Detection and Mitigation of Spam in IP Telephony Networks using Signaling Protocol Analysis

Robert MacIntosh and Dmitri Vinokurov
Alcatel
600 March Rd., Ottawa, ON Canada

***Abstract*-We propose an innovative method to detect and block spam (unsolicited bulk calls) in IP Telephony networks. Rapid adoption of Voice over IP (VoIP) technology introduces new powerful options for spammers and telemarketers to increase their productivity and effectiveness. To date a few concepts have already been proposed in the VoIP spam prevention area. These prior solutions focus mainly on identity control and reliable authentication and follow in this regard the email spam solutions approach. Such measures imply at least strong collaboration of service providers and universal standardization. In this paper we describe a new method based on the analysis of the VoIP signaling messages which can assist service providers in detecting spam activity targeting their customers. This "locally centric" approach would enable a service provider to handle the call before the actual voice spam content reaches the recipient. The detection parameters depend solely on the local service provider's policy; no end-users participation/compliance is required.**

I. Introduction

We define voice spam as unsolicited bulk calls each resulting in a media session, where the content delivered to a phone or voice terminal may include voice, images or video. Therefore, in the context of this paper, by this definition we imply the presence of establishment and termination of interactive spam media sessions. The presented solution addresses the most challenging case when the source of spam is not located in the monitored/controlled network, i.e. it is independent of the policy or level of collaboration of the spammer's service provider. Various kinds of calls are covered: advertisement, telephone polls, telemarketing, etc.; the solution is not focused only on automated distribution of pre-recorded messages and it is not specific for any Voice over IP (VoIP) signaling protocol. The targeted consumer of such solution is a service provider controlling a signaling server for the end users in the provider's network. Similarly, this is also applicable to enterprises with IP telephony gateway. All VoIP signaling traffic going to and from the provider's customers (enterprise employees) is assumed to be observable on this server, which we will further refer as a Gateway.

In this paper, we use various messages from particular Internet telephony signaling protocols as examples and for methodology instantiation purposes, with no loss of generality.

II. Problem Description

A. *Voice Spam Specifics*

The voice spam issue impacts overall VoIP service in terms of consumed voicemail space, enterprise employees' distraction, and service subscriber dissatisfaction. VoIP technology simplifies creation of automated tools for bulk calls generation [1]. Some concerns about the issue are well described in [2].

As all other VoIP calls, spam over Internet telephony (known as SPIT) consists of two parts: signaling and media data. In this concept it is important to differentiate SPIT from denial-of-service (DoS) attacks, in regard to the behavior of VoIP traffic, both signaling and data. Spam does not aim to destroy the service; spammers are interested in its correct functioning. The implication of this is two fold. First, the techniques used for detection of DoS attacks are generally not applicable: there are no malformed packets, incomplete call setups, or floods. Second, call routing information provided in the call setup requests is valid and therefore, can be used in the further analysis.

SPIT detection has its specific challenges which have already been described in [3] and [4]. Here we point to additional issues that may impact the ability to identify a call as SPIT. Analyzing data content may be not only impractical but also not legal in many cases. Any call handling decision must be made in real-time before the actual media session starts. End user upgrades to support a new detection technology are impractical and should not be expected or relied upon. Caller anonymity services may be employed, either maliciously or unintentionally, as illustrated later. The commonly agreed requirement of a SPIT solution is to detect the call as spam before the actual call happens, i.e. during signaling exchange stage.

The approaches currently proposed to address the voice spam issue [3, 4, 5, 6] have their drawbacks. Some solutions essentially rely on the identity of the caller: black/white lists, enforced caller introduction [7], calls rate limiting. Thus, each caller's identity must be authenticated and strong collaboration and trust built between service providers would be required for end-to-end solution [3], whereas some providers or individual enterprises might look for independent turnkey solution. Such statistical metrics as call rates, spacing between calls, and call duration [5] can work with automated calls only and may suffer from false positive alarms caused by some legitimate users or services (e.g., VoIP gateway of broker office) or by malicious disguising floods. Regulative measures like Do Not Call Registry [8] may not be obeyed by the spam sources outsourced off-shore.

B. Anonymity

VoIP technology provides freedom for using aliases and anonymity services. The incoming calls can be anonymous in that fact that the recipient is unable to determine the actual caller. This does not mean that the call recipient cannot authenticate the direct calling party; Figure 1 below illustrates possible scenarios when this would be possible.

