallRank: Combating SPIT Using Call Duration, Social Networks and Global Reputation. In: The Fourth Conference on Email and Anti-Spam (2007)
2. Dantu, R., Kolan, P.: Detecting spam in voip networks. In: Proceedings of the Steps to Reducing Unwanted Traffic on the Internet on Steps to Reducing Unwanted Traffic on the Internet Workshop (2005)
3. Graham-Rowe, D.: A Sentinel to Screen Phone Calls (2006), http://www.technologyreview.com/communications/17300/?a=f
4. Kayote Networks. The Threat of SPIT (2007), http://www.kayote.com/
5. Niccolini, S., Tartarelli, S., Stiemerling, M., Srivastava, S.: SIP Extensions for SPIT identification. draft-niccolini-sipping-feedback-spit-03, IETF Network Working Group (2007) (work in progress)
6. NuVox Communications. Service Provider (2009), http://www.nuvox.com
7. Rosenberg, J., Jennings, C.: The Session Initiation Protocol (SIP) and Spam. RFC 5039, IETF Network Working Group (2008)
8. SIPERA. Sipera IPCS: Products to Address VoIP Vulnerabilities (April 2007), http://www.sipera.com/index.php?action=products,default
9. VOIPSA. Confirmed cases of SPIT. Mailing list (2006), http://www.voipsa.org/pipermail/voipsec_voipsa.org/2006-March/001326.html
10. VOIPSA. VoIP Attacks in the News (2007), http://voipsa.org/blog/category/voip-attacks-in-the-news/
11. Wikipedia. Turing test (2009), http://en.wikipedia.org/wiki/Turing_test
12. Wikipedia. Mahalanobis distance (2010), http://en.wikipedia.org/wiki/Mahalanobis_distance
13. Wu, Y.-S., Bagchi, S., Singh, N., Wita, R.: Spam Detection in Voice-Over-IP Calls through Semi-Supervised Clustering. In: IEEE Dependable Systems and Networks Conference (DSN 2009) (June-July 2009)
14. Rebahi, Y., Al-Hezmi, A.: Spam Prevention for Voice over IP. Technical report (2007), http://colleges.ksu.edu.sa/ComputerSciences/Documents/NITS/ID143.pdf