We consider the scheme based on the Session Initiation Protocol [9] (SIP) as an example to consider which parameters might be included in signaling messages when they reach the recipients of SPIT. The spammer must have registered on Proxy1 his current location under some publicly routable name that we denote as spammer's Registered Location (sRL). Proxy2 belongs to the service provider looking for SPIT solution or could be an enterprise Gateway. GW1 and GW2 are signaling gateways between IP domain and SS7 signaling network. There are few ways to achieve anonymity of the caller for the callee. The spammer can use SIP B2BUA as network privacy service as described in [10]; in this case SPIT would use path 1-2 as in the Figure 1. Also, the spammer can anonymize the *From*: header field value but the correct value of sRL would be set for the *Contact*: header field by Proxy1. This path is identified as 3-4. However, the sRL value could be varied by re-registration of the spammer on the Proxy1. Another way to hide the identity would be to use local proxy and to route the calls' signaling through SS7 network via GW1, and then reach IP enabled recipients with the help of GW2. Ref. [11] defines the mapping between ISUP and SIP signaling that takes place at signaling gateways. The Calling Party's Number (CIN) parameter will be omitted by GW1 from the outbound Initial Address Message (IAM) if the value of *From:* field of SIP INVITE is unusable (e.g., distorted or characters are national-specific). On the GW2, if the CIN in IAM is omitted, the hostname of the GW2 as *sip:gw.carrier.com* (or something else according to the local policy) will be used in the *From:* header. The path of call setup request for this type of scenario would be 3-6-7-8 or 5-6-7-8. Below are the characteristics of the marked key points of the scheme considered in the Figure 1:

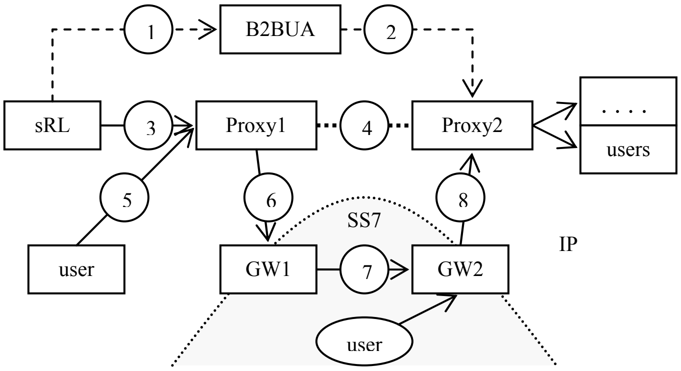

Figure 1. Anonymity in SIP Centrex model.

1: Regular header fields
2: No CallerID, *Contact:* B2BUA
3: *From*: random alias
4: *Contact*: session counterpart (sRL)
5: *From*: anonymized or non-displayable
6: Same as 5
7: No CIN
8: No CallerID, *From*: GW2, *Contact:* GW2

Thus, the call may be anonymous for the recipient but signaling routing data observable on Proxy2 can point either to anonymizer A, signaling gateway GW2 or media session party.

III. SPIT Scenarios and Implications for Signaling

The idea of spam detection we propose is based on three main constituents. First, the observable signaling routing data of the voice spam are valid and may point either to anonymizer, gateway or spammer. Second, spam calls are unidirectional: spammer initiates the calls to the targeted network, but nobody initiates calls to him. Third, spam calls termination behavior is statistically consistent, i.e. these calls are terminated mostly by the same conversation party. Who exactly that party is, recipient or originator depends on the voice spam distribution scenario, possible variants of which we consider below. The statistics we consider are calculated per SPIT source over the number of calls made from this source to the recipients in the monitored network. Besides, SPIT also has the forth distinction: the spammer does not call the same recipient again for some period of time. However, in order to exploit this to the favor of SPIT detection, each call's time and destination must be kept for further analysis, which makes such approach heavy.

A. Persistent telemarketer

The voice spam calls are initiated by the operator who almost never terminates the conversation before the recipient does so first. This could be the case by few reasons. For instance, the caller may be persistent in his offer to recipients hoping to achieve positive results in every call or the calling operator may be not allowed to end the conversation first following the job professional code. Telephone polls are in this category. Thus, for this particular SPIT source, statistically call setup requests go from the spammer to recipients, whereas termination requests flow from recipients to the spammer.

B. Time-conscious spammer

The difference from the previous scenario A is that telemarketer of this type tries to cover as many recipients as possible and consistently hangs up as soon as he figures out from the conversation that his offer is unlikely going to be accepted, which is by far the most likely case. As a result, for this SPIT source, statistically both call setup and termination requests go the same direction from the spammer to recipients.

C. Prerecorded message

SPIT is being distributed by an automated calling engine as a played message. In this case, the person picking up the phone is the one who almost always terminates the call at the stage of playing the message, sooner or later after the message starts. The exception is the small percentage of calls when the recipient follows the dialing instructions that might be in the spam message and gets connected to the operator. After this happens, these calls may fall into category A or B where A is the most likely behavior. Thus, overall for this SPIT source, statistically call setup and termination requests go the same ways as in scenario A.

D. Message deposited to the voice mailbox

Spam call can meet the "busy" or "no answer" conditions on the recipient side. Further spammer's action depends on his policy: he can either leave the message or terminate the session as soon as presence of voice mailbox is detected. In either case, both setup and termination requests go from the spammer to the recipient's side. It is very important to distinguish between regular call and voice mail deposit and count these types of events separately. Otherwise the call termination statistical pattern for scenarios A and C would be mangled as a result of the mix, while the mere fact of many calls made from certain location by itself may not be a sufficient sign of spam.

There is a way to detect voice mail events on the Gateway based on signaling. For instance, SIP signaling for voice message deposit is described in [12]. Proxy2 can recognize the INVITE messages generated and sent towards recipient's voice mail system. Each of these messages has *Contact:* header field set to its value from the original INVITE sent by the caller to the Gateway. This allows to correlate the calls' originator and the number of voice messages this caller has deposited.

E. Calls set by third party

SPIT calls as well as other VoIP calls could be initiated by the third party, when *X* sends some signaling messages to *Y* and sets a media session between Y and Z. X here could be a zombie network element (PC, server, router, etc.), Y is a recipient, and Z is a source of spam media – either telemarketer or media server with prerecorded message played. X should be considered as a spam originator. In SIP, this functionality could be provided by REFER method [13] if it is supported by the recipient's terminal. This is depicted in the Figure 2.

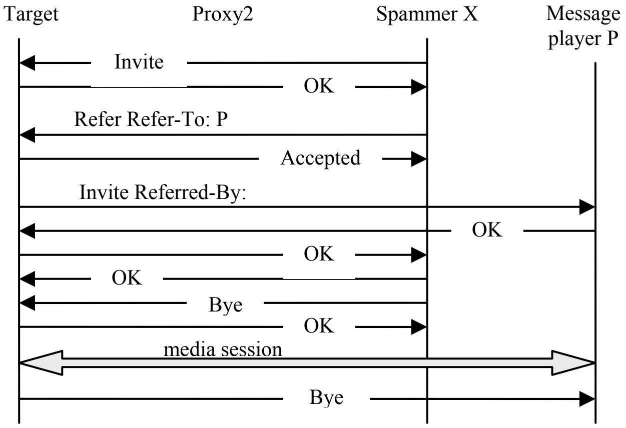

Figure 2. SPIT originated from the different location over SIP.

IV. Statistics for Signaling

Every VoIP signaling protocol has its specific session setup and termination requests. For SIP, these are INVITE and BYE respectively; for H.323 standard [14] and its component H.225.0 [15], these are *Setup* and *ReleaseComplete* messages.

A. Detection statistics

The idea is to monitor the VoIP signaling traffic on the recipients' access domain Gateway. As we said before, we target the problem of "external" spammer only, assuming that preventing spam outgoing from the controlled local network is an easier task. For each "external" identity *X* observed in the signaling routing data, we maintain four counters for the number of times call setup and call termination requests passed out and into the access domain, i.e. made *to X* and received *from X* within a certain time window: $\mathrm{ST}_X^{\mathrm{to}}$, $\mathrm{TR}_X^{\mathrm{to}}$, $\mathrm{ST}_X^{\mathrm{from}}$, $\mathrm{TR}_X^{\mathrm{from}}$. Since the requests are considered a priori as independent events, these counters effectively are sums of independent binomial variables, and the normalized sums of these variables can be approximated by the Normal Distribution, as follows from the Central Limit Theorem [16]. In the simplest case, when every call that *X* participates in can be initiated (terminated) by *X* with the expected probability ½, the basic statistics look like

$$S = (\mathrm{ST}_X^{\mathrm{from}} - ½\mathrm{ST}_X)/(¼\mathrm{ST}_X)^{1/2} \sim N(0,1), \qquad (1)$$
$$T = (\mathrm{TR}_X^{\mathrm{from}} - ½\mathrm{TR}_X)/(¼\mathrm{TR}_X)^{1/2} \sim N(0,1), \qquad (2)$$

where $\mathrm{ST}_X = \mathrm{ST}_X^{\mathrm{to}} + \mathrm{ST}_X^{\mathrm{from}}$, $\mathrm{TR}_X = \mathrm{TR}_X^{\mathrm{to}} + \mathrm{TR}_X^{\mathrm{from}}$. Simple transformations lead (1) and (2) to

$$S = (\mathrm{ST}_X^{\mathrm{from}} - \mathrm{ST}_X^{\mathrm{to}})/(\mathrm{ST}_X^{\mathrm{from}} + \mathrm{ST}_X^{\mathrm{to}})^{1/2} \sim N(0,1), \qquad (3)$$
$$T = (\mathrm{TR}_X^{\mathrm{from}} - \mathrm{TR}_X^{\mathrm{to}})/(\mathrm{TR}_X^{\mathrm{from}} + \mathrm{TR}_X^{\mathrm{to}})^{1/2} \sim N(0,1). \qquad (4)$$

Thus, simultaneous deviation of those counters from their assumed probabilistic averages for expected signaling traffic patterns can indicate spamming activity coming from a particular location *X*. Less probable the deviation is, more likely SPIT activity presents in the traffic. The probability of deviation can be calculated based on the quantiles of the Normal Distribution [16]. The values that should be monitored are *S* and $|T|$, since terminations prevailing in any direction would indicate spam, as follows from the scenarios description. For instance, if overall reliability of detection is set to 0.9999, *S* should exceed 2.33 and $|T|$ should exceed 2.58 at the same time. Practically, this means that if at least *n* out of *N* counted calls made from the location *X* are the ones that conform precisely one of described scenarios (A, B, C or D), *X* would be identified as a SPIT source. The examples of corresponding values of *n* and *N* are:

26 of 100 (26%), 58 of 500 (11.6%), 116 of 2000 (5.8%), 183 of 5000 (3.7%). (5)

In order to resolve the cases when calls with voice mail deposit (scenario D) might be mixed with scenario types A or C so that not to take into account these calls, additional parameter can be introduced. Denote by $\mathrm{ST}_{\mathrm{M}X}^{\mathrm{from}}$ the number of call setup requests passed by the Gateway to its local customers' voice mailboxes. These events can be monitored, as mentioned earlier. Thus, the adjusted statistics T will look like

$$(\mathrm{TR}_X^{\mathrm{from}} - \mathrm{ST}_{\mathrm{M}X}^{\mathrm{from}} - \mathrm{TR}_X^{\mathrm{to}})/(\mathrm{TR}_X^{\mathrm{from}} - \mathrm{ST}_{\mathrm{M}X}^{\mathrm{from}} + \mathrm{TR}_X^{\mathrm{to}})^{1/2}. \quad (6)$$

When the spam media session can be initiated by third party, yet another statistics is needed. We consider the example of SIP. As follows from the Figure 2, each REFER message should decrease $\mathrm{TR}_X^{\mathrm{from}}$ by one to keep the original statistics T intact. Therefore, assuming $\mathrm{RE}_{\mathrm{P}X}^{\mathrm{from}}$ as the number of REFER messages that were passed to the targeted network and had URI X in the *Contact:* header field, (4) takes the form of

$$\mathrm{T} = \frac{\mathrm{TR_X}^{\mathrm{from}} - \mathrm{ST_{MX}}^{\mathrm{from}} - \mathrm{RE_{PX}}^{\mathrm{from}} - \mathrm{TR_X}^{\mathrm{to}}}{\sqrt{\mathrm{TR_X}^{\mathrm{from}} - \mathrm{ST_{MX}}^{\mathrm{from}} - \mathrm{RE_{PX}}^{\mathrm{from}} + \mathrm{TR_X}^{\mathrm{to}}}}. \quad (7)$$

On the other hand, the relationships between the targeted network and SPIT content source P on the Figure 2 have been already covered by monitoring S and $|T|$ for entity P.

Another correction that has to be made in the light of possibility of scenario E is that all those outgoing setup requests contributing in $\mathrm{SET}_X^{\mathrm{to}}$ in (3) must be originated by the sender. This means that the request does not count for statistics if there are signs that it was induced by the third party. For SIP this implies that outgoing SIP INVITE's must not contain *Referred-By:* header field.

B. Reaction to detected SPIT

Reactive spam call handling measures have to be undertaken by a local service provider on the Gateway upon detection of X as a SPIT source. The actions applied to the subsequent incoming call setup requests asking to establish the call with X, can be categorized in three classes:

- "warning": display the text warning on the phone, use special ringing tone
- "call delay": switch the caller to the recipient's voice mail, reject the request and report the callerID and the call at a later time as a missed one
- "call cancellation": drop the call setup on behalf of recipient

In general, any mitigation actions would affect a detection statistics that relies on recipients' reaction. In our case, as soon as a major part of unsolicited call setup requests do not result in media sessions after the reactive measures are applied, corresponding termination requests would be subsequently eliminated from T calculations for X. To continue tracking the spam activity from the identified spam source and for keeping the reactive measures activated, we continue to rely on statistics S only for this particular source X. T may be employed again and reactive measures may be deactivated either after S comes back to the "no-alarm" value for X, after some period of time or by operator command.

C. Limitations of the identity-based statistics

Some legitimate phone services or entities may behave like spam in some or all aspects, such as automated phone notification service (e.g. "Your book is ready for pick up"). Therefore, setup and termination statistical imbalances should be used in combination with white and black lists to provide low false positive and false negative detection rates.

Statistical counters need certain time and material to initialize, since only those "external" entities that accumulate enough events within the given time frame can be monitored. Empirically, the expected average value must be at least 10 to make the approximation by the Normal Distribution valid.

Spammer can try to hide his real identity from the recipients. Possible ways to do so were described in the "Anonymity" section. As a result, the identity that the recipient can deal with could be the one of the anonymizer A, signaling gateway GW2 or sRL. The spammer might be not the only one whose call setup requests assume the identity of A or GW2, but if SPIT constitutes certain percentage (5) of such calls with identities of A or GW2, the SPIT will be detected and second, subsequent reactive measures would affect along with the spammer only those who also hide their real face.

Finally, sRL could be a temporarily assumed username. If the spammer manages to re-register himself often enough (for every few calls as an extreme case), there is no way to build a statistic per this volatile identity. An assumption that could be made is that sRL is constant for a reasonable time period; however this is the most serious limitation for any approach based on statistics per user. Both short-term and long-term monitoring would be needed to maintain on the Gateway to detect also low SPIT activity targeting the covered network.

References

[1] MarTel International, Inc., *PreEmptive Dialer*®.

[2] Privacy Rights Clearinghouse, "Prerecorded Telemarketing Calls to Those with an Existing Business Relationship (EBR): Comments to the Federal Trade Commission," January 2005.

[3] K. Srivastava and H. Schulzrinne, "Preventing Spam for SIP-based Instant Messages and Sessions," Columbia University Technical Report CUCS-042-04, October 2004.

[4] J. Rosenberg, C. Jennings and J. Peterson, "The Session Initiation Protocol (SIP) and Spam," IETF draft, October 2004.

[5] R. Pierce Reid, "Voice Spam Spam, Spamity Spam," *Qovia, Inc.*, white paper, July 2004.

[6] J. Kuthan, "Security in SIP Deployments," proceedings of *International SIP 2005*.

[7] BellSouth Corp., *Privacy Director*® Service

[8] Federal Trade Commission, *National Do Not Call Registry*.

[9] J. Rosenberg et al., "SIP: Session Initiation Protocol," IETF RFC 3261.

[10] J. Peterson, "A Privacy Mechanism for the Session Initiation Protocol (SIP)," IETF RFC 3323.

[11] G. Camarillo, A. B. Roach, J. Peterson and L. Ong, "Integrated Services Digital Network (ISDN) User Part (ISUP) to Session Initiation Protocol (SIP) Mapping," IETF RFC 3398.

[12] B. Campbell and R. Sparks, "Control of Service Context using SIP Request-URI," IETF RFC 3087.

[13] R. Sparks, "The Session Initiation Protocol (SIP) Refer Method," IETF RFC 3515.

[14] International Telecommunications Union, "Packet-based multimedia communications systems," ITU-T Recommendation H.323

[15] International Telecommunications Union, "Call signalling protocols and media stream packetization for packet-based multimedia communication systems," ITU-T Recommendation H.225.0

[16] R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics,* 5th ed., Prentice Hall, 1994, pp.246-258.